Applied Intelligence (2024) 54:6483–6514
https://doi.org/10.1007/s10489-024-05509-8

A survey of explainable knowledge tracing

Yanhong Bai[1] · Jiabao Zhao[1] · Tingjiang Wei[1] · Qing Cai[2] · Liang He[1]

Accepted: 4 May 2024 / Published online: 16 May 2024

Abstract
With the long-term accumulation of high-quality educational data, artificial intelligence (AI) has shown excellent performance in knowledge tracing (KT). However, due to the lack of interpretability and transparency of some algorithms, this approach will result in reduced stakeholder trust and a decreased acceptance of intelligent decisions. Therefore, algorithms need to achieve high accuracy, and users need to understand the internal operating mechanism and provide reliable explanations for decisions. This paper thoroughly analyzes the interpretability of KT algorithms. First, the concepts and common methods of explainable artificial intelligence (xAI) and knowledge tracing are introduced. Next, explainable knowledge tracing (xKT) models are classified into two categories: transparent models and "black box" models. Then, the interpretable methods used are reviewed from three stages: ante-hoc interpretable methods, post-hoc interpretable methods, and other dimensions. It is worth noting that current evaluation methods for xKT are lacking. Hence, contrast and deletion experiments are conducted to explain the prediction results of the deep knowledge tracing model on the ASSISTment2009 by using three xAI methods. Moreover, this paper offers some insights into evaluation methods from the perspective of educational stakeholders. This paper provides a detailed and comprehensive review of the research on explainable knowledge tracing, aiming to offer some basis and inspiration for researchers interested in the interpretability of knowledge tracing.

Keywords Explainable artificial intelligence · Knowledge tracing · Interpretability · Evaluation

1 Introduction

The emergence and application of numerous educational tools, such as profiling and prediction [1], intelligent tutoring systems [2, 3], assessment and evaluation [4], adaptive systems and personalization [5, 6], are transforming traditional methods of teaching and learning. Knowledge tracing (KT) is an important research direction in the field of artificial intelligence in education (AIED) that can automatically track the learning status of students at each stage. KT has been widely used in intelligent tutoring systems, adaptive learning systems and educational gaming [7, 8]. Recently, deep learning-based methods have significantly improved the performance in KT tasks; however, this improvement comes at the cost of interpretability [9, 10]. A lack of explainability is not conducive for stakeholders to understand the reasons behind an algorithm's decisions, which may reduce stakeholders' trust in these tools. For instance, if a knowledge tracing model yields unrealistic predictions, teachers may fail to understand the actual knowledge level of their students, and students may not receive an accurate assessment of their weaknesses. In addition, it is not easy for users or regulators to find defects in black box applications, which may raise security issues, such as learner resistance [11] or an increase in high-risk students [12]. To solve the above issues, researchers have been attempting to improve the interpretability of AI in various educational tasks, such as explainable learner models [13–15], explainable recommender systems [16], explainable

✉ Jiabao Zhao
jbzhao@mail.ecnu.edu.cn

Yanhong Bai
Lucky_Baiyh@stu.ecnu.edu.cn

Tingjiang Wei
mxdlzg@163.com

Qing Cai
qcai@psy.ecnu.edu.cn

Liang He
lhe@cs.ecnu.edu.cn

[1] The School of Computer Science and Technology, East China Normal University, Shanghai 200062, Shanghai, China

[2] The School of Psychology and Cognitive Science, East China Normal University, Shanghai 200062, Shanghai, China

at-risk student prediction [17, 18], and explainable personalized interventions [19].

To the best of our knowledge, there has not been a comprehensive survey of related research on the interpretability of knowledge tracing. This paper aims to fill this gap. Compared to the existing knowledge tracing surveys [11, 20–22], this study primarily focuses on explainable algorithms and knowledge tracing interpretability. The motivation behind this paper is threefold. First, it aims to provide a detailed and comprehensive review of the research on explainable knowledge tracing (xKT). Second, different interpretability methods for knowledge tracing should be compared, and evaluation methods should be explored. Finally, this study aims to provide a foundational and inspirational resource for researchers interested in the field of explainable knowledge tracing.

1.1 Contributions

The contributions of this survey are to provide an in-depth examination of the current status of explainable knowledge tracing. By doing so, it aims to establish a solid foundation of understanding and inspire additional research interest in this rapidly growing area.

Inspired by the classification criteria of xAI for complex object models as delineated by Arrieta et al. [9], this paper offers a novel categorization of knowledge tracing models into two distinct types: transparent models and black-box models. This dichotomy is further explored with a detailed examination of interpretable methods tailored to these models across three critical stages: ante hoc, post hoc, and other dimensions.

Moreover, the current evaluation methods for explainable knowledge tracing are still lacking. In this paper, contrast and deletion experiments are conducted to explain the prediction results of the deep knowledge tracing model on the same dataset by using three XAI methods. Furthermore, this work extends an insightful overview into the evaluation of explainable knowledge tracing, tailored to varying target audiences, and delves into the prospective directions for the future development of explainable knowledge tracing.

1.2 Systematic literature review (SLR) and execution

In this paper, the systematic literature review (SLR) methodology was adopted. This methodology was developed by Kitchenham and Charters [23] and is specifically designed for comprehensive analyzes in software engineering and computer science. The SLR methodology is significant because of its systematic approach to collating and synthesizing literature, providing a comprehensive understanding of the interpretability of knowledge tracing. The survey process commenced with the formulation of specific research questions, focusing on the application and implications of xAI in knowledge tracing. The primary questions addressed were as follows: a) How can the interpretability of knowledge tracing algorithms be improved? b) What are the applications and classifications of xAI in knowledge tracing? c) How can explainable knowledge racing models be effectively evaluated?

Our search strategy involved a comprehensive list of keywords, such as “explainable artificial intelligence”, “xAI”, “explainable”, “explainability”, “explanation”, “interpretable” and “knowledge tracing”. These keywords were chosen based on their prevalence in the current literature and relevance to our research questions. We used the Boolean operators ‘AND’ and ‘’OR’ to construct detailed search strings. We conducted searches in databases such as IEEE Xplore, ACM Digital Library, Science Direct, and SpringerLink. These databases were selected for their extensive coverage of computer science and AI literature. From an initial pool of 1,783 studies, 517 were screened based on their relevance to knowledge tracing. After a full-text review, 59 papers were selected based on our inclusion and exclusion criteria, and the search period ended on November 2023.

Inclusion Criteria: 1) Relevance: Studies must focus on the interpretability of knowledge tracing algorithms, including theoretical analyses, application case studies, and empirical evaluations. 2) Novelty and Contribution: Studies should offer novel insights or approaches in the field of knowledge tracing interpretability. This includes introducing new methodologies, providing unique theoretical perspectives, or presenting novel empirical findings that significantly advance the understanding of the topic. 3) Publication Quality: Studies must be published in peer-reviewed journals or conference proceedings. In cases where no peer-reviewed version is available, but the study is highly relevant to the research topic, non-peer-reviewed versions (e.g., arXiv preprints) will also be considered for inclusion. 4) Language: The study must be written in English. **Exclusion Criteria:** 1) Relevance: Studies not directly addressing the research questions, specifically those not focusing on the interpretability of knowledge tracing algorithms, will be excluded. 2) Empirical and Methodological Rigor: Studies lacking empirical data support or detailed methodological descriptions will be excluded. This includes opinion pieces or conceptual framework studies without specific empirical analysis.

The number of papers published each year is shown in Fig. 1. Research on the interpretability of knowledge tracing has shown a significant increasing trend since 2019, indicating that researchers have realized that high accuracy alone is insufficient to gain the trust of stakeholders when applying AI models to real-world educational scenarios, and improving the interpretability of model decisions is a crucial issue that needs to be addressed.

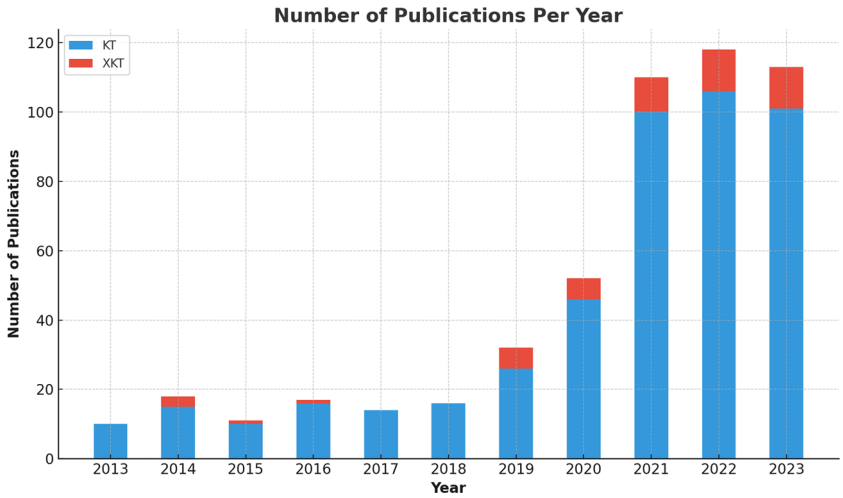

Fig. 1 The number of publications per year: Knowledge tracing and knowledge tracing with interpretability

1.3 Structure

The remainder of this survey is organized as follows. Section 2 discusses the relevant research on explainable artificial intelligence and knowledge tracing while also emphasizing the importance of interpretability in knowledge tracing algorithms. Section 3 delves further into explainable knowledge tracing, offers insights into its various dimensions and implications, and presents a detailed examination of interpretable methods suitable for explaining knowledge tracing models. The focus of Section 4 is on the methodologies for scientifically evaluating the interpretability of knowledge tracing models. Finally, Section 5 outlines future research directions in explainable knowledge tracing, pinpointing key areas that warrant further investigation and innovation. To enhance readers' understanding of this paper's architecture, we depict it in detail in Fig. 2.

2 Background

This section thoroughly examines the developmental background of KT and xAI, presenting the latest frameworks and methodologies in these fields. It delves into the concepts, classifications, and evolution of KT models while revealing the inherent limitations of these models. The section also explores the ongoing challenge of finding a balance between model accuracy and interpretability, discussing how to achieve the optimal compromise between the two. Additionally, it offers an in-depth analysis of

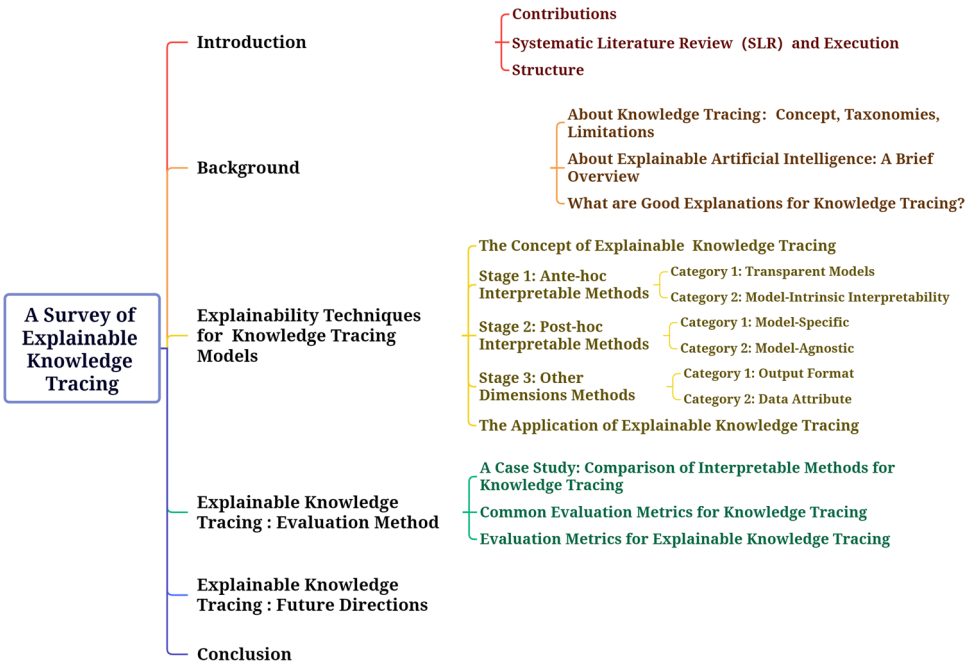

Fig. 2 The concept map of this survey

Table 1 The taxonomies of knowledge tracing

Category	Description	Representative Works
Markov process-based knowledge tracing	Assuming that a student's learning process is representable as a Markov process, it can be modeled with probabilistic models.	BKT [28] DBKT [29]
Logistic knowledge tracing	The logistic function represents the probability of a student answering an exercise correctly, under the premise that this probability is expressible as a mathematical function involving the student and the KC parameter. The logistic model posits that students' binary answers (correct or incorrect) adhere to Bernoulli distributions.	LFA [30] PFA [31] KTMs [32]
Deep learning-based knowledge tracing	DLKT models can simulate changes in students' knowledge states and encapsulate a broad spectrum of complex features, which might be challenging to extract through other methods.	DLKT: [27, 33–40]

xAI's fundamental principles, methodologies, and evaluation mechanisms, underscoring its crucial role in enhancing transparency and comprehension of complex AI systems. Specifically, the section explores the application of tailored explanation methods in the KT domain, proposing targeted solutions for both professional and lay users. By comprehensively analyzing the theoretical foundations and practical applications of KT and xAI, this chapter aims to provide groundbreaking insights into the interpretability of knowledge tracing.

2.1 About knowledge tracing: concepts, taxonomies, evolution, limitations

Knowledge tracing has become a key component of learner models. A large amount of historical learning trajectory information provided by an intelligent tutoring system (ITS) is used to model learners' knowledge states and predict their performance in future exercises [24]. Thus, knowledge tracing provides personalized learning strategies [25] and learning path recommendations [26] for education stakeholders and is a crucial element of adaptive education. Specifically, knowledge tracing is a task for predicting students' performance in future practice according to changes in learners' knowledge mastery in historical practice [27]; this task involves two main steps: 1) modeling learners' knowledge state according to their historical practice sequence and 2) predicting learners' performance in future practice. In other words, this task can be formulated as a supervised time series learning task, where $X_i = \{e_t, r_t\}$ represents a student's answer pair, e_t represents the exercise ID, and r_t represents the answer result for related exercise $e_t, r_t \in \{0, 1\}$ (1 indicates the correct answer and vice versa). Given a student's exercise sequence $X = \{x_1, x_2, x_3, \ldots, x_{(t-1)}\}$ and the next exercise, the task objective is to predict the correct probability $P(r_t = 1|X, e_t)$ of the exercise e_t.

According to the general classification method of knowledge tracing models, existing models can be classified into the following three categories [21]: 1) Markov process-based knowledge tracing, 2) logistic knowledge tracing, and 3) deep learning-based knowledge tracing (DLKT). The taxonomies of knowledge tracing are shown in Table 1. Next, we introduce a series of seminal works on the aforementioned three types of models and outline the timeline of knowledge tracing evolution, as shown in Fig. 3.

In 1994, Corbett et al. proposed Bayesian knowledge tracing (BKT) [28], which is based on a two-state Hidden Markov Model(HMM) that treats student knowledge states as hidden variables [41]. However, since the model uses a shared set of parameters for the same knowledge component (KC), it cannot personalize modeling for students at different levels. To overcome this limitation, researchers have added personalized features to make the model more realistic, leading to the emergence of various variations based on Bayesian knowledge tracing, marking the initial phase of knowledge tracing research. One notable improvement is dynamic BKT (DBKT) [29]. To address the issue of BKT modeling each KC individually, DBKT employs dynamic Bayesian networks to represent multiple KCs jointly in a single model. This approach models the prerequisite hierarchies and relationships within KCs. Both DKT and DBKT are representative models of knowledge tracing based on the Markov process. In 2006, Cen et al. [30] proposed learning factor analysis (LFA), which inherits the Q matrix used in psychometrics to assess cognition and extends the theory of learning curve analysis. An improved LFA model is performance factor analysis (PFA) [31], which was developed in 2009. Additionally, knowledge tracing machines (KTMs) [32] utilize a factorization machine to model all variable interactions. These three methods are representative models of logistic knowledge tracing, and a detailed explanation of each will be provided in Section 3.2. In general, logistic knowledge tracing has achieved better performance than BKT, and knowledge tracing has gradually entered a development period [42]. Since deep knowledge tracing (DKT) [27] was proposed in 2015, deep learning techniques have

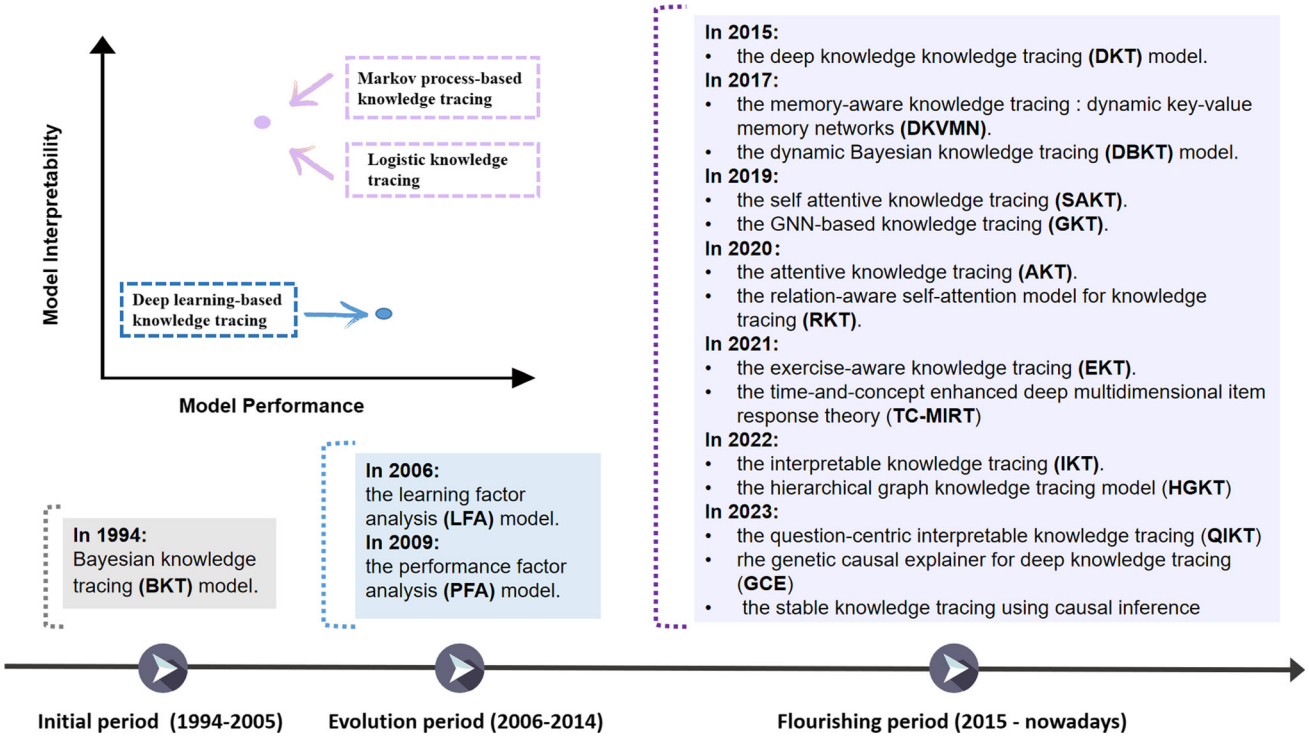

Fig. 3 The timeline of seminal works toward knowledge tracing

shown more vital feature extraction ability in knowledge tracing than the other two types of models. Based on this seminal work, much deep learning-based knowledge tracing (DLKT) has emerged. For example, researchers have applied deep learning techniques to knowledge tracing in various ways. Below, we list several categories of representative work: 1) memory-aware knowledge tracing: DKVMN [33]; 2) attention-aware knowledge tracing: SAKT [35] and AKT [36]; 3) graph-based knowledge tracing: GKT [37] and HGKT [43]; 4) relation-aware knowledge tracing: RKT [38]; 5) exercise-aware knowledge tracing: EKT [39]; and 6) interpretable knowledge tracing: TC-MIRT [44], IKT [40], QIKT [45], GCE [46], and stable knowledge tracing using causal inference [47].

