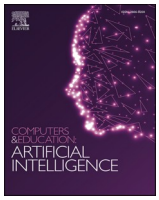
Computers and Education: Artificial Intelligence 3 (2022) 100074

Contents lists available at ScienceDirect

Computers and Education: Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/computers-and-education-artificial-intelligence



Explainable Artificial Intelligence in education

Hassan Khosravi [a,*], Simon Buckingham Shum [b], Guanliang Chen [c], Cristina Conati [d], Yi-Shan Tsai [c], Judy Kay [e], Simon Knight [b], Roberto Martinez-Maldonado [c], Shazia Sadiq [a], Dragan Gašević [c]

[a] The University of Queensland, Brisbane, Australia
[b] University of Technology Sydney, Australia
[c] Monash University, Australia
[d] The University of British Columbia, Australia
[e] The University of Sydney, Australia

ARTICLE INFO

Keywords:
Explainable AI
AI in Education
Open learner models

ABSTRACT

There are emerging concerns about the Fairness, Accountability, Transparency, and Ethics (FATE) of educational interventions supported by the use of Artificial Intelligence (AI) algorithms. One of the emerging methods for increasing trust in AI systems is to use eXplainable AI (XAI), which promotes the use of methods that produce transparent explanations and reasons for decisions AI systems make. Considering the existing literature on XAI, this paper argues that XAI in education has commonalities with the broader use of AI but also has distinctive needs. Accordingly, we first present a framework, referred to as XAI-ED, that considers six key aspects in relation to explainability for studying, designing and developing educational AI tools. These key aspects focus on the stakeholders, benefits, approaches for presenting explanations, widely used classes of AI models, human-centred designs of the AI interfaces and potential pitfalls of providing explanations within education. We then present four comprehensive case studies that illustrate the application of XAI-ED in four different educational AI tools. The paper concludes by discussing opportunities, challenges and future research needs for the effective incorporation of XAI in education.

1. Introduction

Artificial intelligence (AI) has a large and increasing role in education. One important case is of personalised teaching systems which are already well established, with growing evidence of their effectiveness for improving learning (VanLehn, 2011; Kulik & Fletcher, 2016; Steenbergen-Hu & Cooper, 2014, 2013; Ma, Adesope, Nesbit, & Liu, 2014; du Boulay, 2016). AI in education (AIED) systems may also make diverse and sophisticated use of AI to create the interface that is so important for the learning experience. For example, the interface may use natural language processing and generation, speech interfaces, avatars, video analysis of the learner to judge their attention and emotion (see Fig. 12).

These systems collect data about learners as they use the system. This may be collected from the interaction with the teaching interface as part of learning activities, while other more recent systems collect data beyond keyboard/mouse/screen actions, for example, from cameras, microphones and wearable devices. Learners may be more or less aware of the nature of the data being collected. European legislation in the General Data Protection Regulation (GDPR)[1] has led the world in codifying societal values around the governance of data and its use, reflecting a growing concern that people should be in control of technology and its use of their data (Knijnenburg et al., 2022; Wang, Yang, Abdul, & Lim, 2019; Wang et al., 2019). This means that learners should be able to determine how AI works, how that may affect them, and whether it is trustworthy (Drachsler & Greller, 2016; Holmes et al., 2021). It follows that AIED systems are one important context for the need for eXplainable AI (XAI).

While the role and need for XAI in education shares much in common with broader uses of AI (including accountability for accuracy, fairness and privacy management), education has distinctive needs for XAI and the nature of its data poses distinctive challenges. In particular, learning

* Corresponding author.
E-mail address: h.khosravi@uq.edu.au (H. Khosravi).
[1] https://gdpr-info.eu, visited March 2022.

https://doi.org/10.1016/j.caeai.2022.100074
Received 1 August 2021; Received in revised form 20 March 2022; Accepted 20 April 2022
Available online 13 May 2022

data has many *sources* of noise, and *reasoning* about such data is noisy at many levels (Kahneman, Sibony, & Sunstein, 2021). A significant body of work emphasises the importance of giving learners agency and responsibility for their own learning (Berger, Rugen, Woodfin, & Education, 2014; Winne, 1995), and in this context, XAI has the potential to assist (Kay, 2001), with some learning tools harnessing data to create interfaces that support the learner's metacognitive processes of self-monitoring, reflection and planning (Bull, 2020; Bull & Kay, 2013). There are also well-known phenomena that are of particular importance for education, such as the ways that a flawed AI system may introduce misconceptions or encourage students to game a system (Baker et al., 2009).

This paper tackles the particular challenges and mechanisms for XAI in education. It makes three main contributions. Firstly, in Section 3, we present the *XAI in Education (XAI-ED)* framework, that draws on the fields of AI, Human-Computer Interaction and the Cognitive and Learning Sciences. XAI-ED characterises the nature of XAI in Education in terms of questions about six key aspects: *the people involved (stakeholders) and the benefits to each group; how to deliver the explanation; the widely used classes of models used in education; the human-centred design of the AI and interfaces to* support *explanation; and the potential pitfalls of providing explanations.* The second contribution is a set of four case studies, presented in Section 4. These illustrate the importance of explainability in state-of-the art examples with reference to XAI-ED. The third contribution in Section 5 articulates a list of opportunities, challenges and future research needs for advancing XAI and its effective incorporation in education.

2. Background

This section provides an overview of the XAI literature and its association with education. Section 2.1 begins by considering how the AIED field is responding to the broad challenges of FATE. Section 2.2 reviews the current landscape of XAI. Section 2.3 provides an overview of the importance of explanations in human-human interactions in the context of education. Section 2.4 provides an overview of the work on open learner models (OLMs), which is arguably the most well-established application of explanations in human-machine interactions and XAI in education.

2.1. The FATE of AI in education

A range of analyses are now appearing in the literature, which provide conceptual frameworks for understanding the challenge of FATE for AIED. In a commentary on the field, Holmes et al. (2021) surveyed AIED researchers on their perceptions of the challenge of FATE, from which they distill a set of thematic challenges for the community. In more technical detail, Kizilcec and Lee (2022) present an incisive introduction to, and analysis of, the concept of algorithmic "fairness" in education, using the three key steps of "measurement (data input), model learning (algorithm), and action (presentation or use of output)" to illustrate where this can break down. Focusing specifically on machine learning-based AIED, Baker and Hawn (2021) develop a more detailed taxonomy that summarises the various sources of algorithmic bias, and how they can be mitigated in terms of "a framework for moving from unknown bias to known bias and from fairness to equity". We also refer the reader to the recently edited collection by Porayska-Pomsta, Woolf, Holmes, and Holstein (2021), and the forthcoming volume by Holmes & Porayska-Pomsta, (In Press).

To illustrate the kinds of biases that have been studied, we introduce a few examples. It is known that when modelling students' learning performance and identifying at-risk students, AI techniques may provide more accurate predictions for a group of students than the others simply based on the different demographic attributes of the students, e.g., female vs. male (Gardner, Brooks, & Baker, 2019). The authors propose that the differential prediction accuracy between different groups of students can be quantified by using a metric called Absolute Between-ROC Area (ABROCA) to measure the unfairness of a predictive model. With ABROCA, Gardner et al. (2019) evaluated the unfairness displayed by five AI models used to predict to what extent students in massive open online courses were likely to accomplish their studies. Similarly, Hutt, Gardner, Duckworth, and D'Mello (2019) assessed the unfairness of models predicting on-time graduation based on the data provided by students in the setting of college applications. A very different form of analysis by Sha et al. (2021) reported that AI models may be unfair to students when characterizing their posts in discussion forums. For instance, the posts by English-*as*-first-language students were more likely to be accurately classified than those made by non-native English speakers.