As shown in Fig. 3, before the emergence of deep learning in knowledge tracing, Bayesian knowledge tracing and logistic knowledge tracing were widely used because of their relatively simple model structures and powerful interpretability [48]. However, due to the massive and multidimensional nature of online learning data, these two types of models were unable to achieve good performance on big data [49]. Deep learning-based models have a clear advantage in processing large datasets. However, when applied to real-world teaching scenarios, DLKT may face the following challenge: the large number of network layers and parameters in deep networks may limit the interpretability of the generated parameters. Additionally, the lack of interpretability in deep learning-based models can also lead to potential ethical and privacy issues. Stakeholders need to be able to trust the models and understand how the models make decisions.

Overall, DLKT models exhibit strong performance but poor interpretability[11], while simple models with strong interpretability are far weaker than the former, as shown in the upper left corner of Fig. 3. Consequently, the tradeoff between interpretability and performance poses a significant challenge for researchers challenge for researchers. In recent years, researchers have utilized various methods to explain knowledge tracing models and have attempted to maximize transparency while ensuring model performance. In Section 3, we will elaborate on the interpretable methods existing in the above proposed models.

2.2 About explainable artificial intelligence (xAI): a brief overview

The goal of explainable artificial intelligence (xAI) is to provide an understanding of the internal workings of a system in a manner that humans can comprehend. XAI aims to answer questions such as “How did the model arrive at this result?”, “Will different inputs yield the same result?”, and “What is the reliability of the model’s outputs?”. In essence, xAI’s purpose is to provide an explanation to the explainees regarding why the model generates the corresponding output based on the input. Based on the diverse needs of explainees,

explainers offer appropriate types of explanations. The process of explanation provided by xAI enhances explainees' degree of trust in the system, thus increasing the system's utility ratio across various industries. To enhance the readers' understanding of explainable artificial intelligence, we introduce the xAI framework illustrated in Fig. 4. Improving the interpretability of algorithms is important in AIED, and the benefits can be summarized as follows: 1) Developers can enhance the transparency of models in a more scientific way, which can lead to better model optimization. 2) Transparency can help domain experts discover the cognitive rules in the learning process, leading to deeper insights and better decision-making. 3) Transparency can help users better understand the reasons and logic behind AI-driven decisions, which can increase their trust in the technology. 4) Regulatory authorities can use transparency to achieve effective supervision and ensure the safety of intelligent products used in education while also ensuring compliance with the law.

Existing algorithms can be classified as transparent (white box) or black box models, depending on their complexity [67] (details in Section 3.1). Transparent models are characterized by simple internal components and self-interpretability, allowing users to intuitively understand their internal operation mechanism [68]. A black box model refers to a model with a complex, nonlinear relationship between the input and output, with an operating mechanism that is difficult to understand, such as that of a neural network [69]. Based on the characteristics of these two models, xAI methods are typically categorized as ante-hoc interpretable methods or post-hoc interpretable methods [70]. Ante-hoc interpretable methods are mainly applied to models with simple structures and strong interpretability (such as transparent models) or to build interpretable modules in the model to make it intrinsically interpretable. In contrast, post-hoc interpretable methods, such as black box models, develop interpretive techniques to interpret trained machine learning models. Post-hoc interpretable methods are usually subdivided into model-specific approaches and model-agnostic approaches based on their application scope [71]. These methods are introduced in detail in the following sections. Table 2 provides an overview of representative interpretable methods.

There is currently no widely accepted scientific evaluation standard for xAI. Different experts from various disciplines have conducted preliminary investigations based on different evaluation objectives, such as the characteristics of the model being evaluated or the requirements of users and application scenarios [72]. One prominent evaluation method is the three-level approach proposed by Doshi Velez et al. [73], which includes the following steps: 1) application-grounded evaluation, 2) human-grounded evaluation, and 3) functionally-grounded evaluation. These three levels of evaluation provide useful frameworks for evaluating the interpretability of xAI systems in different contexts.

Generally, application-level evaluation is considered an effective approach because the interpretation is applied to the appropriate field and evaluated by professionals, resulting in

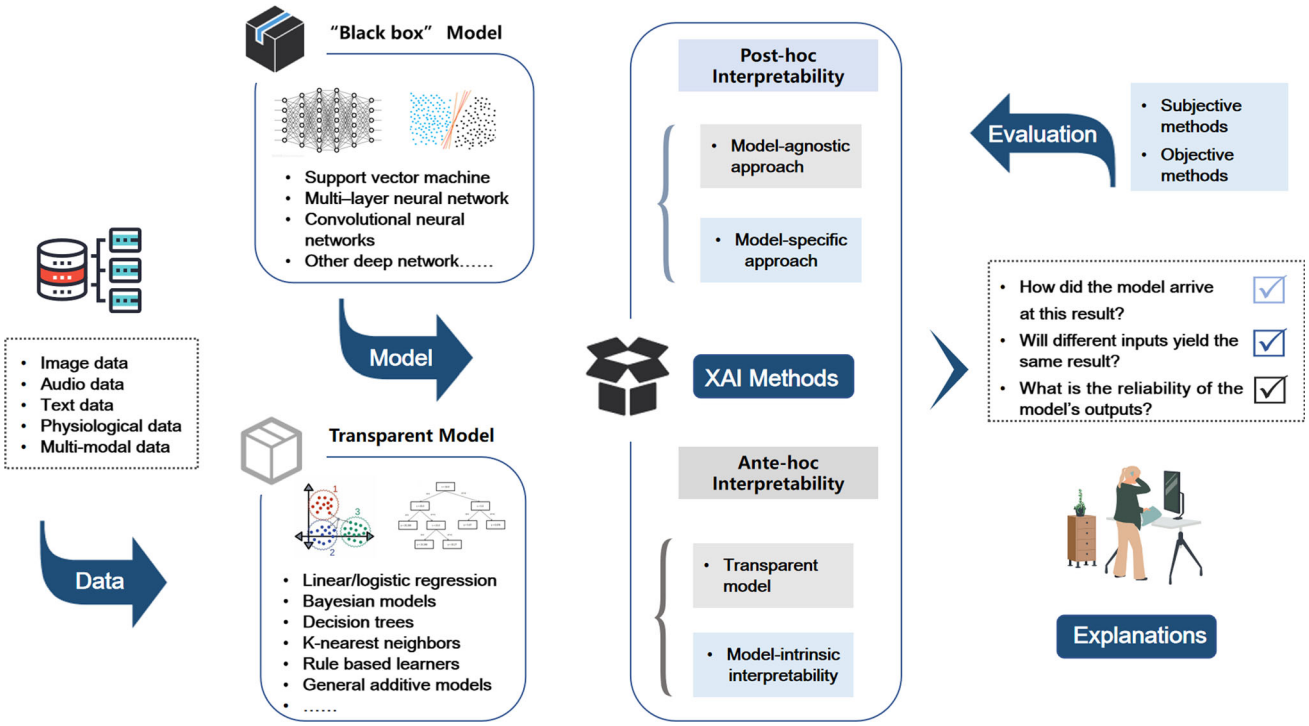

Fig. 4 The framework of explainable artificial intelligence (xAI)

Table 2 A summary of representative interpretable methods

Methods	Year	Stage		Category		Domain	Description
		Anc-hoc	Post-hoc	Model-specfic	Model-Agnostic		
Attention [50]	2014	✓		✓		CV/NLP	Attention weight matrix visualization
Bayes Rule List [51]	2015	✓		✓		–	Trees and Rule-based Models
Generalized additive models (GAMs) [52]	2015	✓		✓		–	The final decision form is obtained by combining each single feature model with linear function
Neural Additive Model [53]	2020	✓		✓		CV	Train multiple deep neural networks in an additive fashion such that each neural network attend to a single input feature
Activation Maximization [54]	2010		✓	✓		CV	Maximize neuronal activation by identifying the optimal input for a neuron at a specific network layer
Gradient-based Saliency Maps [55]	2013		✓		✓	CV	The back propagation mechanism of DNN is used to propagate the decision importance signal of the model from the output layer neurons to the input of the model layer by layer to deduce the feature importance of the input samples
DeConvolution Nets [56]	2014		✓		✓	CV	
Guided Backprops [57]	2015		✓		✓	CV	
SmoothGrad [58]	2017		✓		✓	CV	
Layer-wise Relevance BackPropagation (LRP) [59]	2015		✓		✓	CV/NLP	
Salient Relevance (SR) Map [60]	2019		✓		✓	CV	
Class Activation Mapping (CAM) [61]	2016		✓		✓	CV	The neural network's feature map is utilized to ascertain the significance of each segment of the original image
Grad-CAM [62]	2017		✓		✓	CV	
Grad-CAM++ [63]	2018		✓		✓	CV	
Local Interpretable Model-Agnostic Explanations (LIME) [64]	2016		✓		✓	CV	An interpretable model with simple structure is used to locally approximate the decision result of the model to be explained for an input instance
SHapley Additive exPlanations (SHAP) [65]	2017		✓		✓	CV	Reflects the influence of each feature in the input sample and shows the positive and negative influence
Concept Activation Vectors(CAV) [66]	2018		✓		✓	CV	Measures the relatedness of concepts within the model's output

more convincing results. Currently, the popular classification method for xAI evaluation involves dividing it into subjective and objective evaluations based on whether humans are involved [80, 86]. Table 3 shows the evaluation methods, metrics, and current limitations of both subjective and objective evaluation in xAI. It is important to carefully select appropriate evaluation metrics based on the specific features and goals of the evaluated xAI system. Overall, the combination of subjective and objective evaluation methods may be the most effective approach for assessing the interpretability of xAI systems while balancing cost and performance.

2.3 What are good explanations for knowledge tracing?

According to Merriam-Webster, "interpret" means to present something in understandable terms and explain its meaning [87]. However, what constitutes a good explanation varies across different fields, and experts have attempted to define it in different ways. In computer science, Lipton [88] emphasized the importance of comparative explanations, i.e., whether the predicted outcome Y will change for different inputs X. Physicist Max Tegmark described

Table 3 Classification of xAI Evaluation Methods Based on User Involvement

Angles	Methods	Limitations
Subjective evaluation	Qualitative evaluation based on open-ended questions [74–76]; Quantitative evaluation based on closed-ended questions [77, 78]; A mixed-methods approach that combines both qualitative and quantitative evaluation. [79, 80]	This method may be susceptible to bias and variability owing to individual differences among evaluators. Moreover, the requisite for professional human resources can lead to elevated evaluation costs.
Objective evaluation	Fidelity [81]; Consistency [82]; Stability [80]; Sensitivity [80]; Causality [83]; Complexity [84, 85]	Models might rely on particular evaluation methods, and varying metrics could yield disparate results.

a good explanation as one that answers more questions than asked [89]. Moreover, psychology researchers have highlighted the significance of explanations in learning and inference and how individuals' explanatory preferences can impact explanation-based processes in a systematic way [90]. In certain scenarios of adaptive education, researchers have used verbal and visual explanations [91] or interactive interfaces [92] to provide explanations and have achieved positive outcomes. For example, Cristina Conati et al. [93] added an interactive simulation program to the adaptive CSP (ACSP) applet to provide an explanation function. The research results demonstrated that providing explanations can enhance students' trust in ACPS prompts.

Explanations also play a pivotal role in the process of knowledge tracing, but what are good explanations for knowledge tracing? Research suggests that explanations must be audience-specific and goal-oriented [72, 94]. Stakeholders in knowledge tracing are divided into professional users (developers, researchers) and non-professional users (teachers, students), each requiring tailored explanations [95]. For professionals, explainability enhances understanding, system debugging, model optimization, and credibility [96]. For non-professionals, it facilitates comprehension of learning processes, encourages result acceptance, and boosts model satisfaction [97]. The field has developed methods ensuring both accuracy and transparency, making model operations and decisions clear to all users, thereby improving interpretability. Further details on these methods will be provided in the next section.

3 Explainability techniques for explainable knowledge tracing models

This section delves into the key aspects of enhancing the interpretability of knowledge tracing models. Initially, we categorize and discuss explainable knowledge tracing models, focusing on the critical distinctions between transparent models and complex black-box models. This discussion lays the groundwork for understanding the internal mechanisms of these models. Subsequently, we shift our focus to exploring methods for augmenting the interpretability of knowledge tracing. These methods encompass both ante-hoc and post-hoc strategies, as well as other dimensions. The aim is to reveal how various approaches can enhance the models' transparency and comprehensibility. Finally, we examine the practical implementation of explainable knowledge tracing in real-world applications, such as generating diagnostic reports. The section concludes with a critical discussion evaluating the balance between the models' interpretability, accuracy, and their practical application in educational settings.

3.1 The concept of explainable knowledge tracing

As mentioned in Section 2.2, machine learning models are typically categorized as transparent or black box models according to the complexity of the objects they are intended to explain, based on the criteria of xAI. Transparent models are characterized by high transparency of internal components and self-interpretability, such as, linear/logistic regression [98, 99], bayesian models [100–102], decision trees [103, 104], k-nearest neighbors [105–107], rule based learners [108–110], general additive models [111–113], etc. For transparent models, interpretability can be understood from three perspectives: algorithmic transparency, decomposability, and simulatability [114]. However, models such as multi-layer neural network [115, 116] and other deep network [115, 117, 118], which have complex internal structures and difficult operation mechanisms, are usually referred to as black box models. XAI algorithms are capable of understanding the architecture and layer configurations of transparent models, but they lack the ability to comprehend the operational mechanisms inherent in black box models[119].

Explainable knowledge tracing model taxonomy Inspired by the criteria for classifying model complexity in xAI [9], we propose a novel taxonomy specifically tailored for knowledge tracing models. This framework categorizes models based on their explainability and comprehensibility. We identify models employing Markov processes and logistic regression as "transparent models" due to their straightforward structures and the ease with which users can understand

them. This classification is rooted in the models' interpretability and the transparency of their decision-making processes, emphasizing the accessibility and interpretability of how decisions are made. For example, state transitions in Markov models and parameter settings in logistic regression models are intuitive, making these models' decision-making processes easily traceable and explainable. In contrast, knowledge tracing models based on deep learning, especially those involving complex multi-layered network structures, are considered akin to "black box" by users. The internal mechanisms of these models are difficult to comprehend because of their complexity and rich nonlinear characteristics, obscuring the internal decision-making process. Consequently, these models and their variants are categorized as "black-box models". We define explainable knowledge tracing and show its framework in Fig. 5.

Methodologies for explainable knowledge tracing In the following section, we delve into interpretable methods for the two types of knowledge tracing models mentioned earlier, categorizing them into three phases: 1) Ante-hoc interpretable methods; 2) Post-hoc interpretable methods; and 3) Other dimensions. Ante-hoc interpretable methods that focus on transparent models and model-intrinsic interpretability, which can be achieved by simplifying the model's structure or incorporating intuitive modules to enhance its interpretability. Post-hoc interpretability methods, on the other hand, center around model-specific and model-agnostic approaches, such as using external tools or techniques to elucidate the model's decision-making process. Other dimensions include interpretability methods that are specific to knowledge tracing models but have not yet been widely discussed in the current xAI literature. An example would be leveraging the associative relationships between problems, concepts, or users within the context of knowledge tracing. To aid readers in locating various interpretable methods, we present a framework diagram in Section 3 in Fig. 6. Additionally, all the reviewed explainable knowledge tracing methods are summarized in Table 6.

3.2 Stage 1: ante-hoc interpretable methods

As mentioned in Section 2.2, ante-hoc interpretable methods are primarily used for transparent models or model-intrinsic interpretability [120]. These methods aim to make the model itself capable of interpretation by training a transparent model with a simple structure and strong interpretability or by building interpretable components into a complex model structure [9, 119]. This paper mainly focuses on two types of ante-hoc interpretable methods for xKT: transparent models and model-intrinsic interpretability; these methods are described in the following subsections.

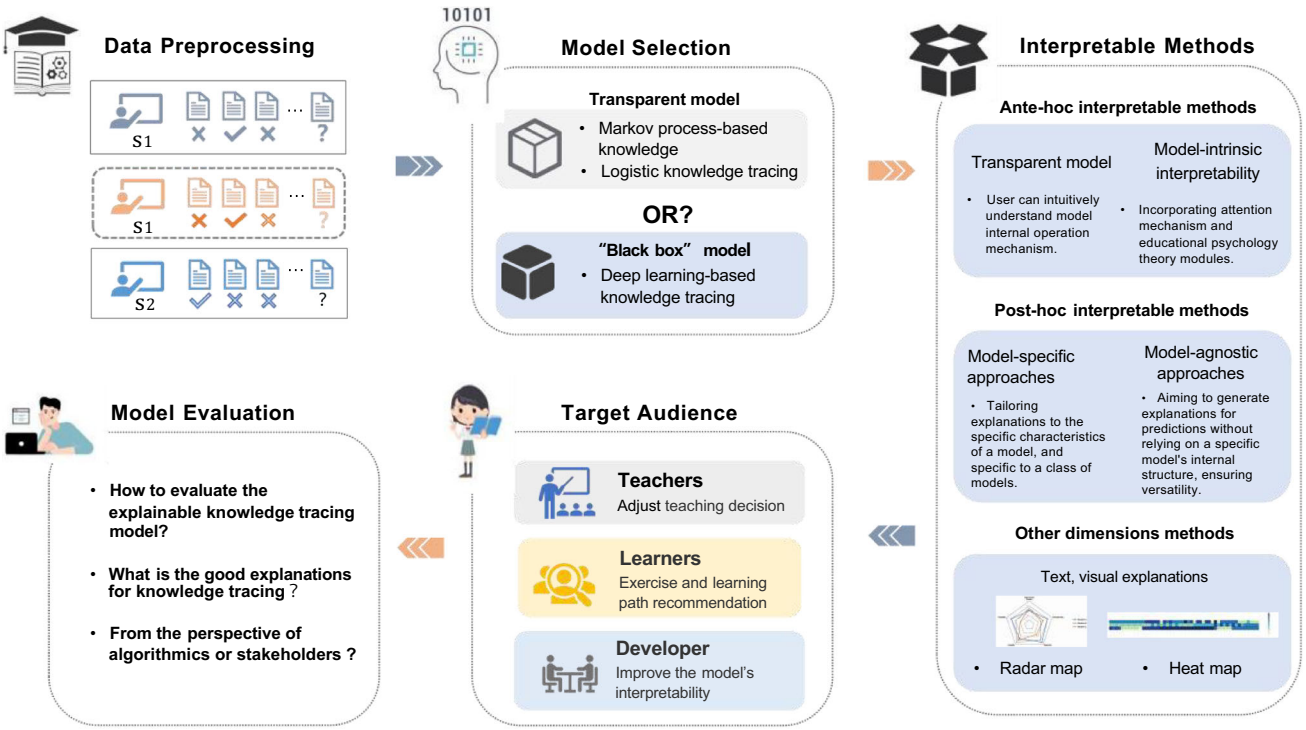

Fig. 5 The framework of explainable knowledge tracing(xKT)

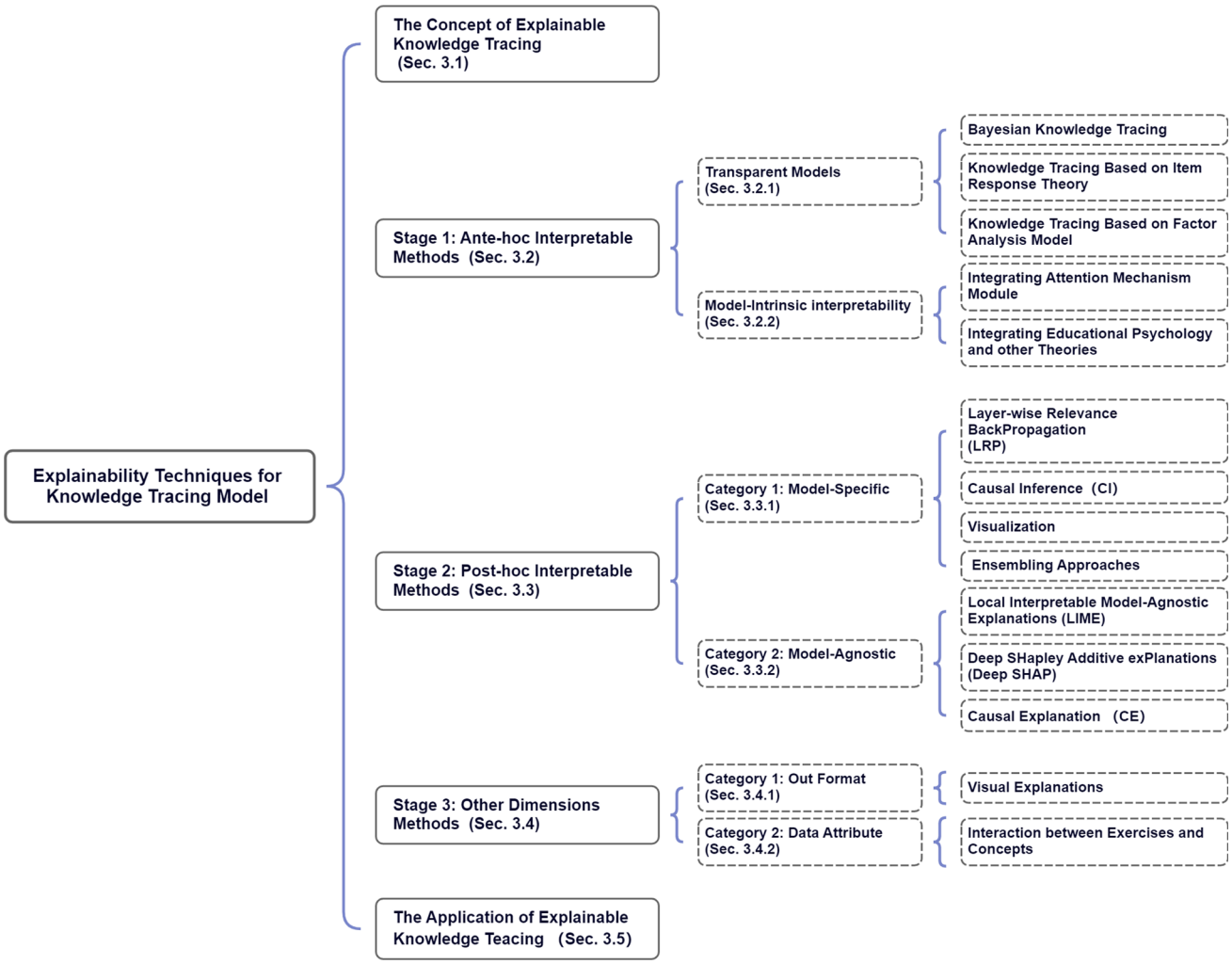

Fig. 6 The framework of explainability techniques for the knowledge tracing model

3.2.1 Category 1: transparent models

Based on the classification criteria in Section 3.1, knowledge tracing models that use the Markov process and logistic regression are considered knowledge tracing transparent models due to their easy-to-understand internal structure and operational process. The following sections elaborate on how the knowledge tracing transparent model achieves interpretability.

Bayesian knowledge tracing (BKT) BKT is a probabilistic model grounded in the Markov process, predicting students' skill mastery by updating beliefs based on performance. Although not as performant as deep knowledge tracing models, BKT's use of the HMM framework ensures a satisfactory and comprehensible explanation of the knowledge tracing. Some extended models of BKT are personalized by incorporating individual student characteristics. For example, Pardos et al. [121] set different initial probabilities for different students to achieve partial personalization. However, differentiating the initial probabilities only partially achieves personalization, and Lee et al. [122] used a separate set of personalization parameters for each student to adequately model interindividual differences. Although the personalized parameters are good for conferring variability across individuals, they do not consider the influence of knowledge concepts. Hawkins et al. [123] proposed BKT-ST to calculate the similarity between current knowledge concepts and those learned previously, enhancing the model's capacity to represent interconnected linkages across concepts. In addition, Wang et al. [124] proposed the use of multigrained-BKT and historical-BKT to model the relationships between different knowledge components (KCs). Moreover, Sun et al. [125] used a genetic algorithm (GA) to optimize the model to solve the exponential explosion problem when tracing multiple concepts simultaneously. Moreover, Moreover, the interpretable knowledge tracing (IKT) model proposed by

Minn et al. [40], which is distinct in its use of tree-augmented naive Bayes and focuses on skill mastery, learning transfer, and problem difficulty, offers greater interpretability and adaptability in student performance prediction. Beyond basic parameters and knowledge concepts, the BKT model is also increasingly taking into account students' emotions and additional behavioral aspects. For examples, Spaulding et al. [126] introduced the concept of affective BKT, integrating students' emotional states into the model. Furthermore, to accurately capture how students' memory retention changes over time, Nedungadi et al. [127] developed the PC-BKT model. This adaptation incorporates a temporal decay function to model the process of forgetting, offering a more nuanced understanding of students' learning and memory retention behaviors. The HMM describes the probabilistic relationship between observable and hidden variables, and the probabilistic relationship varies over time. In Bayesian knowledge tracing, the model estimates students' learning state by observing the results of students' responses to questions related to knowledge concepts. The state transfer probability calculation process and the decision process are transparent, and the change process of the model can be effectively observed through the state transfer diagram, as shown in Fig. 7. Here, $P(M)$ represents the probability of mastery, indicating a student's current understanding of the skill, $P(T)$ represents the probability of transition, which is the rate at which a student transitions from nonmastery to mastery, $P(G)$ is the probability of guessing correctly, accounting for lucky guesses, and $P(S)$ refers to the probability of slipping, where a mastered skill is incorrectly applied. In BKT, knowledge mastery is updated along with the learning parameters. When the probability of students mastering the relevant knowledge concept in the initial knowledge state is greater than 0.95 [128], students have mastered the knowledge concept.

In summary, from an interpretable perspective, the Bayesian knowledge tracing model has good transparency in the computational process as a purely probabilistic model. Observing different state transfers of BKT can provide a basis for modeling decisions for students and teachers. From the model perspective, on the one hand, Bayesian knowledge tracing models do not account for the differences in the initial knowledge levels of different students and lack an assessment of the difficulty of the questions. Although many models have been extended based on BKT, they still cannot be applied to knowledge tracing scenarios with large-scale data. On the other hand, these models assume that students do not forget the knowledge they have mastered, which is not consistent with actual cognitive characteristics. In addition, using binary groups to represent recent knowledge states does not match the real cognitive state situation. It is difficult to adequately predict the relationship between each exercise and specific knowledge concepts due to the ambiguous mapping between hidden states and exercises.

Knowledge tracing based on item response theory (IRT) Item response theory originates from the field of psychometrics and assumes that an underlying trait represents each candidate's ability and can be observed through their response to items. The two models underlying the IRT model are the normal ogive model and the logistic model. However, the logistic regression model is the most common in practical applications [129]. The IRT model is based on four assumptions: 1) monotonicity (the probability of a correct response increases as the level of the trait increases). 2) one-dimensionality (it is assumed that a dominant underlying trait is being measured). 3) local independence (responses to separate items in a given test are independent of each other at a given level of ability). 4) invariance (it is assumed that students' abilities remain constant over time).

Here, θ_i is defined as the individual ability parameter of the $i-th$ student, a_j is defined as the discrimination parameter of question j, b_j is defined as the difficulty parameter of question j, and c_j is defined as the guessing parameter of question j. With one-parameter IRT, it is possible to provide students with interpretable parameters in terms of two dimensions, personal ability and difficulty, as shown in Fig. 8. A two-parameter IRT model uses two parameters (difficulty

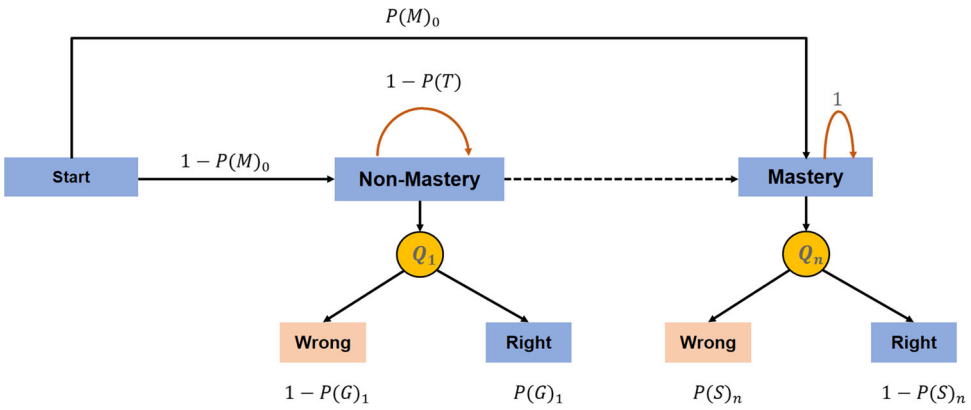

Fig. 7 State transitions for Bayesian knowledge tracing

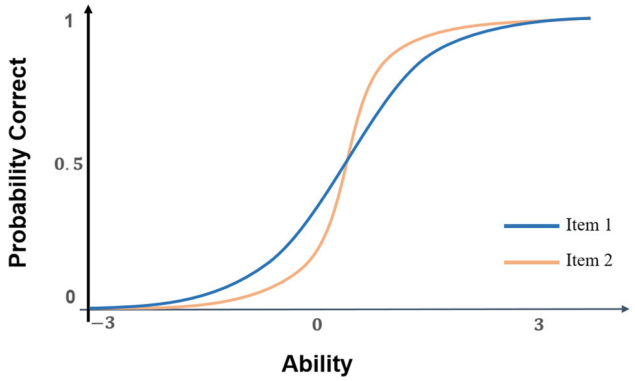

Fig. 8 Parameters in item response theory

and discrimination) to predict the probability of a successful response. Therefore, the discrimination parameter can vary between items and be plotted with different slopes, thus eliminating the explanatory information. A three-parameter IRT model adds a guessing parameter to the two-parameter model. The items answered by guessing indicate that the student's ability is less than the difficulty of the question to which he or she is responding. The three-parameter IRT model can provide explanatory information about guessing behavior.

In IRT, the greater the difficulty of the item is, the greater the corresponding competencies needed by the student. By explicitly defining parameters such as item difficulty and student competencies, the model is transparent in its computational process and has good interpretability. However, the static IRT model assumes that students' abilities remain constant over time, which is particularly unsuitable for long-term knowledge tracing.

Knowledge tracing based on factor analysis model Factor analysis is another critical approach for assessing learners' knowledge mastery. Cen et al. [30] proposed learning factor analysis (LFA), a theory whose primary purpose is to find a more valid cognitive model from students' learning data. Moreover, LFAs inherit the Q matrix used in psychometrics to assess cognition and extend the theory of learning curve analysis, as shown in Fig. 9.

LFA allows researchers to evaluate different representations of knowledge concepts by performing a heuristic search of the cognitive model space. Based on LFA theory, Cen et al. proposed the additive factor model (AFM) [130] and performance factor analysis (PFA) [31]. The AFM is a particular case of PFA and is equivalent when γ_k equals ρ_k. The AFM explains how the difficulty of a student's knowledge points and the number of attempts to solve the problem affect the student's performance, while PFA explains the student's performance in terms of the difficulty of the knowledge points, the number of successes, and the number of failures.

Large-scale factor analysis models have been further developed based on earlier factor analysis models. Using a factor decomposition approach, Vie et al. [32] proposed knowledge tracing machines (KTMs). KTMs use a sparse set of weights for all features to model the learner's correct answer probability. The DASH (difficulty, ability, student interaction history) model is used for memory forgetting and factor analysis [131]. The DAS3H is a newer model that combines IRT and PFA and extends the DASH model by using a time window-based counting function to calculate characteristic factors [131, 132]. With the DAS3H model, the factor analysis method can explain student changes over a continuous time window, thus extending the scope of application of the factor analysis method. Gervet et al. [133] proposed Best-LR based on DAS3H. Unlike DAS3H, Best-LR does not use a window but directly uses the number of successes or failures as an additivity factor. The factors that Best-LR can explain are similar to those that can explain DAS3H. The performance of Best-LR is better than that of DAS3H because Best-LR does not need to calculate window features.

In summary, logistic regression models can explain two main types of features: 1) coded embeddings representing questions and KCs and 2) counting-based features. In Table 4, we compare the factors used by factor analysis models. Counting features summarize the history of students'

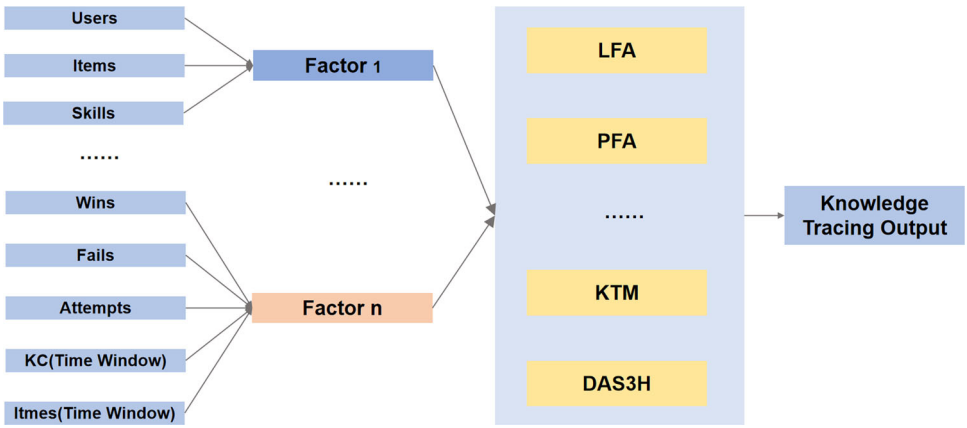

Fig. 9 Learning factor analysis approach

Table 4 Overview of Factors Explained by Factor Analysis Models

Model	Users	Items	Skills	Wins	Fails	Attempts	KC(Time Window)	Items(Time Window)
IRT	✓	✓						
MIRT	✓	✓						
AFM			✓			✓		
PFA			✓	✓	✓			
KTM								
DASH	✓	✓		✓		✓		✓
DAS3H	✓	✓	✓	✓		✓	✓	
Best-LR	✓	✓	✓	✓	✓			

interactions with the system, and counting methods vary among different models, with some even introducing the concept of time windows.