To summarise, FATE covers an extremely broad range of concerns, to which the AIED field is developing a range of responses. A powerful form of mitigation that is often referred in FATE literature is *explainability*, and it is to this intriguing concept that we now turn our attention.

2.2. Explainable AI

One of the emerging methods for increasing trust in AI systems is to use XAI, which promotes the use of methods that "enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" (Gunning, 2017). The initial focus of XAI was predominantly algorithm-centred. At a high level, machine learning models can be categorised based on their level of interpretability, which can be defined as the degree to which a human can understand the cause of a decision or be able to exactly reproduce what the model does (Miller, 2019). As a design criterion, having models with high interpretability is desirable since in principle, this can help mitigate against partiality in decision-making (detecting and correcting various forms of bias), increase robustness against adversarial perturbations (noise added to the input to fool the system while being quasi-imperceptible for humans (Moosavi-Dezfooli, Fawzi, Fawzi, & Frossard, 2017)) that could change the prediction, and improve the employment of meaningful variables and truthful causality in the model reasoning (Arrieta et al., 2020). Some models such as general additive models, rule-based models and decision trees, due to their relatively simple structure, are considered interpretable by design, making it easy to explain how these models work. For example in the case of a simple decision tree, the rules can be explained to a human such that they understand and have the ability to reproduce the decision made by the model. In contrast, some models such as tree ensembles, support vector machines and deep neural networks have complex structures that are not readily interpretable. To make these models explainable to humans, there has been extensive work in the XAI research community on developing post-hoc explainability techniques, which aim to explain how a model produces its predictions without elucidating the structure of the model (Lipton, 2018) (See Section 3.2 for some examples).

Recent advances in XAI have seen a shift and a developmental step towards socially-situated XAI by introducing and exploring a socio-technically informed perspective that incorporates the socio-organizational context into explaining AI-mediated decision-making (Ehsan, Liao, Muller, Riedl, & Weisz, 2021). This shift has leveraged insights from the social sciences where an explanation is seen not just as a product, but as a process that requires social interactions and a knowledge transfer process from an explainer to an explained (Miller, 2019). Srinivasan and Chander (2020) offer a helpful review of types of explanations from a cognitive science perspective, including *trust, troubleshooting or design, education, action, justification, aesthetics*, and *communication* as key aims in explanation. They note that in deciding what kind of explanation to give, the specific stakeholder and task should be considered, highlighting the need for human-centred approaches to generating explanations. The shift has therefore benefited from user-centred approaches and methodologies from the HCI community, which have demonstrated that a range of XAI techniques were

not as effective as assumed in *assisting in sensemaking* (Alqaraawi, Schuessler, Weiβ, Costanza, & Berthouze, 2020), *enhancing user trust* (Yang, Huang, Scholtz, & Arendt, 2020), or *enabling actionable decisions* (Liao, Gruen, & Miller, 2020). These views suggest that XAI in education needs to be sensitive to the learning context and consider the different actors in a learning community as potential audiences.

2.3. Explanations in education

Explanations in educational contexts come in diverse forms depending on the stakeholders and their particular aims or tasks. The need for explanations arises since educators must be accountable (e.g. to students, parents, or the government); when providing individual feedback to students; in giving teachers diagnostic feedback to understand where a class of students needs increased focus; and in parental consultations to help them support their child's learning.

In particular, feedback for students and teachers are a form of explanation in education that is key to educational outcomes. For students, the most common forms of explanations are provided by teachers regarding student performance on particular tasks, suggestions to improve, prompts for self-monitoring and directing, and affect-level comments (Hattie & Timperley, 2007). The main purposes of this type of feedback are to scaffold learning including the development of domain knowledge, self-regulated skills, and a sense of being. Moreover, feedback for students is seen as a relational process through which teachers may encourage positive motivation and help learners build confidence and self-esteem (Price, Handley, Millar, & O'donovan, 2010; Nicol & Macfarlane-Dick, 2006).

For teachers, feedback tends to serve the purpose of assessing the effectiveness of teaching approaches, learning design, and areas of support (particular) students may need. Common materials that may aid teaching reflections include grade distribution of students, class engagement (e.g., attendance), student surveys, parent-teacher communications, and peer evaluation. Feedback for teachers is crucial to the scholarship of teaching (Boyer, 1990), which positions teachers as learners who, through reflections on the feedback, construct instructional knowledge (e.g., instructional design), pedagogical knowledge (e.g., knowing how students learn), and curricular knowledge (e.g., rationales of curriculum design) (Kreber, 2005) that together allow teachers to perform effective teaching.

Another form of explanation that is important to the operation of an educational entity, and its ongoing improvement, involves the use of data to present an overall profile and performance of an institution, e.g., student enrolment data, academic performance (e.g., teaching & research quality), staff profile, performance, and retention, research outputs, institution income, teacher-student ratio, and others. Business intelligence has been employed broadly to manage and gain insights from this form of explanations, which particularly target administrators and managers as key audiences (Drake & Walz, 2018). Across these contexts, the aim is to support people in decision making, and to develop their judgement capacities.

2.4. Open and scrutable learner models

A defining feature of AIED has been the personalisation of teaching (Self et al., 1999). A driving goal for foundational AIED was to create intelligent tutoring systems that could achieve the huge learning benefits of an expert one-to-one human tutor (Bloom, 1984). There is still huge appeal in creating personalised teaching software that can do this at scale.

Such personalisation is driven by a *learner model*. This is often defined as the machine's set of beliefs about the learner's knowledge, goals, preferences and other attributes. Some learner models are detailed cognitive models. Notable cases are the cognitive tutors (Anderson, Corbett, Koedinger, & Pelletier, 1995) which were some of the first widely deployed intelligent tutoring systems (Koedinger, Anderson, Hadley, Mark et al., 1997) and constraint based tutors (Mitrovic, 2003) which were also rigorously evaluated and widely deployed (Mitrovic, 2012). Both these have quite complex learner models and reasoning about the learner. Many other personalised teaching systems have used much simpler, more ad-hoc learner models where the system designer determines the set of Knowledge Components (KCs) (Koedinger, Corbett, & Perfetti, 2012) to model. Commonly, KCs form an ontology, defined by the relationships between them. For example, a genetic graph relates KCs in terms of the generalisation, correction, and refinement that learners make as they progress (Goldstein, 1979). Other ontologies have relationships for prerequisites and hierarchical structure. The learner model also typically represents the value of each KC as a level, such as Bloom (Bloom et al., 1956), SOLO (Biggs & Collis, 2014) or a small integer value. As the learner interacts with a personalised teaching system, the learning data drives changes in the learner model and this, in turn, drives personalised learning.

In many AIED systems, aspects of the learner model are available to the learner via interfaces. Such Open Learner Models (OLMs) typically present a view of key KCs so that the student can monitor their learning progress. OLMs have been designed for multiple purposes (Bull & Kay, 2016). From an XAI perspective, some of these purposes are particularly important for education. The most important is to achieve improved learning outcomes, as demonstrated in work such as (Brusilovsky, Somyürek, Guerra, Hosseini, & Zadorozhny, 2015; Long & Aleven, 2017; Mitrovic & Martin, 2007) and many other examples reviewed in (Bull, 2020). A second valuable role for this form of XAI, is as a foundation for giving learners greater control and responsibility over their learning (Abdi, Khosravi, Sadiq and Gasevic, 2019, 2020).

One class of open learner models and personalised systems has been described as *scrutable* (Kay, 2006). This are designed so that a learner can delve into the way they work by answering questions such as: What "raw data" is kept about me? Where does that data come from? How do I control this inflow? How do I volunteer data about a KC? Similarly, the learner can scrutinise the learner model to answer questions like: What is modelled about me? How does the modelling process work? A more complete set is available in (Kay & Kummerfeld, 2019). There are two main reasons to describe this form of XAI as scrutability. Firstly, scrutability acknowledges the real effort needed to *scrutinise* a system to find the answers to such questions. Secondly, in sophisticated systems, even after scrutinising, the learner may gain understanding of just some aspects but this may fall short of the breadth of understanding implied by terms like transparency and explainability.

The unfairness issues witnessed in AI models, beyond question, greatly slow down the pace of wide adoption of AI in education, and thus calls for more work on XAI in education. In addition to unfairness, it has been reported that instructors often lack enough understanding of AI and so are concerned about it use (Chounta, Bardone, Raudsep, & Pedaste, 2021). This, again, calls for the development of relevant XAI in education presentation methods to communicate the prediction results with instructors and enhance their perception of AI and support them to tackle the challenges they face while using AI for teaching.

3. The XAI-ED framework

This section contextualises the use of XAI in education.

As previously discussed, the role and need for XAI in education has much in common with its broader use. Therefore, as shown in Fig. 1, XAI in education draws insights and best practices from the fields of AI, Human-Computer Interaction and the interdisciplinary and emerging field of Human-Centred AI where much of the initial research on XAI has been conducted. However, there are also distinctive needs for XAI in education that are grounded in theories from cognitive and learning sciences. Some of these needs and their applications have been under exploration in the interdisciplinary and emerging fields of AI in education and learning analytics.

Drawing insights from these related fields, this section characterises

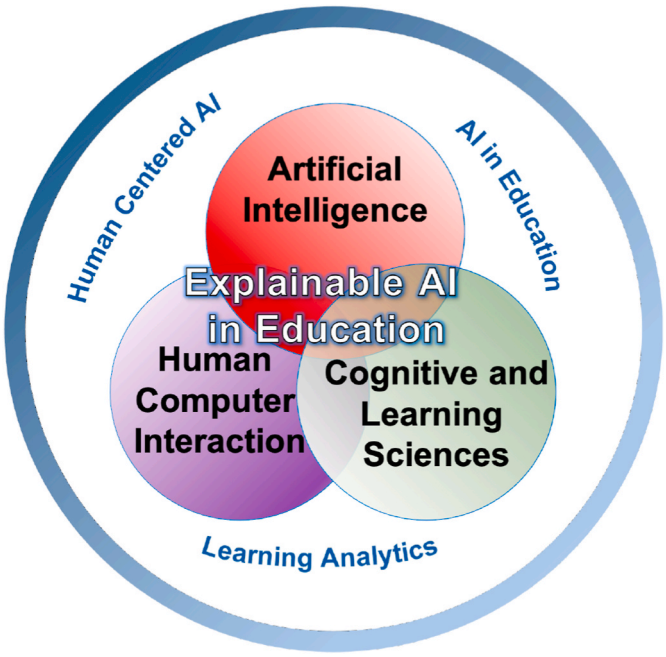


Fig. 1. Related fields to explainable AI in Education.

the nature of XAI in education in the XAI-ED framework that responds to fundamental questions about six key aspects. Section 3.1 presents an overview of the main stakeholders in education and how they may benefit from XAI in their roles and responsibilities. Section 3.2 presents some common approaches for presenting explanations in education. Section 3.3 presents some representative and widely used classes of models for implementing XAI in education. Section 3.4 discusses points for consideration in a human-centred design of the AI and interfaces to support explanation. Finally, Section 3.5 presents common pitfalls and potential shortcomings of incorporating XAI in education and recommendations to avoid them. Fig. 2 visualises the framework and its dimensions.

The XAI-ED aims to provide the means for educational tool developers and researchers to consider these six dimensions and "fill" in the parts of the framework for their particular context, which we hope can contribute to the development of more effective educational XAI systems.

3.1. Stakeholders and potential benefits

Much of the research in XAI has been devoted to providing explanations to system engineers or data scientists; however human-in-the-loop factors are crucial for the enactment of explanations in any given context (Srinivasan & Chander, 2020). The particular people seeking explanations, their intentions, and the explanations are important to the development of understanding of and trust in AI and the decisions AI feeds into (Páez, 2019).

Educational stakeholders include technologists and researchers, including learners, parents, teachers, educational administrators and policy makers, with each diverse and often different needs. All want educational benefits of AI that must also be accountable and trusted. Below, we explore some of the potential benefits of XAI for stakeholders in education.

Agency. Explanations of AI can facilitate conversation among these different stakeholders, further enabling processes of co-design and value co-creation (Dollinger, Liu, Arthars, & Lodge, 2019; Dollinger & Lodge, 2018). In particular, explanations of AI (including the choice of language and details to include) should allow students, teachers, and parents to see personal relevance (Srinivasan & Chander, 2020), thus empowering them to make conscious decisions as to whether or not to adopt AI and how to use AI in ways that may optimise values. Having functional understanding (Páez, 2019) – being able to interpret a decision made by the AI and the relation between an input and an output – will increase the confidence of these stakeholders in AI, and allow them to exert agency through AI (Selwyn, 2021), such as reasoning the credibility of a recommendation by AI and deciding whether or not to act on it.

Student-teacher interactions. On the other hand, XAI is important to facilitate the socio-cultural process of learning where interactions between teachers and students are fundamental to guide learners through zones of proximal development (Rieber & Carton, 1987). Although AI can significantly improve efficiency by automating some mundane work of teaching, the social role of teachers in a learning process remains unique and not replaceable by AI (Miao, Holmes, Huang, & Hui, 2021; Selwyn, 2019). For example, dialogic pedagogies (Lyle, 2008), Moore's theory of transactional distance (Moore, 2013), and the community of inquiry model (Garrison, 2016) highlight the importance of student-teacher interactions in building a learning community where learners can feel supported and belonging. Explanations of AI can prompt dialogue between learners and teachers on insights and implications of AI decisions and possibilities of such decisions to be challenged and improved, which can also feed back to the AI system for further improvement (Moore & Paris, 1992).