3.2.2 Category 2: model-intrinsic interpretability

Due to the invisibility of the internal structure and operation mechanism of the neural network model, model-intrinsic interpretability can only be realized by adding an interpretable module. Deep neural networks $F(\cdot)$ have large network layers and large parameter spaces. An end-to-end process is used to obtain the output prediction $\hat{y}$ from the input sample x. This process is similar to that of a black box. Therefore, researchers have attempted to embed a simple or an easy-to-interpret module inside the model to achieve model-intrinsic interpretability, thus resembling an interpretable model from the outside to provide explanations for the audience, as shown in Fig. 10. For instance, attention mechanisms can provide explanations by visualizing attention weights. As a result, attention mechanisms have become common intrinsic explainable modules in neural networks and are widely used in computer vision [134, 135], sentiment analysis [136, 137], recommendation systems [138, 139], and other fields.

In xAI, this kind of interpretation method is essentially explained by following strict axioms, rule-based final decisions, granular interpretations of decisions, etc. [119]. It is worth noting that this method can only be used for a specific model, which leads to poor transferability. In the xKT model, as shown in Table 5, researchers have attempted to improve the model interpretability by introducing attention mechanisms, educational psychology, and other theories as interpretable modules. These model-intrinsic interpretability methods aim to make the model more transparent and understandable to stakeholders while maintaining good performance. In the following section, we elaborate on these two methods of model-intrinsic interpretability.

Integrating attention mechanism modules The self-attentive knowledge tracing (SAKT) model [35] identifies concepts related to a given concept from historical student interaction data and predicts learners' performance in the next exercise by considering related exercises in past interactions. This process involves sparse data. To address the problem that the attention layer is too shallow to recognize the complex relationships between exercises and responses, separated self-attentive neural knowledge tracing (SAINT) [154], which is based on transformers and stacked two multihead attention layers on the decoder, was proposed to more effectively model the complex relationships between exercises and answers. The above work proved that the introduction of an attention mechanism into knowledge tracing greatly improves the performance of the model. Furthermore, several researchers have studied the construction of an attention mechanism for the knowledge tracing model as an interpretable module to improve the model's explainability.

For example, Liu et al. [39] proposed explainable exercise-aware knowledge tracing (EKT), which utilizes a novel attention mechanism to deeply capture the focusing information of students on historical exercises. This technique can track students' knowledge states on multiple concepts and visualize knowledge acquisition tracing and student performance prediction to ensure the interpretability of the

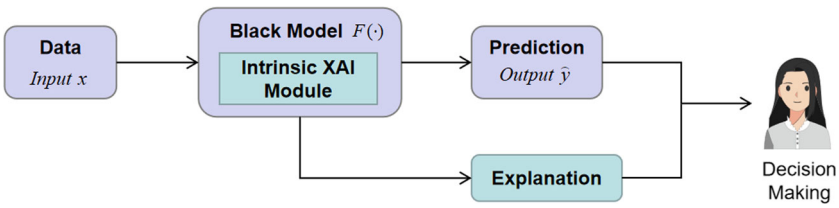

Fig. 10 The framework of the model-intrinsic interpretability method

Table 5 Model-intrinsic interpretability for knowledge tracing

Model-intrinsic module	References
Attention mechanism	[36, 38, 39, 140–143]
Educational psychology and other theories	Item response theory (IRT) [144–147]
	Multidimensional item response theory (MIRT) [44]
	Constructive learning theory [148, 149]
	Learning curve theory and forgetting curve theory [150]
	Finite state automaton (FSA) [151]
	Classical test theory (CTT) [152]
	Monotonicity theory [153]

model. Context-aware attentive knowledge tracing (AKT) [36] combines interpretable components into a novel monotonic attention mechanism and uses the Rasch model to regularize concepts and exercises; this approach has been proven to have excellent interpretability via experiments. Moreover, Zhao et al. [140] proposed a novel personalized knowledge tracing framework with an attention mechanism that uses learner attributes to explain the prediction of mastery. The relation-aware self-attention model for knowledge tracing (RKT) [38] uses interpretable attention weights to help visualize the relationships between interactions and temporal patterns in the human learning process. Similarly, Zhang et al. [141] introduced a new vector to capture additional information and used attention weights to analyze the importance of input features, making it easier for readers to understand the predicted results. Recently, Li et al. [142] regarded the attention mechanism as an effective interpretability module for constructing a new knowledge tracing model, effectively improving the interpretability and predictive ability of the model. Yue et al. [155] , based on ability attributes and an attention mechanism, provided explanations through an inference path. Zu et al. introduced CAKT[156], an innovative model that merges contrastive learning and attention networks to enable interpretable knowledge tracing.

Attention mechanisms, through visualized attention weights, explain aspects of decision-making in models. However, the interpretability of these tools is contingent upon the complexity of the model and the expertise of the interpreter. While elucidating certain decisions in simpler models, attention weights may become less transparent in more complex architectures, where multiple layers and nonlinear interactions obscure the interpretability of the information. Therefore, despite their utility, attention mechanisms should be integrated with complementary techniques for more holistic interpretability in sophisticated deep learning models.

Integrating educational psychology and other theories Item response theory [157] is a modern psychometric theory in which "items" refer to the questions in students' papers and "item responses" refer to students' answers to specific questions. As the parameters of IRT are interpretable, many scholars have combined IRT with deep learning methods, which have powerful feature extraction capabilities for enhancing interpretability. Deep-IRT [144] integrates dynamic key-value memory networks (DKVMNs) with IRT for knowledge training. The DKVMN captures learners' trajectories, inferring their abilities and item difficulties via neural networks, which are subsequently utilized in IRT to predict answer correctness. This model combines the predictive strength of the DKVMN model with the interpretability of IRT, enhancing both the performance and insight into learner and item profiles.

Even though IRT can utilize predefined interpretable parameters to describe students' behavior, students' ability to solve problems is not limited; therefore, one-dimensional IRT parameters cannot be used to effectively explain students' complex behaviors in a real-world scenario. To address this issue, enhanced deep multidimensional item response theory (TC-MIRT) [44] integrates the parameters of a multidimensional item response theory into an improved recurrent neural network, which enables the model to predict students' states and generate interpretable parameters in each specific knowledge field. Inspired by the powerful interpretability of IRTs, many studies have integrated them into model frameworks to improve the model reliability in recent years. For example, knowledge interaction-enhanced knowledge tracing (KIKT) [145] uses the IRT framework to simulate learners' performance and obtains an interpretable relationship between learners' proficiency and project characteristics. Geoffrey Converse et al. [146] improved the model interpretability by transforming the representation of high-dimensional student ability from a deep learning model to an interpretable IRT representation at each time step; leveled attentive knowledge tracing (LANA) [147] uses the interpretable Rasch model to cluster students' ability levels, thus using leveled learning to assign different encoders to different groups of students. Recently, Chen et al. [45] developed QIKT, a question-centric KT model, improved knowledge tracing interpretability using question-centric

representations and an interpretable item response theory layer.

In addition, constructivist learning theory [158] is a classical cognitive theory that emphasizes knowledge mastery differences as the result of knowledge internalization. Based on this theory, the ability boosted knowledge tracing (ABKT) model [148] utilizes continuous matrix factorization to simulate the knowledge internalization process for enhancing the model's interpretability. PKT [149] was designed based on constructivist learning and item response theories and features interpretable and educationally meaningful parameters. The forgetting curve theory [159] indicates that a decrease in students' memory during learning usually reduces their proficiency in knowledge concepts. The learning curve [160] regards knowledge acquisition as a mathematical expression in the process of human learning; that is, students can acquire knowledge after each practice. According to the above two types of pedagogical research, Zhang et al. [150] constructed learning and forgetting factors at the learner level as additional factors to better trace and explain changes in learners' knowledge levels. Moreover, some researchers have attempted to integrate a mathematical compression model into the KT model to enhance the interpretability of the model. For example, Wang et al. [161] utilized finite state automation (FSA) to interpret the hidden state transition of DKT when receiving inputs. In addition, MonaCoBERT [152] uses a classical test theory-based (CTT-based) embedding strategy to consider the difficulty of an exercise to improve the performance and interpretability of the model. Recently, the counterfactual monotonic knowledge tracing (CMKT) [153] method enhances interpretability by integrating counterfactual reasoning with the emonotonicity theory in knowledge acquisition, demonstrating superior performance across real-world datasets.

Incorporating educational psychology theories into models offers interpretability through psychological frameworks and parameters. However, the efficacy of these methods in complex real-world educational contexts is limited and often constrained by the specificity and scope of the underlying theories. While providing insights into controlled scenarios, these approaches may struggle to encapsulate the multifaceted and dynamic nature of learning processes. Consequently, their application necessitates a nuanced and broadened perspective, blending theoretical insights with empirical data analysis to enhance the overall interpretability of the model in diverse educational environments.

3.3 Stage 2: post-hoc interpretable methods

Post-hoc interpretability techniques are applied to pretrained machine learning models, especially those considered "black box" models, to explain their decisions, as shown in Fig. 11. Unlike ante-hoc methods, which are built into the model during development for inherent interpretability, post-hoc methods are employed after model creation, mainly to clarify the predictions [162, 163]. However, some post-hoc methods, such as knowledge distillation [164, 165], rule extraction [166], and activation maximization [167], extend beyond explaining outputs; they also attempt to uncover the model's internal mechanisms. In this paper, we delve into post-hoc interpretable methods for knowledge tracing from two angles: model-specific and model-agnostic methods. Model-specific methods are tailored to particular types of knowledge tracing models, reflecting their unique architectures and learning algorithms. Conversely, model-agnostic methods offer broader applicability, allowing for interpretability across various knowledge tracing models, regardless of their specific designs. This distinction is crucial for developing a comprehensive understanding of how different interpretability methods can be leveraged to demystify the predictions of knowledge tracing models, thereby enhancing their utility and trustworthiness in educational applications.

3.3.1 Category 1: model-specific

Most model-specific methods focus on the interpretability of deep learning, mainly for a certain type of model. It is important to note that model-specific approaches are not necessarily model-based but specific to a class of models. At present, model-specific methods are mainly used to explain the following two categories of models: 1) ensemble-based models [168–170]; and 2) neural networks [171, 172]. Knowledge tracing based on deep learning is a black box that is difficult to understand due to the large parameter space inside the neural network. At present, researchers generally use visualization methods to explain this type of neural network in models.

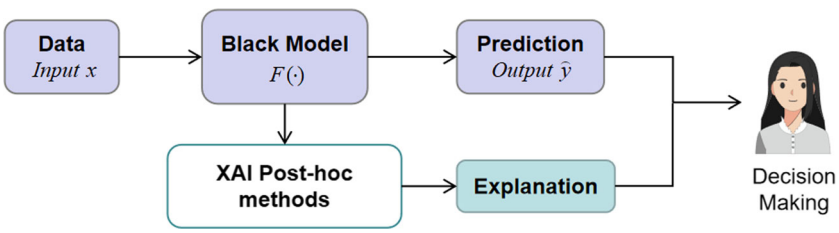

Fig. 11 The framework of the model-intrinsic interpretability method

Layer-wise relevance backpropagation (LRP) The LRP technique was proposed by Bach et al. [59] to calculate the correlation scores of single features in the input data by decomposing the output prediction of the deep neural network. It uses a specially designed set of propagation rules to operate a neural network by backpropagating predictions, where the inputs can be images, videos, or texts [59, 173]. In recurrent neural networks (RNNs), correlations are propagated to hidden states and memory units. Several researchers have applied the LRP method to the knowledge tracing model to enhance its interpretability. For example, Lu et al. [174] proposed solving the interpretability of deep DLKT by adopting the LRP method, which backpropagates the relevance from the output layer of the model to its input layer to explain the RNN-based DLKT model. Wang et al.[175] studied whether the same post hoc interpretable method could be applied to the extensive dataset EdNet and achieved particular effectiveness. However, its effectiveness decreases with increasing learner size and learner practice sequence. Subsequently, Lu et al. [176] used the classical LRP method to interpret the output forecast variables of the DLKT model from the ready-made inputs of the DLKT model, which captured skill-level semantic information. The model output is progressively mapped to the input layer using a backward propagation mechanism, and the interpretation method assigns relevant values to each input sky-answer pair.

The LRP provides valuable insights into neural network-based KT models by mapping predictions back to input features. However, its reliance on correlation for attribution measurements raises concerns about fidelity, as it may be influenced by spurious correlations. This limitation, along with scalability issues in handling large datasets and complex learner sequences, as indicated in [175], questions its applicability and relevance in diverse educational settings. The effectiveness of this method in treating KT thus requires careful consideration of these potential drawbacks.

Causal inference Causal inference is a method for analyzing causal relationships in observational data, attempting to determine whether different treatments (such as different strategies or methods in an experiment) lead to different outcomes [177, 178]. The focus is on distinguishing true causal effects from mere correlations, especially when dealing with confounding variables [179]. Causal inference enhances the transparency and interpretability of AI models by clarifying the "why" behind AI decisions and distinguishing between direct causal relationships and spurious associations. Zhu et al. [47] focused on applying causal inference to the field of knowledge tracing. By adjusting confounding variables within a causal inference framework, they aimed to enhance the prediction accuracy and stability of knowledge tracing models. This approach takes into account key factors, such as confounding variables, to improve the models' ability to predict students' knowledge states accurately. Furthermore, the temporal and causal-enhanced knowledge tracing (TCKT) model [180] integrates causal self-attention with temporal dynamics. This integration not only enhances prediction accuracy and interpretability in educational settings but also effectively mitigates dataset bias by employing causal inference to model the student learning process more accurately.

Causal inference, which is critical in distinguishing between correlation and causation, is invaluable in KT for analyzing the impact of educational interventions. The challenges, as outlined in [177], lie in the need for robust statistical frameworks and the management of confounding variables, which can be daunting in practical educational contexts. Its application in KT requires a careful balance between theoretical robustness and practical feasibility.

Visualization A classic approach to interpret black box models is visualization, which provides intuitive explanations through analysis of the model's training process. Based on their previous work [181], Ding et al. [182] tried to open the "box" of the deep knowledge tracing model. First, they used the larger dataset EdNet to visually analyze the behavior of the DKT model in high-dimensional space, tracked the changes in activation over time, and analyzed the influence of each skill relative to other skills, which solved the problem that interpretation methods were not intuitive.

Visualization techniques provide an intuitive means of interpreting complex KT models. However, they necessitate a high level of expertise in both the model's workings and the data represented. The risk here, particularly with high-dimensional data, is the potential for oversimplification or misinterpretation of the model's dynamics, leading to incorrect conclusions about the learning process.

Ensembling approaches Several researchers have attempted to use ensemble approaches to improve the interpretability of knowledge tracing models on big data. For example, Tirth Shah et al. [183] used a combination of 22 models to predict whether students can answer given questions correctly and discovered that an ensembling approach can improve the prediction performance and interpretability of knowledge tracing tasks. EnKT [184] is based on BKT and DKT and represents student concepts and student questions using learning and performance parameters, respectively, to improve the interpretability of the model.

Ensembling approaches combine multiple models to enhance both predictive accuracy and interpretability in KT. However, the increased complexity of these methods can obscure the contributions of individual models within the ensemble. This complexity poses a significant challenge in KT, where understanding the specific influence of different factors on learning outcomes is crucial (Table 6).

Table 6 A summary of different types of explainable knowledge tracing models

Models	Year	Taxonomy		Interpretable Methods				Other Dimensions
		Transparent Model	Black-box model	Ante-hoc		Post-hoc		
				Self-explanatory	Model-Intrinsic	Model-Specfic	Model-Agnostic	
BKT [28]	1994	✓		✓				
Pardos et al. [121]	2010	✓		✓				
Lee et al. [122]	2012	✓		✓				
BKT-ST [123]	2014	✓		✓				
Wang et al. [124]	2016	✓		✓				
Sun et al. [125]	2022	✓		✓				
IKT [40]	2022	✓						
Affective BKT [126]	2015	✓		✓				
PC-BKT [127]	2014	✓		✓				
LFA [30]	2006	✓		✓				
AFM [130]	2008	✓		✓				
PFA [31]	2009	✓		✓				
KTM [32]	2019	✓		✓				
DASH [131]	2014	✓		✓				
DAS3H [132]	2019	✓		✓				
Best-LR [133]	2020	✓		✓				
EKT [39]	2019		✓		✓			
AKT [36]	2020		✓		✓			
RKT [38]	2020		✓		✓			
Zhao et al. [140]	2020		✓		✓			
Zhang et al. [141]	2021		✓		✓			
Li et al. [142]	2022		✓		✓			
Yue et al. [155]	2023		✓		✓			
CAKT [156]	2023		✓		✓			
MonaCoBERT [152]	2022		✓		✓			
CMKT [153]	2023		✓		✓			
deep-IRT [144]	2019		✓		✓			
TC-MIRT [44]	2011		✓		✓			
KIKT [145]	2020		✓		✓			
Geoffrey Converse et al. [146]	2021		✓		✓			
LANA [147]	2021		✓		✓			
QIKT [45]	2023		✓		✓			
ABKT [148]	2022		✓		✓			
PKT[149]	2023		✓		✓			
Zhang et al. [150]	2021		✓		✓			
Wang et al. [161]	2023		✓		✓			
EAKT [185]	2022		✓		✓			
Zhu et al. [151]	2022		✓		✓			
Ding et al. [181]	2019		✓			✓		
Ding et al. [182]	2021		✓			✓		
Zhu et al.[47]	2023		✓			✓		
TCKT [180]	2024		✓			✓		

Table 6 continued

Models	Year	Taxonomy		Interpretable Methods				
		Transparent Model	Black-box model	Ante-hoc		Post-hoc		Other Dimensions
				Self-explanatory	Model-Intrinsic	Model-Specfic	Model-Agnostic	
Tirth Shah et al. [183]	2020		✓			✓		
EnKT [184]	2022		✓			✓		
Lu et al. [174]	2020		✓			✓	✓	
Valero et al. [186]	2023		✓				✓	
Wang et al. [175]	2021		✓			✓	✓	
Lu et al. [176]	2022		✓			✓	✓	
Varun Mandalapu et al. [187]	2021		✓				✓	
Wang et al. [188]	2022		✓				✓	
GCE [46]	2023		✓				✓	
SPDKT [189]	2021		✓					✓
CoKT [190]	2021		✓					✓
Lee et al. [191]	2019		✓					✓
HGKT [43]	2022		✓					✓
GKT [37]	2019		✓					✓
SKT [192]	2020		✓					✓
Zhao et al. [193]	2022		✓					✓
JKT [194]	2021		✓					✓

3.3.2 Category 2: model-agnostic

Model-agnostic techniques separate explanations from model outputs and are applicable to any machine learning model [162]. It only acts on the input and output of the neural network, providing explanations by perturbing the input or simplifying the model. Because model-agnostic techniques are not limited to a specific model, most researchers currently prefer model-agnostic approaches over model-specific approaches.