AI literacy. While AI-based innovations, such as learning analytics,

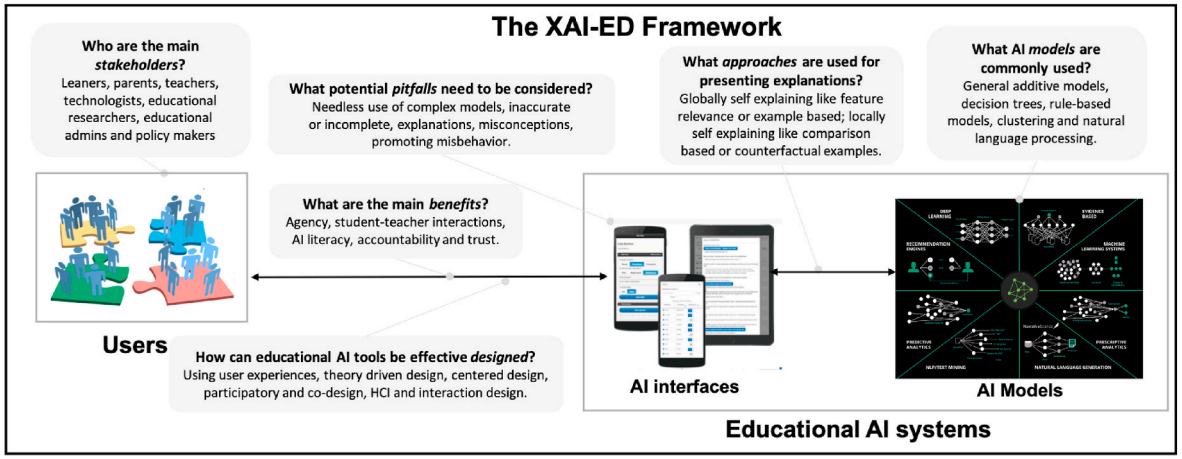


Fig. 2. The XAI-ED framework.

open up new ways to understand and support learning, the huge influence (often seen as disruptive) of AI on how people think and act calls for a new set of skills important to our navigation in a world of AI - AI literacy. Broadly speaking, AI literacy includes an understanding of what AI is, the ability to learn with AI, and the skill to collaborate with AI in a world that is increasingly integrated with AI (Miao et al., 2021). Long and Magerko (2020) defines AI literacy as a set of 17 competencies that enable "individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" (p. 598). They propose 14 design considerations to support AI developers and educators in creating learner-centred AI. Among these, explainability is the first to consider when designing an AI system, e.g., through the use of graphic visualisations, simulations, explanations of agent decision-making process, and iterative demonstrations to aid the learner's understanding of AI. Explainability of AI is also important for the development of AI literacy among teachers who play an orchestrator role in deciding when and how to best use AI tools to support learning in the classroom (Miao et al., 2021).

Accountability and trust. Motivations behind AI adoption in education are not purely educational but also socio-political. Higher education, in particular, faces constant pressure to provide evidence of education quality not only for quality assurance, but also to compete locally and internationally to attract students, staff, and funding. As a result, AI is not only perceived as promising in helping students develop essential skills as described above, but also to enhance institutional performance in areas such as teaching quality, student progression and retention, student satisfaction, graduate employment, and international reputation. This political agenda has attracted various parties outside the education domain (e.g., think tanks and EdTech industry) to offer solutions and shape education futures, resulting in outcries about education being manipulated by non-education experts (Williamson, 2018). In addition to this, several socio-political issues related to the way data is sourced, collected, processed, and transformed into seemly arbitrary decisions have caused fear and distrust in AI. Prominent issues include surveillance, privacy, bias, and context relevance (e.g., educational constructs) among others. Learning analytics, for example, were found to suffer from trust issues with reasons such as numbers being subjective, threat to agency and autonomy, and the design and implementation of learning analytics failing to meet needs or negatively affecting student well-being (Tsai, Whitelock-Wainwright, & Gašević, 2021). In light of the potential danger of AI, UNESCO published Beijing Consensus (UNESCO, 2019) outlining 44 recommendations to harness AI in education.

In summary, XAI can be a catalyst for several educational benefits and desirable futures of education. It is important to consider whom the explanations are for, what the purposes are, and how to effectively communicate the explanations to different stakeholders who need varying levels of understanding (Páez, 2019) to benefit directly or help others benefit from AI. We posit that XAI can play a crucial role in enabling collective efforts of various stakeholders to shape desirable education futures through responsible, fair, and effective use of AI. In particular, XAI needs to communicate algorithmic transparency effectively to educational administrators (e.g., institutional leaders and legal officers) who may ensure model compliance and downstream implications, to policy makers who are responsible for the governance and continuous monitoring of ethical, legal, and desirable use of AI, and to teachers who can provide feedback to enhance AI systems based on their domain expertise. In this way, XAI can help keeping educational entities and service providers accountable and address some trust issues around the use of AI, thus realising and scaling the potential benefits.

3.2. Approaches

Here we present some of the common approaches to explainability in AI that can help various stakeholders gain insights into the details of machine learning models. These approaches often rely on simplification techniques to generate proxies with reduced complexity. Fig. 3 presents an overview of these approaches.

Approaches to explainability in AI are generally classified according to two main criteria: (1) global approaches that explain the entire model (approaches a, b, and c) vs. local approaches that explain an individual prediction (approaches d, e, and f) and (2) self-explainable models that have a single structure (a and d) vs. post-hoc approaches which explain how a model produces its predictions without elucidating the structure of the model (b, c, e, f) (Arrieta et al., 2020). and chapters 5 and 6 of (Molnar, 2019) provide a more detailed view of these common XAI approaches.

Global explanations. This approach focuses on explaining how different features/variables affect predictions. Global explanations are generally model-specific and work best for models that are interpretable by design. The example given in Fig. 3-a illustrates a decision tree that can be used to predict the risk level of every student in a class. Many teacher-facing systems (e.g. (Lakkaraju et al., 2015)) use global explanations to inform instructors of the performance of their students. For models that are less interpretable by design, one approach is to implement a global surrogate by learning an interpretable model to approximate the predictions of a black box model. However, this method needs to be used with care as the surrogate model may provide an accurate representation for one subset of the dataset, but diverge widely for another subset.

Feature relevance. One common XAI approach presents the computed relevance, in terms of a score, for each feature in the prediction process (as shown in Fig. 3-b. A comparison of the scores among different features demonstrates the importance granted by the model to each of the features when producing its output. The scores can be computed using a variety of models such as linear or logistic regression or more recently Shapley values (Shapley, 2016) that take a coalitional game theory approach to fairly distribute the "payout" among the features. The relevance of each feature can be shared by simply reporting the scores or visualising them in various graphs. This method has been used within predictive learning analytics applications (e.g. (Lakkaraju et al., 2015)) to present the importance of each feature in determining the "at-riskness" of the learners.

Example-based. Example-based explainable approaches select particular instances to explain the model. A main advantage of these explanations is that they are generally model-agnostic and can be used to make any machine learning model more interpretable. They can provide both local explanations (described below) and global explanations. Fig. 3-c illustrates this approach for identifying boundaries of a risk model based on students' midterm grades at a global level. This is done by visualising the entire set of instances, highlighting the examples and articulating them.

Local explanations. This approach focuses on explaining a particular instance, independent of what might be happening at the model level. Local explanations are only generally model-agnostic; however, they work particularly well with rule-based models and decision trees. The example in Fig. 3-d illustrates how risk factor for an individual student is computed using a interpretable model. Many student-facing systems and open learner models (e.g., Abdi et al. (2020)) use local explanations to inform students of their progress without overwhelming the students about the entire model and system. For models that are less interpretable by design, one approach is to implement a series of local surrogate models by learning an interpretable model locally around the prediction (Ribeiro, Singh, & Guestrin, 2016).

Comparison. This approach selects particular instances to locally explain the outcome of other instances. It is particularly well suited to non-parametric models, such as K nearest neighbours, where the model structure is determined by the instances. The example in Fig. 3-e compares essays of two students with high similarity for academic integrity purposes. Many plagiarism detection systems use comparison to show the evidence for their prediction (Foltýnek, Meuschke and Gipp, 2019).

Counterfactual explanation: This approach generally describes a

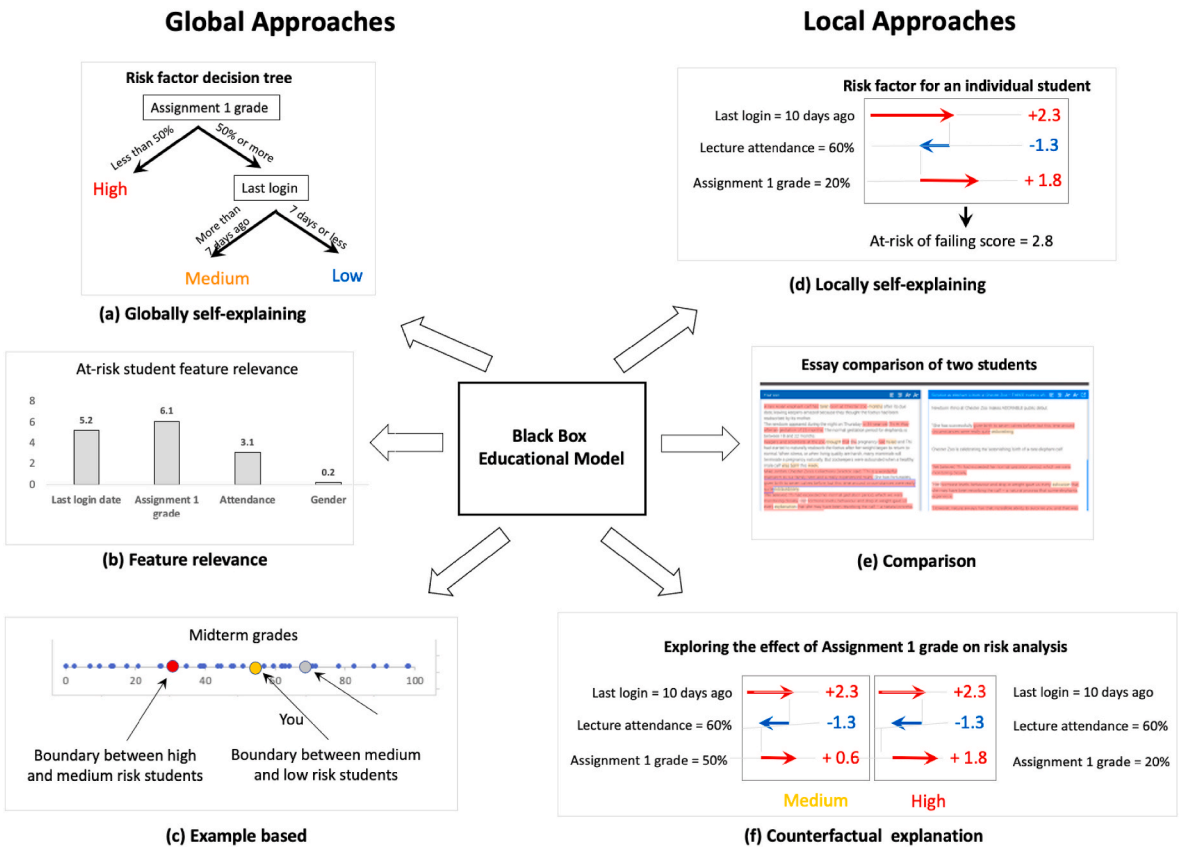


Fig. 3. Common explainability approaches.

causal situation in the form: "If X had not occurred, Y would not have occurred". Counterfactual explanations can be used to explain and demonstrate the smallest change to the feature values that changes the prediction to a predefined output. The example given in Fig. 3-f demonstrates the case of two students with similar values across features with the exception of their Assignment 1 grade which has led to one being categorised at medium risk while the other is categorised as high risk. The difference in outcome explains and provides evidence that the Assignment 1 grade plays an important part in the student risk modelling. There are two main challenges associated with counterfactual explanation (Molnar, 2019): (1) Identification of instances that provide counterfactual explanations can sometimes be a challenging task and (2) there may be multiple counterfactual explanations where each tells a different "story" of the importance of other features which may come across as contradictory or confusing.

3.3. Models

In this subsection, we present some representative AI models that are commonly used education. Some of the most promising use cases of AI in education include adaptive learning systems that tailor instruction based on student needs (Aleven, McLaughlin, Glenn, & Koedinger, 2016), adaptive testing systems that tailor assessment items based on students' mastery (Mills, Potenza, Fremer, & Ward, 2005), learner models and in particular open learner models that allow students to better regulate their learning (Bull, 2020), automated feedback tools that provide instant feedback on students' academic writing tasks (Knight, Abel, et al., 2020; Roscoe & McNamara, 2013) as well as intelligent learning analytics dashboards that help teachers with sense-making and identifying students in need of attention (Khosravi, Shabaninejad, et al., 2021). It should be noted that the majority of the examples we identified in the literature incorporated interpretable models for explainability rather than using post-hoc explanations alongside non-interpretable models.

Generalized Additive Model (GAM) assumes a linear relationship between the dependent variable and independent variables, and attempts to predict the value of the dependant variable by aggregating a number of smooth functions which take the independent variables as input. Given the underlying linear nature of GAM, we can easily verify the importance of an independent variable by evaluating how its corresponding smooth function affects the predicted value. For example, Dikaya, Avanesian, Dikiy, Kirik, and Egorova (2021) aimed to depict the relationship between the psychological traits of students and their attitudes toward remote learning during the COVID-19 pandemic. By using GAM, it was illustrated that the interpersonal communicative skills of students (e.g., being manipulative or shy) correlated highly with their attitudes towards remote learning. Many of the well-established interpretable learner models such as Adaptive Factor Models (AFM) (Cen, Koedinger, & Junker, 2006) and Performance Factor Analysis (PFA) (Pavlik Jr, Cen and Koedinger, 2009) fall into this category. It is worth noting that there is a of class learner models that use deep learning such as deep knowledge tracing (DKT) (Piech et al., 2015) which are not interpretable.

Decision Trees are one of the most widely-used machine learning techniques in the field of learning analytics and educational data mining. Different from GAM, decision trees aim to learn a set of decision rules, which are organized in a hierarchical tree-like manner, to determine the value of the dependant variable. For instance, when predicting students' competence in collaborative problem solving, Cukurova, Zhou, Spikol, and Landolfi (2020) and Pardo et al. (2016) built decision trees to provide a set of useful decision rules to capture factors that were essential to students' performance. Due to their hierarchical structure, decision trees have been widely regarded as a useful technique to enable XAIED. However, if the model has a large number of complex features,

the size of decision tree can be very large and incomprehensible to a human user.

Rule-based learning is similar to decision trees and also aims at using a set of decision rules to generate predictions (Liu, Gegov, & Cocea, 2015). However, the rules are not necessarily structured in a hierarchical tree-like manner. The rules can take the form of simple conditional statements (e.g., *"if X then Y"*) plus more complex combinations. For instance, Engin et al. (2014) applied rule-based learning to develop two educational expert systems to recommend courses and scholarships to students. The prediction accuracy of rule-based learning often depends on the coverage (or amount) and the specificity (length) of the adopted rules. Rule-based systems can explain their entire chain of reasoning, which can be used to generate explanations that an analyst can use to diagnose and debug, and which in suitably simplified form can be understood by non-technical stakeholders. However, the more rules a model uses, the less understandable it becomes to regular educational stakeholders; similarly, the specificity of the rules also tends to hinder its users from interpreting how it works.