Local interpretable model-agnostic explanations (LIME) The LIME was proposed by Ribeiro et al. [64]. This method trains local surrogate models to explain a single prediction a global black-box model gives. LIME partially replaces complex models with simpler models to provide local explanations. Specifically, since the perturbed data will affect the model's output, LIME trains a local interpretable model to learn the mapping relationship between the perturbed data and the model's output and uses the similarity between the perturbed input and the original input as the weight. Finally, the essential K features are selected from the local interpretable model for interpretation. This approach can provide a very effective local approximation to the black box model. In xKT, Varun Mandalapu et al. [187] utilized LIME to understand the impact of various features on best-performing model predictions.

LIME provides microlevel insights into specific predictions of KT models by training local interpretable surrogate models. Its major strength lies in revealing the influence of particular features in specific instances. However, LIME's focus on local explanations may not capture the model's global behavior, particularly in KT, where diverse learning paths can significantly influence model decisions. Additionally, the dependence of LIME on perturbation strategies and the choice of local models might affect the consistency and accuracy of its interpretations.

Deep shapley additive explanations (Deep SHAP) By combining DeepLIFT [195] with Shapley values [196], Lundberg and Lee [65] proposed a fast method to approximate Shapley values for CNNs called Deep SHAP. Deep SHAP decomposes the prediction of the deep learning model into the sum of feature contributions through backpropagation and obtains the reference specific contribution of each feature to the prediction through the backpropagation prediction difference. Several studies [197] have concentrated on using the DKT model to predict test scores based on skill mastery and then assessed the influence of each skill on the predicted score using SHAP analysis. Inspired by this idea, Valero-Lea et al. [186] aimed to explain learners' skill mastery by analyzing past interactions using a SHAP-like method to determine the importance of these interactions. Wang et al. [188] proposed a four-step procedure to interpret the DLKT model and

obtained effective explicable results. The four-step procedure is as follows: 1) Given a sample x to be interpreted, reference samples are selected, and predictions are made about the last questions; 2) the difference between each reference sample and sample x is calculated; 3) the prediction is then backpropagated from the output layer to the input layer to calculate the reference-specific feature contribution between each reference sample and the interpreted sample; and 4) the contribution of each question-answer in sample x to the prediction of the DLKT model is obtained.

Deep SHAP, which combines DeepLIFT with Shapley values, elucidates feature contributions to predictions in deep learning models. KT helps us understand the relative importance of various features, such as prior performance or interaction frequencies. While effective at revealing individual feature impacts, Deep SHAP may struggle with high-dimensional feature spaces and overlook complex inter-feature interactions, which are crucial in KT with diverse learning trajectories.

Causal explanation(CE) Li et al. [46] addressed the explainability issue of DLKT models by proposing a genetic causal explainer (GCE) based on genetic algorithms (GAs). The GCE established a causal framework and a specialized coding system, effectively resolving the issues of spurious correlations caused by reliance on gradients or attention scores, thereby enhancing the accuracy and readability of explanations. Additionally, the GCE is a post hoc explanation method that can be applied to various DLKT models without interfering with model training, offering a flexible and effective means of explanation.

GCE based on genetic algorithms addresses explainability in DLKT models by establishing a causal framework. While GCE offers a novel approach to understanding deep causal relationships in KT, establishing causal connections requires precise data modeling and hypothesis validation. The complexity and computational demands of GCE, particularly for large datasets, pose significant challenges.

3.4 Stage 3: other dimensions

Beyond the mainstream ante-hoc and post-hoc methods, there are interpretatable approaches specific to the knowledge tracing domain, yet underrepresented in current xAI literature. These approaches exploit unique data features in KT, such as interconnected relationships among questions, concepts, or users. Upcoming sections will provide a detailed exploration of these approaches and their role in enhancing interpretability.

3.4.1 Category 1: output format

Visual explanations To verify the interpretability of the proposed model, many researchers use radar charts or heat maps to provide readers with visual explanations. In the following, several representative works will be introduced. Self-paced deep knowledge tracing (SPDKT) [189] reflects the difficulty of the problem by assigning different weights to the problem, visualizing the difficulty of the problem, and improving the interpretability of the model. Similarly, collaborative embedding method for Knowledge Tracing (CoKT) [190] provided an interpretable question embedding by visualizing the distance between question embeddings that share the same concepts and those that do not. Lee et al. [191] proposed a knowledge query network (KQN) model, which uses the dot product between the knowledge state vector and skill vector to define knowledge interaction and uses a neural network to encode students' responses and skills into vectors of the same dimension. Moreover, the KQN can query students' knowledge of different skills and subsequently enhance the interpretability of the model by visualizing the interaction between two types of vectors.

Visual explanations enhance surface-level interpretability in KT but often fail to delve into the internal mechanisms of models. The explanatory power of these models heavily relies on the quality of the data representation; inaccurate representations may lead to misleading interpretations. Moreover, complex visual outputs, such as relationships in high-dimensional spaces, can be challenging for general users to understand, limiting their practical effectiveness. Therefore, these methods require further refinement and development to provide more in-depth and thorough explanations.

3.4.2 Category 2: data attribute

Interaction between exercises and concepts. Several researchers have shown that the relationship between exercises and concepts may be structured into a graph by analyzing learners' learning data. For example, one exercise may involve multiple concepts, and one concept may also correspond to multiple exercises. In addition, two relationships exist among concepts, prerequisite and similarity, as shown in Fig. 12. The prerequisite is that the mastery of concept y requires the mastery of concept x, for example, by adding before multiplying; the similarity relationship means that knowledge y and knowledge x belong to the same category, such as addition. Some studies show that incorporating this graph structure into the knowledge tracing model as

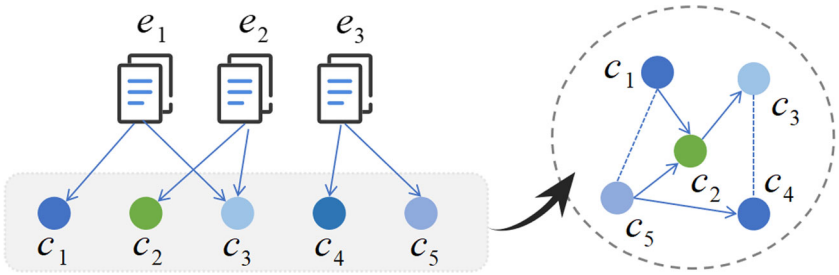

Fig. 12 (a) Relationships between the exercises and the concepts. (b) Relationships between concepts: the solid line represents the prerequisite relationship, and the dashed line represents the similarity relationship

a relation induction deviation can enhance interpretability in the prediction process. Generally, there are exercises-to-exercises, concepts-to-concepts, exercises-to-concepts, and other potential relationships in the graph structure constructed from learner learning data. Next, this subsection describes in detail the works that have been done to increase the interpretability of the model by using the underlying graph structure in the data.

At the level of exercise-to-exercise relationships, hierarchical graph knowledge tracing (HGKT) [43] constructs a hierarchical exercise graph according to the latent hierarchical relationships (direct relationships and indirect relationships) between exercises and introduces a problem schema to explore the dependencies of exercise learning, which enhances the interpretability of knowledge tracing. On the level of concept-to-concept relationships, the graph-based knowledge tracing model (GKT) [37] synchronously learns the latent relationships between concepts in the prediction process. The model updates the knowledge state related to the current exercises each time, thus realizing interpretability at the concept level. Educational entities emphasize the significance of knowledge structure. Structure-based knowledge tracing (SKT) [192] utilizes two different propagation models to track the influence of prerequisites or similarity relationships between concepts and has been used in many experiments to prove interpretability. This work visually demonstrated the interpretability of the models in the manner described in 3.1, such as through heatmaps and radar maps. On the level of exercise-to-concept relationships, Zhao et al. [193] used a graph attention network to learn the underlying graph structure between concepts in the answer record and input information from the model containing the relationship information between the exercises and the concept, which enhances the interpretability of the model. To dig deeper into the relationship between exercise-to-exercise and concept-to-concept, a joint graph convolutional network-based deep knowledge tracing (JKT) [194] framework was used to model the multidimensional relationships of the above two factors into a graph and fuse them with "exercise-to-concept" relationships. The model connects exercises under cross-concepts and helps capture high-level semantic information, which increases the interpretability of the model.

Graph-based approaches in knowledge tracing offer intricate insights into the relationships among exercises, concepts, and hierarchical interdependencies. While these methods enhance interpretability by mapping complex educational theories onto graph structures, they also present challenges in terms of complexity and accessibility. Their reliance on sophisticated graph representations and computational models may limit usability for nontechnical users such as teachers and students, hindering their practical application in diverse educational settings. Moreover, the assumptions inherent in these graph-based models about learning processes and relationships might not fully align with the dynamic and varied nature of real-world learning, raising questions about their generalizability and effectiveness.

3.5 Explainable knowledge tracing: application

In this section, we explore the practical application of xKT in educational settings. Our focus is on its role in generating diagnostic reports, personalized learning, resource recommendations, and knowledge structure discovery. This section examines how xKT algorithms are used to track and predict learners' knowledge states, enabling the creation of dynamic, personalized educational pathways. We discuss the balance between algorithmic complexity and the need for clear, interpretable results in educational settings.

Knowledge tracing in diagnostic reports and visualization In the realm of educational technology, knowledge tracing primarily manifests in the generation and visualization of learning diagnostic reports. Algorithms such as BKT and DLKT have been pivotal in this regard. BKT utilizes probabilistic modeling to continually update a student's knowledge state, adjusting the likelihood of concept mastery after each educational interaction. DLKT, leveraging neural network architectures, excels in capturing complex learning patterns, offering nuanced insights into student performance. Despite its interpretability, BKT sometimes struggles with complex learning scenarios, whereas DLKT, though proficient at deciphering intricate behaviors, compromises clarity for the sake of complexity.

To enhance the interpretability of diagnostic reports, models such as KSGKT [198] integrate knowledge structures

with graph representations, employing attention mechanisms to accurately predict learner performance. HGKT [43] uses a hierarchical graph neural network to analyze learner interactions, enabling detailed categorization of exercises for a deeper understanding of learner knowledge and problem-solving skills. Both models aim to provide granular diagnostic reports to support personalized learning paths. Additionally, visual explanations, such as intuitive graphs and heatmaps [189], significantly augment the interpretability of KT algorithm outputs, transforming complex data into actionable insights for tailored educational strategies.

Knowledge tracing in personalized learning and resource recommendation The field of personalized learning and resource recommendation has greatly benefited from advancements in knowledge tracing algorithms. These algorithms are adept at tailoring learning pathways and recommending appropriate learning resources based on a student's current state of knowledge and learning preferences [199]. The introduction of deep learning technologies, such as DKT, has further enhanced the precision and personalization of these recommendations. However, one limitation is the often reduced explainability of these sophisticated models, which can make it challenging for educators to understand the rationale behind specific recommendations.

Several studies have made notable efforts to enhance the accuracy and interpretability of personalized recommendations. The integration of concept tags with the DKVMN model, as seen in [200], marks a significant step in improving exercise recommendations by accurately tracing students' knowledge states. Building upon this, Zhao et al. [201] introduce attention mechanisms and learner attributes to refine mastery predictions, yielding more personalized and interpretable activity recommendations. Adopting a dynamic approach, Cai et al. [202] combined reinforcement learning with knowledge tracing, adapting learning paths in real time to align with the learner's evolving understanding. Similarly, the ER-KTCP [203] innovatively merges knowledge state tracking with concept prerequisites for exercise selection, demonstrating marked improvements in student performance. Furthermore, Wang et al. [204] focused on small private online courses (SPOCs), employing learning behavior dashboards and a modified DKVMN model to emphasize student engagement and concept mastery in a specific educational setting. Collectively, these studies contribute to a more nuanced understanding of data-driven models in educational technology, paving the way for adaptive, personalized learning experiences.

Knowledge tracing in knowledge structure discovery In knowledge structure discovery, knowledge tracing algorithms clarify the relationships between problems and concepts. They analyze students' learning behaviors and performances to identify connections, aiding educators in understanding the foundational concepts for advanced problems. For example, an algorithm can show that mastering basic mathematical skills is essential before complex concepts are grasped. These insights are vital for creating effective teaching strategies and curricula, allowing educators to logically sequence lessons and ensuring that students master fundamentals before progressing to advanced topics.

Advancements such as HGKT [43] and GKT [37] have significantly improved the interpretability of learning models. HGKT reveals complex interdependencies between exercises, enhancing the interpretability of exercise-related learning progress. Moreover, GKT delves into the latent relationships between concepts, providing clear insights at the concept level. Complementary approaches such as SKT [192] and graph attention networks further augment this clarity by tracing relationships (both pre-requisite and similarity) between concepts. These methods, along with the JKT [194], collectively enhance the overall interpretability of knowledge structures, making the connections within the learning process more understandable and accessible.

3.6 Discussion

In this comprehensive study, we have evaluated various xKT models, emphasizing interpretability, accuracy, computational efficiency, and applicability in real-world scenarios. This analysis elucidates the distinct characteristics and constraints of different xKT models, pivotal for enhancing educational technology tools.

Interpretability Within KT models, the spectrum of interpretability ranges from transparent, ante hoc methods to intricate, post hoc techniques. Transparent models such as BKT offer straightforward interpretability due to their simple probabilistic frameworks, which are beneficial in scenarios demanding clarity [28]. In contrast, post hoc methods such as the LRP [176] and LIME [187] methods provide insights into more complex models suitable for detailed analytical requirements. However, these methods can be challenging for nontechnical users to interpret due to their complexity.

Accuracy DLKT models, such as those employing neural networks, exhibit high accuracy in modeling complex student interactions but require substantial tuning and expertise [36, 44]. These models, while powerful, can be prone to overfitting and opaque, making their predictions difficult to interpret. On the other hand, simpler models such as the BKT, despite being more interpretable, may not capture complex learning behaviors effectively, thus limiting their accuracy in more nuanced educational settings.

Computational efficiency The computational demands of KT models vary widely. Simpler structures such as those in BKT models are computationally efficient and align well with resource-constrained environments. Advanced models, particularly ensembling approaches [184], demand significant computational power, making them suitable for well-resourced scenarios but impractical for more constrained scenarios.

Real-world applicability The applicability of a KT model heavily depends on the educational context. Transparent models [126] are ideal in settings where quick and clear feedback is essential. In contrast, environments that require deep insights into intricate learning patterns require more sophisticated models. However, these advanced models, while offering detailed analysis capabilities, often involve the downside of higher computational requirements and potential overfitting issues [148].

In conclusion, the choice of a xKT model requires a delicate balance between interpretability, educational environment complexity, accuracy needs, and computational resource availability. Future developments in xKT should aim to integrate these factors, pursuing models that provide both clarity and depth in understanding diverse learning patterns adaptable across various educational contexts.

4 Explainable knowledge tracing: evaluation method

Although many researchers have shown interest in explainable knowledge tracing, most model evaluation metrics focus on evaluating the model performance, while existing research evaluating interpretability is scarce [22]. The process of imparting knowledge from teachers to learners needs to be highly explanatory and understandable. As an intelligent auxiliary tool in the teaching process, an AI model cannot be trusted by users only by its high accuracy [205, 206]. Based on the above, it is also worth evaluating these existing explainable knowledge tracing models.

The following sections begin with a brief introduction to the common evaluation methods in xAI and then explore how we can develop a standardized and reasonable interpretable evaluation system for educational models on knowledge training tasks. The goal is to improve the user's understanding and trust in an education model and realize the wide application of intelligent education products in education.

4.1 A case study: comparison of interpretable methods for knowledge tracing

As mentioned in Section 3, transparent models such as Bayesian knowledge tracing are explained by their internal parameters, while deep learning-based knowledge tracing requires additional specific interpretation methods. "Are all models in all defined-to-be-interpretable model classes equally interpretable [207]?" To compare the unified interpretation results for the same model and dataset, we use three common post hoc xAI interpretable methods, LRP, LIME, and SHAP, to explain the deep knowledge tracing in ASSISTment2009 [208], as shown in Fig. 13. For this dataset, we selected two interactive sequences, each with a length of 25, from two different learners to be explained. Using the aforementioned methods, we calculated and visualized the interpretable features of the interactive sequences for each respondent. This comparison provides insights into the effectiveness and interpretability of each method, helping us to better understand how the model can make predictions for deep knowledge tracing.