Clustering methods such as K-means clustering (Krishna & Murty, 1999) group data points into multiple clusters by calculating their similarities with each other, which is often measured based on the distance between these data points. These clusters can be used to reveal and explain different data patterns, e.g., exploratory learning behaviours specific to different groups of students. For instance, the FUMA framework presented in Section 4.2 and Conati, Barral, Putnam, and Rieger (2021) incorporate clustering to generate interpretable and personalised hints that guide student learning.

Natural Language Processing has commonly been used for detecting potential plagiarism and identifying misconduct cases. For instance, Turnitin (Heckler, Rice, & Hobson Bryan, 2013) is a well-recognized tool used for plagiarism detection. It works by comparing one student submission against an archive of relevant documents (e.g., internet articles and academic publications) and produces a report indicating where the text within the submission matches another source. This can be viewed by instructors to help determine whether it is a plagiarism case or not. It can also be presented to a student to improve the submission. Batane (2010) demonstrated that, with the aid of Turnitin, fewer plagiarism cases occurred in an undergraduate class.

3.4. XAI designs

Providing appropriate information to help people understand AI can be considered a human-AI interaction *design* challenge (Liao et al., 2020). Simply opening the algorithmic 'black box' is not enough to understand the implications of AI in the sociotechnical system at large. Some of the most accurate AI models (e.g. deep learning algorithms) are very complex and hard even for data scientists to understand (Sejnowski, 2020). It is also very challenging for people without formal data analysis training to comprehend and trust even simple, rule-based algorithms. This is precisely the case for most end-users of data-intensive educational innovations (i.e., students and their teachers). In this subsection, we present some design approaches and areas of research and development that can be considered when creating AIED systems.

User experience (UX) is the process design teams follow to create products and systems that provide meaningful and relevant experiences to *end-users* (Hassenzahl & Tractinsky, 2006). User experience itself is subjective as it includes people's perceptions of utility, accessibility, efficiency and ease of use about a system. Yet, the design attributes that contribute to such user experience are objective. This means that researchers and designers can maximise the opportunities for people to have an effective experience with a system. UX design is key for the development of effective XAI systems. In fact, a recent study demonstrated that effectively designing causability and explanatory cues in AI systems help people understand the decision-making process of algorithms by bringing transparency and accountability into such systems (Shin, 2021). However, designers of AI innovations are facing unique challenges that are not present when designing in other contexts. For example, UX designers need to collaborate effectively with AI developers to create a joint vision of the ideal impact of the models, their evaluation and the technical feasibility in addressing end-user needs (Yang, Scuito, Zimmerman, Forlizzi, & Steinfeld, 2018). The current reality is that UX design is not always a key component in the development of AI algorithms. To address this, UX designers need to be included in the development cycle of the algorithms to understand the capabilities of such algorithms and improve the explainability of the system to end-users (Dove, Halskov, Forlizzi, & Zimmerman, 2017). This suggests specific AI literacy capabilities that need to be developed along with design core knowledge and practices.

User-centred design (UCD). It is starting to be acknowledged that a prerequisite to create effective AI and learning analytics innovations in education is to understand the authentic needs of students, educators and other educational stakeholders while designing AI innovations (Buckingham Shum, Ferguson, & Martinez-Maldonado, 2019), which is now establishing itself as an active stream of learning analytics research. Yet, it is critical to understand to what extent the voices of these key educational stakeholders can influence the design of XAI. From a classic bottom-up, UCD perspective, the target user (e.g. teachers or students) is an object of study (Sanders & Stappers, 2008). The researcher or designer observes or interviews the educational stakeholder, often bringing knowledge from theories. Then the person in charge of the design receives this knowledge, adds an understanding of the technology capabilities, and creates a product. In the context of XAIED systems, there is no guarantee that the target users will be able to understand the XAI explanations. According to UCD, a design needs to provide a comprehensible explanation about the AI algorithm or its output, based on target users' needs and capabilities (Xu, 2019). Thus, the same XAI version targeted at researchers or developers cannot be the same as that targeted at non-expert users.

Top-down, theory driven design. Top-down guidelines have been proposed to help designers choose relevant XAI techniques that may work effectively in a certain domain. For example, Arrieta et al. (2020) proposed taxonomies of explainability techniques related to specific machine learning algorithms which can be used by other researchers working with such algorithms. This way, there is no need to deeply engage with the user values in making design decisions about XAI. Educational theory can provide the necessary foundations to inform the design of AIED and learning analytics research and the interpretation of its results (Er et al., 2021, pp. 196–206). It may therefore be possible to align XAI design guidelines with well-established educational theories to effectively design XAI techniques that support explanation and reasoning. For example, Wang et al. (2019) drew on educational psychology theory to map from conceptualisations about how people reason and explain to the ways to design XAI to support such processes.

Participatory design and co-design are aimed at giving an active voice to the end-users. In contrast to seeing end-users as passive objects of study, they can have an active role in the design process and they are given the position of 'expert of their own experience' (Sanders & Stappers, 2008). In the development of AIED and learning analytics systems, there has been a growing call for considering educational stakeholders as equal partners instead of simply conducting consultation or evaluation sessions with them (Prieto-Alvarez, Martinez-Maldonado, & Anderson, 2018). Yet, this bring additional challenges in terms of how users without data analysis training can effectively communicate needs that can be addressed through XAI innovations. A way to address this is by identifying the types of explanations needed by different stakeholders and, then involving such stakeholders in the design process of the XAI itself. For example, Liao, Pribić, Han, Miller, and Sow (2021) proposed a question-driven process for XAI UX design to engage end-users in the formulation of modelling solutions. Toolkits to conduct design studies with users to identify XAI needs and capabilities are also emerging for XAI practitioners (e.g., https://www.uxai.design/).

HCI and Interaction Design (IxD). While XAI research has so far been dominated by AI and machine learning experts, a rapidly growing community of HCI and IxD researchers/professionals is forming to contribute to the design of the XAI interfaces and bring psychological theories of explanations to XAI research (Xu, 2019). HCI and IxD professionals can take advantage of their interdisciplinary approaches to help AI experts to create effective UIs, contribute to design algorithms that enable explainability and involve users in the design process, as highlighted above. Inroads are being made through human-computer interaction (HCI) studies and interaction design (IxD) experiences that are trying to bridge the AI and human perspectives (e.g. Gunning, 2017). AI innovations in education require a balance between learning theory, data science and design (Gašević, Kovanović, & Joksimović, 2017). We need to consider what is possible with current XAI methods to be used throughout the design process. At the same time, there is value in identifying the educational stakeholders of the sociotechnical system from early stages in the design process. Only then, would it be possible to give an active voice to learners, educators and other potential end-users, and identify what to explain and how to explain it.

3.5. Pitfalls and how to avoid them

While there are many benefits in increasing the explainability of educational AI systems, there are also challenges and pitfalls that need to be considered. This section describes some of the key pitfalls as well as pointing out potential solutions and principles for avoiding them.

Needless use of complex models. A common misjudgment is to use over-complex models in cases where the use of an interpretable model would have delivered a comparable or even superior performance. As an example, in the context of student modelling and knowledge tracing, Gervet, Koedinger, Schneider, Mitchell et al. (2020) report that logistic regression outperforms deep learning models on datasets of moderate size or containing a very large number of interactions per student. Therefore, designers should start with simple interpretable models and increase complexity in a step-wise manner where performance in both accuracy and interpretability are measured and compared (Molnar et al., 2020). Nevertheless, in some cases more complex models would significantly outperform interpretable models. To follow the same example, Gervet et al. (2020) report that the use of deep knowledge tracing approaches outperform more interpretable approaches on datasets of large size or where precise temporal information matters most. In these cases, one approach would be to complement the complex model with interpretable models for explainability. As an example, Ghosh, Heffernan, and Lan (2020) propose a learner model which couples complex attention-based neural network models with a series of novel, interpretable model components inspired by cognitive and psychometric models for explainability.

Inaccurate explanations. Explanations that are inaccurate or even incorrect may lead to unfavourable outcomes for the stakeholders. A common underlying cause of inaccurate explanations is the use of models that are poor, potentially due to under- or over-fitting or noisy data. For example, a lenient open learner model that overestimates the mastery level of students may encourage them to develop a false sense of confidence of their abilities, resulting in poor performance in summative assessments and exams. Given that an interpretation can only be as good as its underlying model, it is crucial to develop the model rigorously by conducting best practices for model selection, hyperperameter tuning and evaluation before the adoption of AI tools (Molnar et al., 2020). Additionally, with the increasing demand for model explainability, some AI tool designers may be tempted to provide plausible and convincing interpretations that may be inaccurate or even incorrect. One way to combat this issue is to only incorporate sound explanations based on the underlying mathematical foundations which are truthful in describing the underlying system (Kulesza, Burnett, Wong, & Stumpf, 2015).

Incomplete explanations. Given the complexity of many AI models, some system developers may be tempted to provide incomplete and simplified explanations that disguise the entire complexity of the model. While the incorporation of incomplete explanations may appeal to a broad base of users, it provides them with a false sense of comprehension which may again lead to unfavourable outcomes. One way to combat this issue is to only incorporate complete explanations that expose all aspects of the relevant mechanisms (Kulesza et al., 2015). An alternative method would be to clearly flag their incompleteness, and the rationale, to warn the user.

Misconceptions. Given the complexity of many AI systems, it is reasonable to assume that some users may misinterpret or not fully understand sound and complete explanations of the system. Kulesza et al. (2015) propose the use of iterations for providing sound and complete explanations without overwhelming the users. They suggest that "explanations could take the form of concise, easily consumable "bites" of information—if a user is interested in learning more about the system, he or she can attend to many of these explanations to iteratively build a higher-fidelity mental model" (p. 127). Iterative cycle of explanations can focus on breadth (i.e., various aspects of the model) or depth (i.e., drilling down into the details of the model).

Promoting dysfunctional behaviour. The challenge in many of the previously mentioned pitfalls was due to users developing a false sense understandably. However, another line of concern is that if users have access to information about how a decision or recommendation has been made, they may be able to game the system by altering their behaviour to gain a more favourable outcome. For example, by having access to multiple rounds of submitting assignments to plagiarism detection software, students may use a strategy of making minimal changes to "bypass" the system rather than developing the disposition not to engage in academic misconduct. In this case, the issue is with how the system is implemented rather than its AI functionality or interpretability. Taking best practices from game theory mechanism design may help in developing systems in which users are incentivised to avoid malicious gaming and other dysfunctional behaviours (Maskin, 2008).

4. Case studies

In this section, we present four comprehensive case studies that illustrate the application of XAI-ED for studying a diverse range of AIED systems. Section 4.1 discusses the use of explanations within an adaptive learning system powered by high-quality learnersourced content. Section 4.2 discusses the effect of adding explanation functionality to an adaptive system that helps students learn an algorithm for constraint satisfaction problems. Section 4.3 discusses the use of explanations in the context of feedback within a writing analytics tool. Finally, Section 4.4 discusses the case of using explanations for data storytelling through team work analytics in educational healthcare systems. Table 1 demonstrates the relation between each of the case studies and the dimensions of XAI-ED.

4.1. RiPPLE: A learnersourced adaptive educational system

4.1.1. Overview

Adaptive educational systems (AESs) make use of data about students and learning processes to adapt the level or type of instruction for each student (Aleven et al., 2016). To provide such adaptivity, AESs require access to a large pool of instructional material and learning resources. These resources are commonly created by domain experts, which makes AESs expensive to develop and challenging to scale (Aleven et al., 2016). RiPPLE (Khosravi, Kitto, & Joseph, 2019) is an AES that takes the crowdsourcing approach of partnering with students, also referred to as learnersourcing (Khosravi, Demartini, Sadiq, & Gasevic, 2021; Kim, 2015), to create the resource repository.

4.1.2. Stakeholders and benefits

RiPPLE is predominantly a course-level platform that instructors can incorporate in their teaching. Therefore, the platform should first and

Table 1
Relationship between the case studies and dimensions of XAI-ED.

	Main Stakeholders	XAI Benefits	Approaches	Models	XAI design	Pitfalls
RiPPLE	- Educators - Students - Educational researchers	- Accountability and trust - AI literacy - Agency	- Comparison - Local explanations - Example based	- NLP functions - Graph-based trust propagation - Content-based recommender systems - Elo rating system	- Co-design - User experience	- Disengagement - Dysfunctional behaviour
FUMA	- Educators - Students	- Trust - Sensemaking	- Global explanations - Local explanations	- Clustering - Classification - Association rule mining	- User-centred design	- Overly complexed models - May not benefit all students
AcaWriter	- Educators - Students	- Agency - Trust - AI literacy	- Local explanations - Comparisons	- Rule-based NLP	- Co-design in design-research cycles	- Narrowness of rules-based systems - Context sensitivity
TeamWork Analytics	- Educators - Students	- Accountability and Trust - Agency - Sensemaking	- Global explanations - Local explanations	- Rule-based learning	- Co-design	- Incomplete explanations - Dysfunctional behaviour

foremost appeal to the *instructors*. However, the majority of the features of the platform are designed to be student-facing and to help *students* with their learning. A secondary mission of developing RiPPLE is to support *educational researchers* in conducting ethical, sound, large scale educational research. As detailed below, several attempts have been made to employ XAI in various parts of RiPPLE. The intention has been to (1) help instructors make sense of how the platform is operating, *trust* its decisions, and provide oversight and raise concerns as needed; (2) enable students to develop *AI literacy* and to have *agency* to regulate their learning and trust the system; and (3) support educational researchers to evaluate the platform and propose changes to improve the system and increase *accountability*.

4.1.3. Approaches and models

Here we discuss various ways used to incorporate XAI in RiPPLE.