In Fig. 13, Row 1 indicates the question ID, Row 2 indicates the correctness of the corresponding question, where 1 represents the correct answer, and 0 represents an incorrect answer, Row 3 is the model prediction, which indicates the probability of correct answers, and Rows 4-6 represent the eigenvalues calculated by LRP, LIME, and SHAP, respectively. The model correctly predicts position 24 for student 1, and the LIME method better handles the phenomenon of incorrect prediction at positions 14, 18, and 21. Similarly, when the model predicts the problem at position 24 with a low probability (0.56) for Student 2, all three methods can assign different eigenvalues to the irrelevant problem at position 23. Through the above two examples, it is found that the LRP method tends to transition the feature contribution of continuous problems smoothly, the LIME method distributes the feature contribution more evenly, and the SHAP method is more focused on sharing the features of current problems.

In the deletion experiment, as shown in Fig. 14, three lines representing different deletion strategies-random deletion (green), deletion based on the LIME method-for which the feature importance was calculated (yellow) and deletion based on the LRP method-for which the relevance was calculated (blue) are observed. The x-axis represents the deletion of 0 to 10 pairs of the original input data, and the y-axis represents the LSTM model's predicted values. Before x=4, the LRP line is below the green line, the green line is below the LIME line, and the LIME line shows the steepest decrease. The difference between the green and LRP lines is approximately 0.5, and the difference between the green and LIME lines is approximately 1-2. After x=4, the LIME line continues to decline rapidly, reaching a plateau after x=6. The descent of the LRP line accelerates after x=4, decreases below the green line by approximately 0.5, and plateaus after x=8. The green line remains above the other lines, exhibiting fluctuations after x=7 but converging with the other lines at approximately x=10.

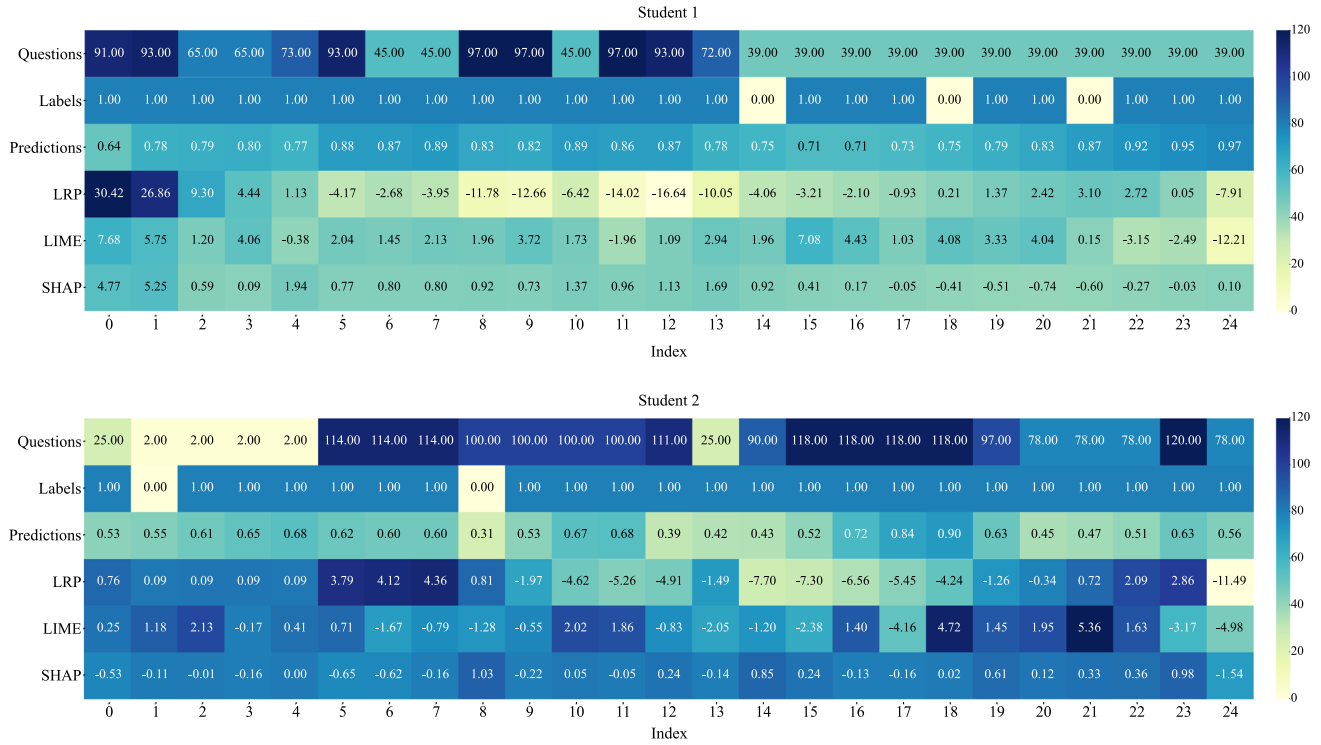

Fig. 13 Comparison of post-hoc interpretability methods on deep knowledge tracing

In conclusion, different deletion strategies significantly impact the model interpretability and robustness. The LIME method shows high sensitivity to small deletions but plateaus for larger counts. The LRP method performs better at capturing model changes with larger deletions. Random deletion shows relative robustness but may not capture complex model changes. The results emphasize the importance of choosing appropriate feature deletion strategies for interpreting model behavior. Further research could explore combining different interpretation methods for a comprehensive understanding of model behavior.

All three approaches explain deep knowledge tracing models, but which approach is closest to reality? The interpretability of models has become an urgent problem. The following sections briefly introduce the common evaluation methods used in KT and xAI and then explore how we can develop a standardized and reasonable interpretable evaluation system for educational models on knowledge tracing tasks. The goal is to improve the user's understanding and trust in the education model and realize the wide application of intelligent education products in education.

4.2 Common evaluation metrics for knowledge tracing

The accuracy (ACC) and area under the curve (AUC) are the two main metrics commonly used to evaluate the performance of knowledge tracing models. The accuracy represents the proportion of correct prediction results among all the results. The AUC is the area under the ROC curve, and a lower coordinate axis indicates that the probability of a positive prediction is greater than that of a negative prediction. Therefore, the higher the AUC value is, the better the model being evaluated can achieve classification. However, it's important to note that while these metrics are instrumental in evaluating the model's predictive accuracy, they do not contribute to the evaluation of the model's interpretability.

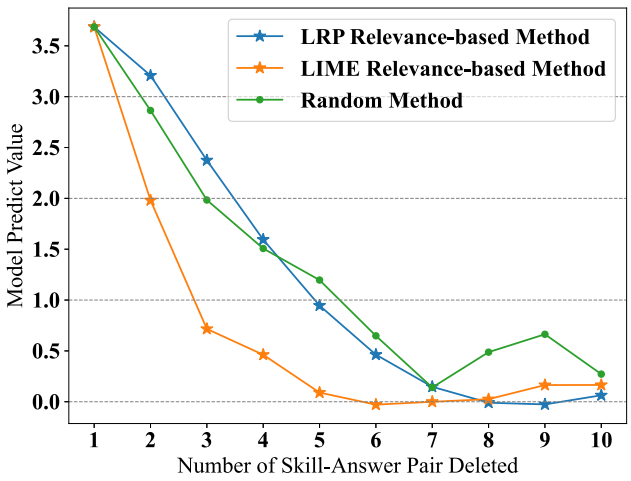

Fig. 14 Deletion experiment for LRP and LIME methods

Table 7 Summary of Evaluation Metrics for xKT

User Category	Stakeholders	Evaluation Metrics	References
Professional Users	Developers	Objective metric(stability, fidelity, sensibility, ect.)	[80, 81]
		Human-machine interaction	[209, 210]
		Multi-disciplines integration	[211–213]
Non-professional Users	Educators, Learners	Subjective metrics (interviews, questionnaires, scale analyses)	[214, 215]
		Social experiment	[216]

4.3 Evaluation metrics for explainable knowledge tracing

Inspired by the classification of xAI evaluation methods [80] in the previous section, in this section, the evaluation methods for explainable knowledge tracing are presented, summarised in Table 7. As discussed above, "humans" are the main component of the whole educational loop. Therefore, to evaluate xKT, in this paper, models are evaluated from a subjective perspective, that is, focusing on evaluating xAI systems concerning the target audience and specific interpretability goals. Therefore, we divide the educational subjects into three categories in the task of knowledge tracing: KT educators, learners, and developers. Considering the different interpretability goals of the three types of stakeholders in the educational process, in this section, we elaborate on the interpretability of each category of subjects in the follow-up.

4.3.1 From the perspective of professional users

From the perspective of KT developers, according to the xAI user classification standard in the previous section [80], developers are considered professional users. In general, developers are usually AI scientists and data engineers who design machine learning models and interpretable techniques for xAI systems. Compared with nonprofessional users, professional users are clearer about the operation mechanism of the model. How to provide reasonable and scientific explanations according to the different needs and abilities of different end users is a problem that professional users need to consider.

Adopt objective metrics Developers should leverage quantitative evaluation methodologies within the realm of xAI for evaluating xKT processes. These methodologies include the deployment of standardized objective metrics, such as stability [80], fidelity [80], and sensibility [81], to appraise the congruence in xKT's interpretation of proximate or analogous data instances and the accuracy in approximating black-box model predictions. However, it is pertinent to acknowledge that the technical nature of these methods may render them less accessible to laypersons, potentially impeding their efficacy in broader evaluative contexts.

Human-machine interaction We advocate for the integration of advanced human-machine interaction [209, 210] technologies to facilitate a dynamic interaction loop between users and AI models. In this loop, the AI system should adapt its outputs based on feedback from the user, who may function as a teacher or learner, utilizing actions such as modifying data labels or assessing the validity of decisions made by the model. This approach fosters a more immersive evaluation of the model interpretability, striving to harmonize human intuition with artificial intelligence insights for more effective assessment. Essential to this process is the active engagement between developers and end users, as user-centric feedback is critical for the iterative refinement of explanatory mechanisms offered by the model.

Multi-disciplines integration Finally, a collaborative approach with specialists in cognitive psychology or educational theory [211] is recommended. Cognitive psychology principles, particularly mental model theories [212], can be instrumental in conceptualizing a framework for understanding human-machine interaction and behavior. This understanding is crucial for an effective evaluation of AI interpretability. Furthermore, incorporating insights from educational measurement theories [213] enables developers to ascertain whether AI model predictions align with established cognitive learning patterns, thereby facilitating a more robust evaluation of the model interpretability.

4.3.2 From the Perspective of Non-professional Users

From the perspective of evaluating the interpretability of a model with teachers and learners as the subject, according to the xAI user classification standard in the previous section, teachers and learners are considered nonprofessional users, who do not understand the internal structure and operation mechanism of the model and consider an AI model a "black box". Therefore, how to improve the model transparency and the user's reliance is a problem that requires increased attention.

Adopt subjective metrics Researchers can choose the subjective evaluation method in xAI. For example, interviews, questionnaires, and scale analyzes are used to evaluate the validity of explanations provided models and the satisfaction

and trust of users so that nonprofessional users can understand and trust model decisions [214, 215]. Before this process, we should improve the AI literacy of teachers and learners in advance so that they can provide reasonable and scientific feedback.

Social experiment It is also possible to attempt to design a reasonable social experiment of intelligent education to evaluate the interpretability of a model [216]. In brief, researchers can recruit some stakeholders involved in knowledge tracing tasks and conduct small social experiments to empirically study the interpretability of a model. For example, participants can evaluate the interpretability of a model by reviewing the interpretation results.

Generally, in this section, we first analyze the limitations of the current evaluation metrics for knowledge tracing models. Next, inspired by the evaluation methods of xAI, we review the evaluation methods of the xKT model. Specifically, the evaluation methods proposed in this paper are human-centered and can be subdivided into methods for professional users (developers) and nonprofessional users (educators and learners). According to the characteristics of the two types of target users, corresponding evaluation methods have been proposed. The goal of this section is to provide some ideas for evaluating explainable knowledge tracing.

5 Explainable knowledge tracing : future directions

xKT is emerging at the crossroads of xAI and educational analysis, driven by the need for effective, comprehensible, and ethically sound models in education. We will explore four key future directions that have been identified for their potential to substantially enhance xKT: Balancing model performance with interpretability to create sophisticated yet transparent algorithms, making advanced models accessible through user-friendly explainable methods, integrating causal inference to shift from prediction to understanding learning dynamics, and addressing ethical and privacy concerns in the data-driven educational era. These areas collectively aim to enhance xKT, aligning technical prowess with evolving educational and ethical demands.

The trade-off between model performance and interpretability in knowledge tracing In knowledge tracing, the critical future challenge is to balance model accuracy with interpretability. Achieving this balance requires creating algorithms that are both precise in prediction and intuitive in understanding. Future research is expected to focus on refining deep learning architectures to simplify structures and integrate advanced attention mechanisms, aiming to balance high performance with better interpretability [152]. Additionally, a growing trend is the integration of post hoc interpretability tools such as LIME [187] and SHAP [186, 188, 197], which offer clearer explanations for complex model decisions and uncover the underlying drivers of behavior. Moreover, complex, accurate models are likely to be blended with simpler, more interpretable models using advanced ensemble learning techniques [183, 184]. This blend aims to improve the prediction accuracy while maintaining decision transparency, promoting enhanced educational quality and personalized learning in knowledge tracing applications.

User-friendly explainable methods As knowledge tracing technology evolves, user-friendly explainable methods have emerged as a core issue [90]. Future research should focus on designing explainability mechanisms that are transparent not only to data scientists but also accessible and friendly to educators and learners. This level of explainability requires models to produce predictions that are easy to understand, in addition to clear logic and reasoning processes. Leveraging advanced natural language processing technology, models can generate detailed and comprehensible explanations, clearly articulating the logic behind their predictions. Furthermore, dynamic and interactive visualization technologies [91] play a crucial role in intuitively presenting learners' knowledge states and learning paths, significantly enhancing educators' and learners' understanding and acceptance of model feedback. Additionally, when designing these models, the intuitiveness of the user interface should be considered [92, 93], enabling nonexperts to easily interpret and utilize the model outputs.

Integrating causal inference into knowledge tracing models Integrating causal inference into knowledge tracing models is a vital direction for future research. This approach aims to uncover the actual causal relationships within learning processes, moving beyond the limitations of correlation-based analysis prevalent in many machine learning models [46]. By applying techniques such as counterfactual reasoning [217], researchers can explore various hypothetical scenarios, and such alternative learning strategies might lead to diverse learning outcomes. This method enables a more profound understanding of the direct impact of specific learning activities on educational results. Such in-depth causal analysis not only improves the scientific rigor and accuracy of knowledge tracing models but also offers valuable insights for the development of effective and personalized educational interventions [46, 47]. Consequently, knowledge tracing technology has advanced from simply predicting outcomes to providing actionable insights for enhancing educational practices.

Ethical and privacy considerations in model explainability research In the realm of knowledge tracing, as efforts intensify to enhance model explainability, parallel emphasis must be placed on ensuring ethical and privacy considerations [218]. Research should aim to design explainable models that not only provide transparent and understandable predictions but also rigorously protect user data and maintain ethical integrity. This research will involve developing techniques that balance the need for clarity in how models process and interpret personal data with robust measures to secure data privacy. Approaches such as differential privacy [219, 220], which anonymizes data to prevent the identification of individuals, can be integrated with explainable Al frameworks. These approaches ensure that while models remain interpretable and that their decisions are transparent to users, they also adhere to strict privacy and ethical guidelines. Such research would necessitate a nuanced approach where explainability does not compromise privacy and ethical standards guide the transparency of the models.

6 Conclusion

This survey is the first to provide a comprehensive survey of explainable knowledge about multiple dimensions, including concepts, methods, and evaluations. Specifically, according to the xAI classification criteria for the complexity of explainable object models, we classify the related models of explainable knowledge training into two categories: 1) transparent models and 2) black box models. Then, representative explainable methods are reviewed in three stages: ante-hoc stage, post-hoc stage, and other dimensions. Additionally, includes an investigation into the applications of explainable knowledge tracing. Furthermore, to fill the gap in the evaluation methods of explainable knowledge tracing, we consider evaluation methods from the perspective of education stakeholders. Finally, future research directions for explainable knowledge tracing are explored. The field of explainable knowledge tracing is booming, and we aim to draw researchers' attention to the interpretability of algorithms, improve the transparency and reliability of algorithms, and provide a foundation and insight for researchers who are interested in interpretable knowledge tracing.

Author Contributions Yanhong Bai: Conception and design of the survey and writing the original draft. Jiabao Zhao: Conception and design of the survey, writing, and review. Tingjiang Wei: Writing and performing experiments. Qing Cai: Supervision and review. Liang He: Project administration, Resources.

Availability of data and access All data generated or analysed during this study are included in this published article [208].

Declarations

Competing interests This work was supported by the National Natural Science Foundation of China (Grant number [62207013] and [6210020445]).

References

1. Ouyang F, Zheng L, Jiao P (2022) Artificial intelligence in online higher education: A systematic review of empirical research from 2011 to 2020. Educ Inf Technol 27(6):7893–7925
2. Mousavinasab E, Zarifsanaiey NR, Niakan Kalhori S, Rakhshan M, Keikha L, Ghazi Saeedi M (2021) Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods. Interact Learn Environ 29(1):142–163
3. Xu J, Huang X, Xiao T, Lv P (2023) Improving knowledge tracing via a heterogeneous information network enhanced by student interactions. Expert Syst Appl 120853
4. Xu X, Xie J, Wang H, Lin M (2022) Online education satisfaction assessment based on cloud model and fuzzy topsis. Appl Intell 52(12):13659–13674
5. Xiao X, Jin B, Zhang C (2023) Personalized paper recommendation for postgraduates using multi-semantic path fusion. Appl Intell 53(8):9634–9649
6. Ahmad HK, Qi C, Wu Z, Muhammad BA (2023) Abine-crs: course recommender system in online education using attributed bipartite network embedding. Appl Intell 53(4):4665–4684
7. Chrysafiadi K, Papadimitriou S, Virvou M (2022) Cognitive-based adaptive scenarios in educational games using fuzzy reasoning. Knowl-Based Syst 250:109111
8. Hooshyar D, Huang Y-M, Yang Y (2022) Gamedkt: Deep knowledge tracing in educational games. Expert Syst Appl 196:116670
9. Arrieta AB, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, García S, Gil-López S, Molina D, Benjamins R et al (2020) Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. Inform Fusion 58:82–115
10. Gunning D, Aha D (2019) Darpa's explainable artificial intelligence (xai) program. AI Mag 40(2):44–58
11. Abdelrahman G, Wang Q, Nunes B (2023) Knowledge tracing: A survey. ACM Comput Surv 55(11):1–37
12. Diedrick M, Clements-Nolle K, Anderson M, Yang W (2023) Adverse childhood experiences and clustering of high-risk behaviors among high school students: a cross-sectional study. Public Health 221:39–45
13. Long Y, Aleven V (2017) Enhancing learning outcomes through self-regulated learning support with an open learner model. User Model User-adap Inter 27(1):55–88
14. Castiglioni I, Rundo L, Codari M, Di Leo G, Salvatore C, Interlenghi M, Gallivanone F, Cozzi A, D'Amico NC, Sardanelli F (2021) Ai applications to medical images: From machine learning to deep learning. Physica Medica 83:9–24
15. Clancey WJ, Hoffman RR (2021) Methods and standards for research on explainable artificial intelligence: lessons from intelligent tutoring systems. Appl AI Lett 2(4):53
16. Barria-Pineda J, Akhuseyinoglu K, Želem-Ćelap S, Brusilovsky P, Milicevic AK, Ivanovic M (2021) Explainable recommendations

in a personalized programming practice system. In: International conference on artificial intelligence in education, pp 64–76. Springer
17. Jang Y, Choi S, Jung H, Kim H (2022) Practical early prediction of students' performance using machine learning and explainable ai. Educ Inf Technol 1–35
18. Melo E, Silva I, Costa DG, Viegas CM, Barros TM (2022) On the use of explainable artificial intelligence to evaluate school dropout. Educ Sci 12(12):845
19. Hur P, Lee H, Bhat S, Bosch N (2022) Using machine learning explainability methods to personalize interventions for students. Int Educ Data Mining Soc
20. Liu Q, Shen S, Huang Z, Chen E, Zheng Y (2021) A survey of knowledge tracing
21. Liu T (2022) Knowledge tracing: A bibliometric analysis. Comput Educ: Artif Intell 100090
22. Song X, Li J, Cai T, Yang S, Yang T, Liu C (2022) A survey on deep learning based knowledge tracing. Knowl-Based Syst 258:110036
23. Keele S, et al (2007) Guidelines for performing systematic literature reviews in software engineering. Technical report, ver. 2.3 ebse technical report. ebse
24. Qiu L, Zhu M, Zhou J (2024) Opkt: Enhancing knowledge tracing with optimized pretraining mechanisms in intelligent tutoring. IEEE Trans Learn Technol 17:841–855
25. Abdelrahman G, Wang Q (2023) Learning data teaching strategies via knowledge tracing. Knowl-Based Syst 269:110511
26. Ma Y, Wang L, Zhang J, Liu F, Jiang Q (2023) A personalized learning path recommendation method incorporating multi-algorithm. Appl Sci 13(10):5946
27. Piech C, Spencer J, Huang J, Ganguli S, Sahami M, Guibas L, Sohl-Dickstein J (2015) Deep knowledge tracing. Computer. Science 3(3):19–23
28. Albert T, CorbettJohn R (1994) Anderson: Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model User Adapt Int
29. Käser T, Klingler S, Schwing AG, Gross M (2017) Dynamic bayesian networks for student modeling. IEEE Trans Learn Technol 10(4):450–462
30. Cen H, Koedinger K, Junker B (2006) Learning factors analysis–a general method for cognitive model evaluation and improvement. In: International conference on intelligent tutoring systems, pp 164–175. Springer
31. Pavlik Jr, PI, Cen H, Koedinger KR (2009) Performance factors analysis–a new alternative to knowledge tracing. Online Submission
32. Vie J-J, Kashima H (2019) Knowledge tracing machines: Factorization machines for knowledge tracing. In: Proceedings of the AAAI conference on artificial intelligence, vol 33, pp 750–757
33. Zhang J, Shi X, King I, Yeung D-Y (2017) Dynamic key-value memory networks for knowledge tracing. In: Proceedings of the 26th international conference on world wide web, pp 765–774
34. Su Y, Liu Q, Liu Q, Huang Z, Yin Y, Chen E, Ding C, Wei S, Hu G (2018) Exercise-enhanced sequential modeling for student performance prediction. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
35. Pandey S, Karypis G (2019) A self attentive model for knowledge tracing. In: Proceedings of the 12th international conference on educational data mining, EDM 2019, Montréal, Canada, July 2-5, 2019
36. Ghosh A, Heffernan N, Lan AS (2020) Context-aware attentive knowledge tracing. In: Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining, pp 2330–2339
37. Nakagawa H, Iwasawa Y, Matsuo Y (2019) Graph-based knowledge tracing: modeling student proficiency using graph neural network. In: IEEE/WIC/ACM international conference on web intelligence, pp 156–163
38. Pandey S, Srivastava J (2020) Rkt: relation-aware self-attention for knowledge tracing. In: Proceedings of the 29th ACM international conference on information & knowledge management, pp 1205–1214
39. Liu Q, Huang Z, Yin Y, Chen E, Xiong H, Su Y, Hu G (2019) Ekt: Exercise-aware knowledge tracing for student performance prediction. IEEE Trans Knowl Data Eng 33(1):100–115
40. Minn S, Vie J-J, Takeuchi K, Kashima H, Zhu F (2022) Interpretable knowledge tracing: Simple and efficient student modeling with causal relations. In: Proceedings of the AAAI conference on artificial intelligence, vol 36, pp 12810–12818
41. Baum LE, Petrie T (1966) Statistical inference for probabilistic functions of finite state markov chains. Ann Math Stat 37(6):1554–1563
42. Schmucker R, Wang J, Hu S, Mitchell T (2022) Assessing the performance of online students - new data, new approaches, improved accuracy. J Educ Data Mining 14(1):1–45
43. Tong H, Wang Z, Zhou Y, Tong S, Han W, Liu Q (2022) Introducing problem schema with hierarchical exercise graph for knowledge tracing. In: Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval, pp 405–415
44. Su Y, Cheng Z, Luo P, Wu J, Zhang L, Liu Q, Wang S (2021) Time-and-concept enhanced deep multidimensional item response theory for interpretable knowledge tracing. Knowl-Based Syst 218:106819
45. Chen J, Liu Z, Huang S, Liu Q, Luo W (2023) Improving interpretability of deep sequential knowledge tracing models with question-centric cognitive representations. In: Thirty-seventh AAAI conference on artificial intelligence, pp 14196–14204
46. Li Q, Yuan X, Liu S, Gao L, Wei T, Shen X, Sun J (2023) A genetic causal explainer for deep knowledge tracing. IEEE Trans Evol Comput
47. Zhu J, Ma X, Huang C (2023) Stable knowledge tracing using causal inference. IEEE Trans Learn Technol
48. Zhou T, Sheng H, Howley I (2020) Assessing post-hoc explainability of the bkt algorithm. AIES '20, pp 407–413. Association for Computing Machinery, New York, NY, USA
49. Fischer C, Pardos ZA, Baker RS, Williams JJ, Smyth P, Yu R, Slater S, Baker R, Warschauer M (2020) Mining big data in education: Affordances and challenges. Rev Res Educ 44(1):130–160
50. Bahdanau D, Cho K, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International conference on learning representations, ICLR 2015
51. Letham B, Rudin C, McCormick TH, Madigan D (2015) Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model
52. Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N (2015) Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In: Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1721–1730
53. Agarwal R, Melnick L, Frosst N, Zhang X, Lengerich B, Caruana R, Hinton GE (2021) Neural additive models: Interpretable machine learning with neural nets. Adv Neural Inf Process Syst 34:4699–4711
54. Erhan D, Courville A, Bengio Y (2010) Understanding representations learned in deep architectures
55. Simonyan K, Vedaldi A, Zisserman A (2014) Deep inside convolutional networks: Visualising image classification models and saliency maps. In: 2nd International Conference on Learning Representations, ICLR 2014
56. Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Computer vision–ECCV 2014: 13th

european conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13, pp 818–833. Springer
57. Springenberg JT, Dosovitskiy A, Brox T, Riedmiller MA (2015) Striving for simplicity: The all convolutional net. In: 3rd International Conference on Learning Representations, ICLR 2015
58. Smilkov D, Thorat N, Kim B, Viégas F, Wattenberg M (2017) Smoothgrad: removing noise by adding noise. arXiv:1706.03825
59. Bach S, Binder A, Montavon G, Klauschen F, Müller K-R, Samek W (2015) On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS One 10(7):0130140
60. Li H, Tian Y, Mueller K, Chen X (2019) Beyond saliency: understanding convolutional neural networks from saliency prediction on layer-wise relevance propagation. Image Vis Comput 83:70-86
61. Zhou B, Khosla A, Lapedriza A, Oliva A, Torralba A (2016) Learning deep features for discriminative localization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2921–2929
62. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE international conference on computer vision, pp 618–626
63. Chattopadhay A, Sarkar A, Howlader P, Balasubramanian VN (2018) Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In: 2018 IEEE winter conference on applications of computer vision (WACV), pp 839–847
64. Ribeiro MT, Singh S, Guestrin C (2016) " why should i trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp 1135–1144
65. Lundberg SM, Lee S-I (2017) A unified approach to interpreting model predictions. Adv Neural Inf Process Syst 30
66. Kim B, Wattenberg M, Gilmer J, Cai C, Wexler J, Viegas F, et al (2018) Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In: International conference on machine learning, pp 2668–2677. PMLR
67. Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42
68. Agarwal C, Krishna S, Saxena E, Pawelczyk M, Johnson N, Puri I, Zitnik M, Lakkaraju H (2022) Openxai: Towards a transparent evaluation of model explanations. In: Advances in neural information processing systems, vol 35, pp 15784–15799
69. Zhang Y, Xu F, Zou J, Petrosian OL, Krinkin KV (2021) Xai evaluation: Evaluating black-box model explanations for prediction. In: 2021 II International conference on neural networks and neurotechnologies (NeuroNT), pp 13–16
70. Kamath U, Liu J (2021) Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning. Springer, ???
71. Adadi A, Berrada M (2018) Peeking inside the black-box: A survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
72. Mohseni S, Zarei N, Ragan ED (2021) A multidisciplinary survey and framework for design and evaluation of explainable ai systems. ACM Trans Interact Intell Syst (TiiS) 11(3–4):1–45
73. Doshi-Velez F, Kim B (2018) Considerations for evaluation and generalization in interpretable machine learning. Explain Interpret Models Comput Vis Mach Learn 1:3–17
74. Hoffman RR, Mueller ST, Klein G, Litman J (2018) Metrics for explainable ai: Challenges and prospects. arXiv:1812.04608
75. Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inf 113:103655
76. Balog K, Radlinski F (2020) Measuring recommendation explanation quality: The conflicting goals of explanations. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 329–338
77. Lage I, Chen E, He J, Narayanan M, Kim B, Gershman SJ, Doshi-Velez F (2019) Human evaluation of models built for interpretability. In: Proceedings of the AAAI conference on human computation and crowdsourcing, vol 7, pp 59–67
78. Markus AF, Kors JA, Rijnbeek PR (2021) The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies. J Biomed Inf 113:103655
79. Kumarakulasinghe NB, Blomberg T, Liu J, Leao AS, Papapetrou P (2020) Evaluating local interpretable model-agnostic explanations on clinical machine learning classification models. In: 2020 IEEE 33rd international symposium on computer-based medical systems (CBMS), pp 7–12. IEEE
80. Vilone G, Longo L (2021) Notions of explainability and evaluation approaches for explainable artificial intelligence. Inf Fusion 76:89–106
81. Alvarez Melis D, Jaakkola T (2018) Towards robust interpretability with self-explaining neural networks. Adv Neural Inf Process Syst 31
82. Robnik-Šikonja M, Bohanec M (2018) Perturbation-based explanations of prediction models. Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent 159–175
83. Moraffah R, Karami M, Guo R, Raglin A, Liu H (2020) Causal interpretability for machine learning-problems, methods and evaluation. ACM SIGKDD Explorations Newsletter 22(1):18–33
84. Wu M, Hughes M, Parbhoo S, Zazzi M, Roth V, Doshi-Velez F (2018) Beyond sparsity: Tree regularization of deep models for interpretability. In: Proceedings of the AAAI conference on artificial intelligence, vol 32
85. Fan L, Liu C, Zhou Y, Zhang T, Yang Q (2020) Interpreting and evaluating black box models in a customizable way. In: 2020 IEEE international conference on big data (Big Data), pp 5435–5440. IEEE
86. Coroama L, Groza A (2022) Evaluation metrics in explainable artificial intelligence (xai). In: International conference on advanced research in technologies, information, innovation and sustainability, pp 401–413. Springer
87. Merriam-Webster (2019) Merriam-Webster's Collegiate Dictionary. Merriam-Webster, Springfield, MA
88. Lipton P (1990) Contrastive explanation. R Inst Philos Suppl 27:247–266
89. Brockman J (2013) This Explains Everything: 150 Deep, Beautiful, and Elegant Theories of How the World Works. Harper Collins, ???
90. Lombrozo Tania (2016) Explanatory preferences shape learning and inference. Trends Cogn Sci 748–759
91. Williamson K, Kizilcec RF (2021) Effects of algorithmic transparency in bayesian knowledge tracing on trust and perceived accuracy. Int Educ Data Mining Soc
92. Ghai B, Liao QV, Zhang Y, Bellamy R, Mueller K (2021) Explainable active learning (xal) toward ai explanations as interfaces for machine teachers. Proc ACM Human-Comput Inter 4(CSCW3):1–28
93. Conati C, Barral O, Putnam V, Rieger L (2021) Toward personalized xai: A case study in intelligent tutoring systems. Artif Intell 298:103503
94. Phillips PJ, Hahn CA, Fontana PC, Broniatowski DA, Przybocki MA (2020) Four principles of explainable artificial intelligence. Gaithersburg, Maryland 18
95. Ferreira, J.J., Monteiro MS (2020) What are people doing about xai user experience? a survey on ai explainability research and practice. In: Design, user experience, and usability. Design for

contemporary interactive environments: 9th international conference, DUXU 2020, Held as Part of the 22nd HCI International Conference, HCII 2020, Copenhagen, Denmark, July 19–24, 2020, Proceedings, Part II 22, pp 56–73. Springer
96. Khosravi H, Shum SB, Chen G, Conati C, Tsai Y-S, Kay J, Knight S, Martinez-Maldonado R, Sadiq S, Gašević D (2022) Explainable artificial intelligence in education. Comput Educ Artif Intell 3:100074
97. Hoffman RR, Mueller ST, Klein G, Litman J (2023) Measures for explainable ai: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-ai performance. Front Comput Sci 5:1096257
98. Harrell FE, et al (2001) Regression Modeling Strategies: with Applications to Linear Models, Logistic Regression, and Survival Analysis vol. 608. Springer, ???
99. Peng C, Ingersoll L (2002) An introduction to logistic regression analysis and reporting. J Educ Res 96(1):3–14
100. Carlin BP, Chib S (1995) Bayesian model choice via markov chain monte carlo methods. J R Stat Soc Series B (Methodological) 57(3):473–484
101. Lye A, Cicirello A, Patelli E (2021) Sampling methods for solving bayesian model updating problems: A tutorial. Mech Syst Signal Process 159:107760
102. Raftery AE (1995) Bayesian model selection in social research. Sociol Methodol 111–163
103. Quinlan JR (1987) Simplifying decision trees - sciencedirect. Int J Man-Mach Stud 27(3):221–234
104. Rivest HRL (1976) Constructing optimal binary decision trees is np-complete. Inf Process Lett
105. Guo G, Wang H, Bell D, Bi Y, Greer K (2004) An knn model-based approach and its application in text categorization. In: International conference on intelligent text processing and computational linguistics, pp 559–570. Springer
106. Nefeslioglu H, Sezer E, Gokceoglu C, Bozkir A, Duman T (2010) Assessment of landslide susceptibility by decision trees in the metropolitan area of istanbul, turkey. Math Prob Eng 2010
107. Imandoust SB, Bolandraftar M et al (2013) Application of k-nearest neighbor (knn) approach for predicting economic events: Theoretical background. Int J Eng Res Appl 3(5):605–610
108. Setnes M, Babuska R, Verbruggen HB (1998) Rule-based modeling: Precision and transparency. IEEE Trans Syst Man Cybern Part C (Applications and Reviews) 28(1):165–169
109. Núñez H, Angulo C, Català A (2002) Rule extraction from support vector machines. In: Esann, pp 107–112
110. Nunez H, Angulo C, Catala A (2006) Rule-based learning systems for support vector machines. Neural Process Lett 24(1):1–18
111. Hastie T, Tibshirani R (1987) Generalized additive models: some applications. J Am Stat Assoc 82(398):371–386
112. Hastie TJ (2017) Generalized additive models. In: Statistical models in S, pp 249–307. Routledge, ???
113. Wood SN (2006) Generalized Additive Models: an Introduction with R. chapman and hall/CRC, ???
114. Lipton ZC (2018) The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57
115. Zilke JR, Mencía EL, Janssen F (2016) Deepred - rule extraction from deep neural networks. Springer International Publishing
116. Shrikumar A, Greenside P, Shcherbina A, Kundaje A (2016) Not just a black box: Learning important features through propagating activation differences
117. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhutdinov R, Zemel R, Bengio Y (2015) Show. Neural Image Caption Generation with Visual Attention, Attend and Tell
118. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D (2020) Grad-cam: Visual explanations from deep networks via gradient-based localization. Int J Comput Vis 128(2):336–359
119. Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (xai): A survey. arXiv:2006.11371
120. Sarkar A, Vijaykeerthy D, Sarkar A, Balasubramanian VN (2022) A framework for learning ante-hoc explainable models via concepts. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp 10286–10295
121. Pardos ZA, Heffernan NT (2010) Modeling individualization in a bayesian networks implementation of knowledge tracing. In: International conference on user modeling, adaptation, and personalization, pp 255–266. Springer
122. Lee JI, Brunskill E (2012) The impact on individualizing student models on necessary practice opportunities. Int Educ Data Mining Soc
123. Hawkins WJ, Heffernan NT (2014) Using similarity to the previous problem to improve bayesian knowledge tracing. In: EDM (Workshops)
124. Wang Z, Zhu J, Li X, Hu Z, Zhang M (2016) Structured knowledge tracing models for student assessment on coursera. In: Proceedings of the third (2016) ACM conference on learning@ Scale, pp 209–212
125. Sun S, Hu X, Bu C, Liu F, Zhang Y, Luo W (2022) Genetic algorithm for bayesian knowledge tracing: A practical application. In: International conference on sensing and imaging, pp 282–293. Springer
126. Spaulding S, Breazeal C (2015) Affect and inference in bayesian knowledge tracing with a robot tutor. In: Proceedings of the Tenth Annual ACM/IEEE international conference on human-robot interaction extended abstracts, pp 219–220
127. Nedungadi P, Remya M (2014) Predicting students' performance on intelligent tutoring system–personalized clustered bkt (pc-bkt) model. In: 2014 IEEE frontiers in education conference (FIE) proceedings, pp 1–6. IEEE
128. Corbett AT, Anderson JR (1994) Knowledge tracing: Modeling the acquisition of procedural knowledge. User Model User-adapted Inter 4(4):253–278
129. Baylor C, Hula W, Donovan NJ, Doyle PJ, Kendall D, Yorkston K (2011) An introduction to item response theory and rasch models for speech-language pathologists
130. Cen, H., Koedinger, K., Junker, B.: Comparing two irt models for conjunctive skills. In: International Conference on Intelligent Tutoring Systems, pp. 796–798 (2008). Springer
131. Lindsey RV, Shroyer JD, Pashler H, Mozer MC (2014) Improving students' long-term knowledge retention through personalized review. Psychol Sci 25(3):639–647
132. Choffin B, Popineau F, Bourda Y, Vie J (2019) DAS3H: modeling student learning and forgetting for optimally scheduling distributed practice of skills. In: Proceedings of the 12th international conference on educational data mining, EDM 2019. International Educational Data Mining Society (IEDMS), ???
133. Gervet T, Koedinger K, Schneider J, Mitchell T et al (2020) When is deep learning the best approach to knowledge tracing? J Educ Data Mining 12(3):31–54
134. Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R, Bengio Y (2015) Show, attend and tell: Neural image caption generation with visual attention. In: International conference on machine learning, pp 2048–2057. PMLR
135. Chefer H, Gur S, Wolf L (2021) Transformer interpretability beyond attention visualization. In: Computer vision and pattern recognition
136. Letarte G, Paradis F, Giguére P, Laviolette F (2018) Importance of self-attention for sentiment analysis. In: Empirical methods in natural language processing
137. Ito T, Tsubouchi K, Sakaji H, Yamashita T, Izumi K (2020) Word-level contextual sentiment analysis with interpretability. Proc AAAI Conf Artif Intell 34(4):4231–4238

138. Wu Q, Zhang H, Gao X, He P, Weng P, Gao H, Chen G (2019) Dual graph attention networks for deep latent representation of multifaceted social effects in recommender systems. In: The World Wide Web Conference, pp 2091–2102
139. Fan J, Jiang Y, Liu Y, Zhou Y (2022) Interpretable mooc recommendation: a multi-attention network for personalized learning behavior analysis. Int Res Electron Netw Appl Pol (2):32
140. Zhao J, Bhatt S, Thille C, Zimmaro D, Gattani N (2020) Interpretable personalized knowledge tracing and next learning activity recommendation. In: Proceedings of the seventh ACM conference on learning@ scale, pp 325–328
141. Zhang M, Zhu X, Ji Y (2021) Input-aware neural knowledge tracing machine. In: International conference on pattern recognition, pp 345–360. Springer
142. Li J, Jiang C, Ye S (2022) A knowledge tracing model based on attention mechanism. In: 2022 International conference on machine learning, cloud computing and intelligent mining (MLCCIM), pp 219–224. IEEE
143. Zu S, Li L, Shen J (2023) Cakt: Coupling contrastive learning with attention networks for interpretable knowledge tracing. In: 2023 International joint conference on neural networks (IJCNN), pp 1–8
144. Yeung C (2019) Deep-irt: Make deep learning based knowledge tracing explainable using item response theory. In: Proceedings of the 12th international conference on educational data mining. international educational data mining society (IEDMS), ???
145. Gan W, Sun Y, Sun Y (2020) Knowledge interaction enhanced knowledge tracing for learner performance prediction. In: 2020 7th International conference on behavioural and social computing (BESC), pp 1–6. IEEE
146. Converse G, Pu S, Oliveira S (2021) Incorporating item response theory into knowledge tracing. In: International conference on artificial intelligence in education, pp 114–118. Springer
147. Zhou Y, Li X, Cao Y, Zhao X, Ye Q, Lv J (2021) LANA: towards personalized deep knowledge tracing through distinguishable interactive sequences. In: Proceedings of the 14th international conference on educational data mining, EDM 2021
148. Liu S, Yu J, Li Q, Liang R, Zhang Y, Shen X, Sun J (2022) Ability boosted knowledge tracing. Inf Sci 596:567–587
149. Sun J, Wei M, Feng J, Yu F, Zou R, Li Q (2023) Progressive knowledge tracing: Modeling learning process from abstract to concrete. Expert Syst Appl 122280
150. Zhang L (2021) Learning factors knowledge tracing model based on dynamic cognitive diagnosis. Math Prob Eng 2021
151. Zhu J, Yu W, Zheng Z, Huang C, Tang Y, Fung GPC (2020) Learning from interpretable analysis: Attention-based knowledge tracing. In: International conference on artificial intelligence in education, pp 364–368. Springer
152. Lee U, Park Y, Kim Y, Choi S, Kim H (2022) Monacobert: Monotonic attention based convbert for knowledge tracing. arXiv:2208.12615
153. Zhang M, Zhu X, Zhang C, Qian W, Pan F, Zhao H (2023) Counterfactual monotonic knowledge tracing for assessing students' dynamic mastery of knowledge concepts. In: Proceedings of the 32nd ACM international conference on information and knowledge management, pp 3236–3246
154. Choi Y, Lee Y, Cho J, Baek J, Kim B, Cha Y, Shin D, Bae C, Heo J (2020) Towards an appropriate query, key, and value computation for knowledge tracing. In: Proceedings of the seventh ACM conference on learning@ scale, pp 341–344
155. Yue Y, Sun X, Ji W, Yin Z, Sun C (2023) Augmenting interpretable knowledge tracing by ability attribute and attention mechanism. arXiv:2302.02146
156. Zu S, Li L, Shen J (2023) Cakt: Coupling contrastive learning with attention networks for interpretable knowledge tracing. In: 2023 International joint conference on neural networks (IJCNN), pp 1–8. IEEE
157. Drasgow F, Hulin CL (1990) Item response theory
158. Fosnot CT (2013) Constructivism: Theory, Perspectives, and Practice. Teachers College Press, ???
159. Averell L, Heathcote A (2011) The form of the forgetting curve and the fate of memories. J Math Psychol 55(1):25–35
160. Anzanello MJ, Fogliatto FS (2011) Learning curve models and applications: Literature review and research directions. Int J Ind Ergon 41(5):573–583
161. Wang X, Zheng Z, Zhu J, Yu W (2023) What is wrong with deep knowledge tracing? attention-based knowledge tracing. Appl Intell 53(3):2850–2861
162. Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (xai). IEEE Access 6:52138–52160
163. Madsen A, Reddy S, Chandar S (2022) Post-hoc interpretability for neural nlp: A survey. ACM Comput Surv 55(8):1–42
164. Gou J, Yu B, Maybank SJ, Tao D (2021) Knowledge distillation: A survey. Int J Comput Vis 129(6):1789–1819
165. Wang L, Yoon K-J (2021) Knowledge distillation and student-teacher learning for visual intelligence: A review and new outlooks. IEEE Trans Pattern Anal Mach Intell 44(6):3048–3068
166. Xue G, Chang Q, Wang J, Zhang K, Pal NR (2022) An adaptive neuro-fuzzy system with integrated feature selection and rule extraction for high-dimensional classification problems. IEEE Trans Fuzzy Syst
167. Casarotto S, Fecchio M, Rosanova M, Varone G, D'Ambrosio S, Sarasso S, Pigorini A, Russo S, Comanducci A, Ilmoniemi RJ et al (2022) The rt-tep tool: real-time visualization of tms-evoked potentials to maximize cortical activation and minimize artifacts. J Neurosci Methods 370:109486
168. Hara S, Hayashi K (2018) Making tree ensembles interpretable: A bayesian model selection approach. In: International conference on artificial intelligence and statistics, pp 77–85. PMLR
169. Obregon J, Kim A, Jung J-Y (2019) Rulecosi: Combination and simplification of production rules from boosted decision trees for imbalanced classification. Expert Syst Appl 126:64–82
170. Konstantinov AV, Utkin LV (2021) Interpretable machine learning with an ensemble of gradient boosting machines. Knowl-Based Syst 222:106993
171. Craven MW, Shavlik JW (2014) Learning symbolic rules using artificial neural networks. In: Proceedings of the tenth international conference on machine learning, pp 73–80
172. Zeltner D, Schmid B, Csiszár G, Csiszár O (2021) Squashing activation functions in benchmark tests: Towards a more explainable artificial intelligence using continuous-valued logic. Knowl-Based Syst 218:106779
173. Arras L, Horn F, Montavon G, Müller K-R, Samek W (2017) " what is relevant in a text document?": An interpretable machine learning approach. PloS One 12(8):0181142
174. Lu Y, Wang D, Meng Q, Chen P (2020) Towards interpretable deep learning models for knowledge tracing. In: Artificial intelligence in education: 21st international conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21, pp 185–190. Springer
175. Wang D, Lu Y, Meng Q, Chen P (2021) Interpreting deep knowledge tracing model on ednet dataset. arXiv:2111.00419
176. Lu Y, Wang D, Chen P, Meng Q, Yu S (2022) Interpreting deep learning models for knowledge tracing. Int J Artif Intell Educ, 1–24
177. Yao L, Chu Z, Li S, Li Y, Gao J, Zhang A (2021) A survey on causal inference. ACM Trans Knowl Discovery Data (TKDD) 15(5):1–46
178. Pearl J (2010) Causal inference. Causality: objectives and assessment, 39–58

179. Dinga R, Schmaal L, Penninx BW, Veltman DJ, Marquand AF (2020) Controlling for effects of confounding variables on machine learning predictions. 2020–08
180. Huang C, Wei H, Huang Q, Jiang F, Han Z, Huang X (2024) Learning consistent representations with temporal and causal enhancement for knowledge tracing. Expert Syst Appl 245:123128
181. Ding X, Larson EC (2019) Why deep knowledge tracing has less depth than anticipated. Int Educ Data Mining Soc
182. Ding X, Larson EC (2021) On the interpretability of deep learning based models for knowledge tracing. arXiv:2101.11335
183. Shah T, Olson L, Sharma A, Patel N (2020) Explainable knowledge tracing models for big data: Is ensembling an answer? arXiv:2011.05285
184. Sun J, Zou R, Liang R, Gao L, Liu S, Li Q, Zhang K, Jiang L (2022) Ensemble knowledge tracing: Modeling interactions in learning process. Expert Syst Appl, 117680
185. Pu Y, Wu W, Peng T, Liu F, Liang Y, Yu X, Chen R, Feng P (2022) Embedding cognitive framework with self-attention for interpretable knowledge tracing. Sci Rep 12(1):1–11
186. Valero-Leal E, Carlon MKJ, Cross JS (2023) A shap-inspired method for computing interaction contribution in deep knowledge tracing. In: International conference on artificial intelligence in education, pp 460–465. Springer
187. Mandalapu V, Gong J, Chen L (2021) Do we need to go deep? knowledge tracing with big data. arXiv:2101.08349
188. Wang D, Lu Y, Zhang Z, Chen P (2022) A generic interpreting method for knowledge tracing models. In: International conference on artificial intelligence in education, pp 573–580. Springer
189. Dai H, Zhang Y, Yun Y, Shang X (2021) An improved deep model for knowledge tracing and question-difficulty discovery. In: Pacific rim international conference on artificial intelligence, pp 362–375. Springer
190. Sun J, Zhou J, Zhang K, Li Q, Lu Z (2021) Collaborative embedding for knowledge tracing. In: International conference on knowledge science, engineering and management, pp 333–342. Springer
191. Lee J, Yeung D-Y (2019) Knowledge query network for knowledge tracing: How knowledge interacts with skills. In: Proceedings of the 9th international conference on learning analytics & knowledge, pp 491–500
192. Tong S, Liu Q, Huang W, Hunag Z, Chen E, Liu C, Ma H, Wang S (2020) Structure-based knowledge tracing: an influence propagation view. In: 2020 IEEE international conference on data mining (ICDM), pp 541–550. IEEE
193. Zhao Z, Liu Z, Wang B, Ouyang L, Wang C, Ouyang Y (2022) Research on deep knowledge tracing model integrating graph attention network. In: 2022 Prognostics and health management conference (PHM-2022 London), pp 389–394. IEEE
194. Song X, Li J, Tang Y, Zhao T, Chen Y, Guan Z (2021) Jkt: A joint graph convolutional network based deep knowledge tracing. Inf Sci 580:510–523
195. Shrikumar A, Greenside P, Kundaje A (2017) Learning important features through propagating activation differences. In: International conference on machine learning, pp 3145–3153. PMLR
196. SHAPLEY L (1997) A value for n-person games[1]. Classics in Game Theory, 69
197. Kim S, Kim W, Jang Y, Choi S, Jung H, Kim H (2021) Student knowledge prediction for teacher-student interaction. In: Proceedings of the AAAI conference on artificial intelligence, vol 35, pp 15560–15568
198. Gan W, Sun Y, Sun Y (2022) Knowledge structure enhanced graph representation learning model for attentive knowledge tracing. Int J Intell Syst 37(3):2012–2045
199. Muñoz JLR, Ojeda FM, Jurado DLA, Peña PFP, Carranza CPM, Berríos HQ, Molina SU, Farfan ARM, Arias-Gonzáles JL, Vasquez-Pauca MJ (2022) Systematic review of adaptive learning technology for learning in higher education. Eurasian J Educ Res 98(98):221–233
200. Ai F, Chen Y, Guo Y, Zhao Y, Wang Z, Fu G, Wang G (2019) Concept-aware deep knowledge tracing and exercise recommendation in an online learning system. In: Proceedings of the 12th international conference on educational data mining, 2019. International Educational Data Mining Society (IEDMS), ???
201. Zhao J, Bhatt S, Thille C, Zimmaro D, Gattani N (2020) Interpretable personalized knowledge tracing and next learning activity recommendation. In: Proceedings of the seventh ACM conference on learning @ scale, pp 325–328. Association for Computing Machinery, ???
202. Cai D, Zhang Y, Dai B (2019) Learning path recommendation based on knowledge tracing model and reinforcement learning. In: 2019 IEEE 5th international conference on computer and communications (ICCC), pp 1881–5
203. He Y, Wang H, Pan Y, Zhou Y, Sun G (2022) Exercise recommendation method based on knowledge tracing and concept prerequisite relations. CCF Trans Pervasive Comput Interact 4(4):452–464
204. Wan H, Zhong Z, Tang L, Gao X (2023) Pedagogical interventions in spocs: Learning behavior dashboards and knowledge tracing support exercise recommendation. IEEE Trans Learn Technol
205. Fiok K, Farahani FV, Karwowski W, Ahram T (2022) Explainable artificial intelligence for education and training. J Def Model Simul 19(2):133–144
206. Liu H, Wang Y, Fan W, Liu X, Li Y, Jain S, Liu Y, Jain A, Tang J (2022) Trustworthy ai: A computational perspective. ACM Trans Intell Syst Technol 14(1):1–59
207. Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv:1702.08608
208. Feng M, Heffernan N, Koedinger K (2009) Addressing the assessment challenge with an online system that tutors as it assesses. User Model User-adapted Inter 19:243–266
209. Kosch T, Karolus J, Zagermann J, Reiterer H, Schmidt A, Woźniak PW (2023) A survey on measuring cognitive workload in human-computer interaction. ACM Comput Surv
210. 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. Artif Intell Rev 56(4):3005–3054
211. Rapp DN, Braasch JL (2023) Processing Inaccurate Information: Theoretical and Applied Perspectives from Cognitive Science and the Educational Sciences. MIT Press, ???
212. Andrews RW, Lilly JM, Srivastava D, Feigh KM (2023) The role of shared mental models in human-ai teams: a theoretical review. Theor Issues Ergon Sci 24(2):129–175
213. Kubiszyn T, Borich GD (2024) Educational Testing and Measurement. John Wiley & Sons, ???
214. Lakkaraju H, Bach SH, Leskovec J (2016) Interpretable decision sets: A joint framework for description and prediction. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, pp. 1675–1684. Association for Computing Machinery, ???
215. Balog K, Radlinski F (2020) Measuring recommendation explanation quality: The conflicting goals of explanations. In: Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval, pp 329–338. Association for Computing Machinery, New York, NY, USA
216. Alam, A.: Contemplative pedagogy: An experiment with school students for demystifying the philosophy of contemplative education. Resilience and Transformation in Global Restructuring, 289–300 (2022)
217. Cornacchia G, Anelli VW, Biancofiore GM, Narducci F, Pomo C, Ragone A, Di Sciascio E (2023) Auditing fairness under unaware-

ness through counterfactual reasoning. Inf Process Manag 60(2):103224
218. Adams C, Pente P, Lemermeyer G, Rockwell G (2023) Ethical principles for artificial intelligence in k-12 education. Comput Educ Artif Intell 4:100131
219. Yang M, Cheng H, Chen F, Liu X, Wang M, Li X (2023) Model poisoning attack in differential privacy-based federated learning. Inf Sci 630:158–172
220. Vasa J, Thakkar A (2023) Deep learning: Differential privacy preservation in the era of big data. J Comput Inf Syst 63(3):608–631