1. *Explainable automated feedback on poor quality peer reviews.* To effectively utilise a learnersourced repository of content, there is a need for a selection process to separate high-quality from low-quality resources. RiPPLE uses an evaluation process, which relies on crowdsourcing, where students review and evaluate existing resources. A common challenge with incorporating peer review and feedback is that some students may not have the incentive or the ability to provide high-quality feedback. To increase the quality of the provide reviews, RiPPLE uses a range of *NLP models* (Darvishi, Khosravi, & Sadiq, 2020) and *comparison* based explanation approaches to flag comments that are too similar to previously provided reviews or lack the required depth (as shown in Fig. 4-left). A set of tips (as shown in Fig. 4-right) provide training on providing better feedback. Data collected by system shows that around 35% of users whose reviews are flagged revise do go on to resubmit their comments while the other 65% ignore the hint and submit anyway (Darvishi, Khosravi, Abdi, Sadiq, & Gasevic, 2022).
2. *Explainable consensus approach.* RiPPLE assigns each resource for evaluation to multiple moderators, which then requires a consensus approach of optimally integrating the decisions made by multiple people towards an accurate final decision. Summary statistics such as mean or median aggregation are a common explainable consensus approach. However, these models are quite fragile against users with diverse abilities or interests (El Maarry, Güntzer, & Balke, 2015). To increase the accuracy, RiPPLE uses a *graph-based trust propagation approach* (Darvishi, Khosravi, & Sadiq, 2021) that infers the reliability of each moderator. The final decision is computed as a weighted average of the evaluation ratings given by the peer evaluators, which is easy to interpret. Fig. 5 shows an example of a *locally self-explaining* approach for sharing the evaluations and the inferred outcome with the author and moderators. They are first invited to vote on the helpfulness of each moderation. They are then asked to determine whether they agree with the outcome and provide further feedback. Interestingly, data captured by the platform shows that only 2.2% of the submitted responses disagree with the inferred outcome (Darvishi, Khosravi, Sadiq, & Gasevic, 2022).
3. *Explainable spot-checking.* Despite efforts to increase the quality of the peer reviews, the outcome of the consensus approach based on judgements from students, as experts-in-training, cannot be wholly trusted. Given the limited availability of instructors, RiPPLE incorporates a *content-based recommender system* that identifies and recommends resources that most need expert review (Darvishi et al., 2021). When flagging a resource for spot checking, RiPPLE uses absolute and relative points of *comparison* to help instructors make sense of the recommendation. Fig. 6 is an example of a resource

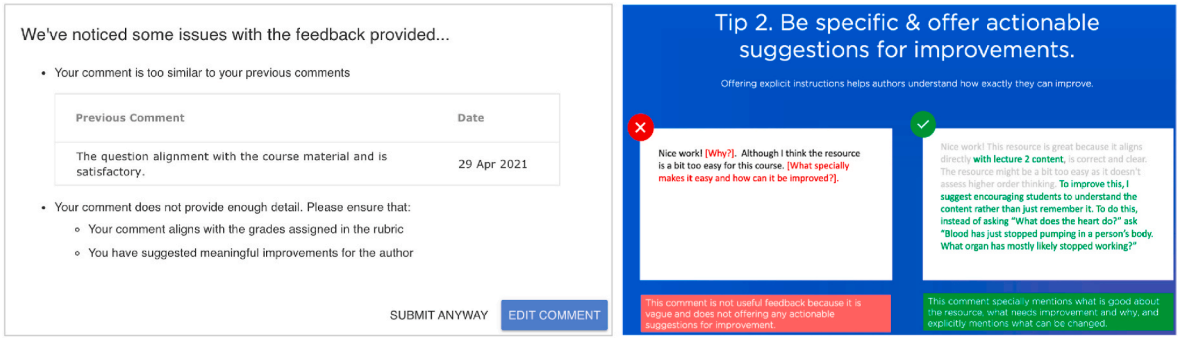


Fig. 4. Explainable suggestions for improving quality of reviews.

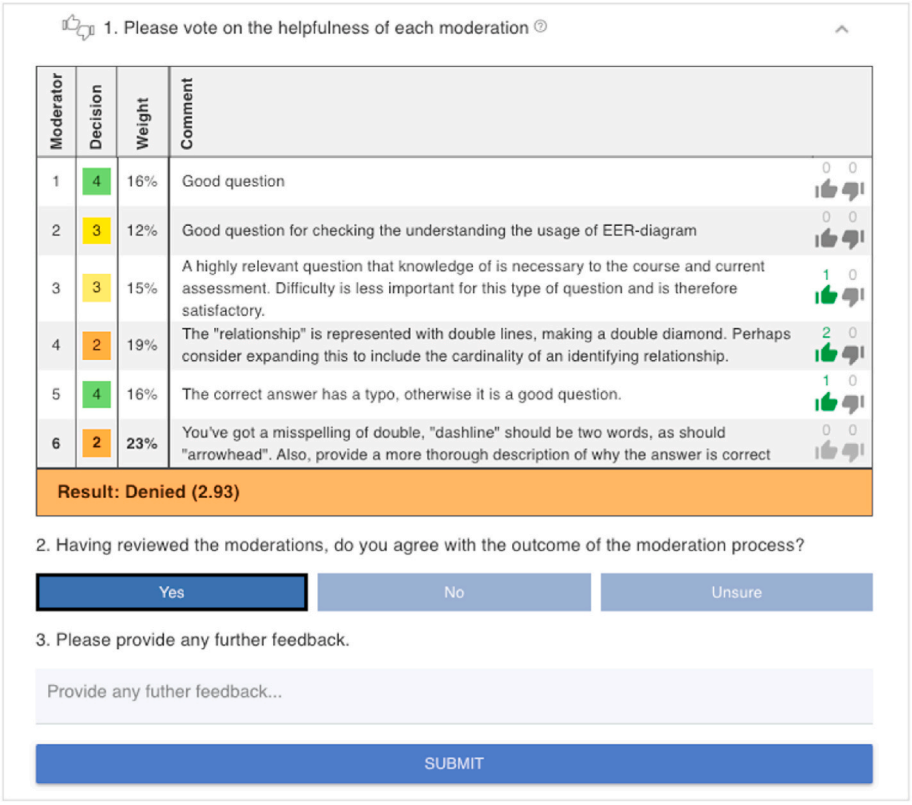


Fig. 5. Explainable consensus approach.

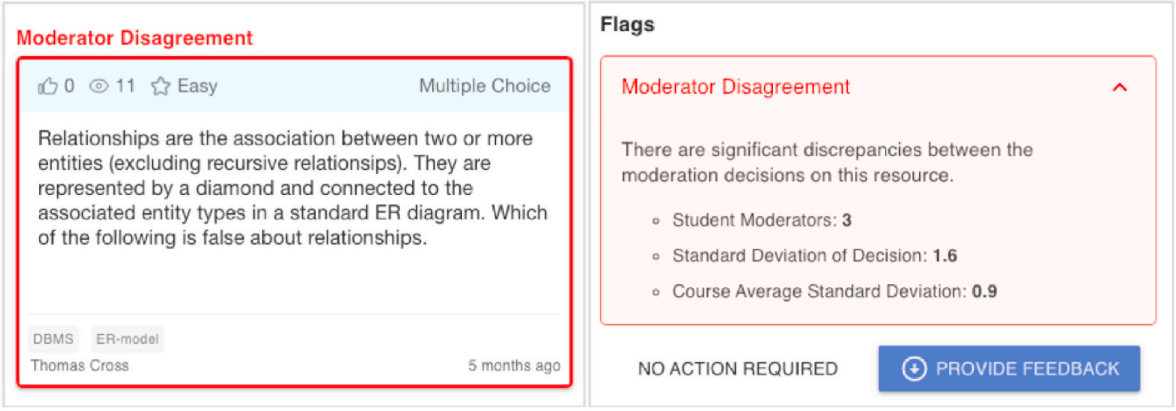


Fig. 6. Explainable spot-checking.

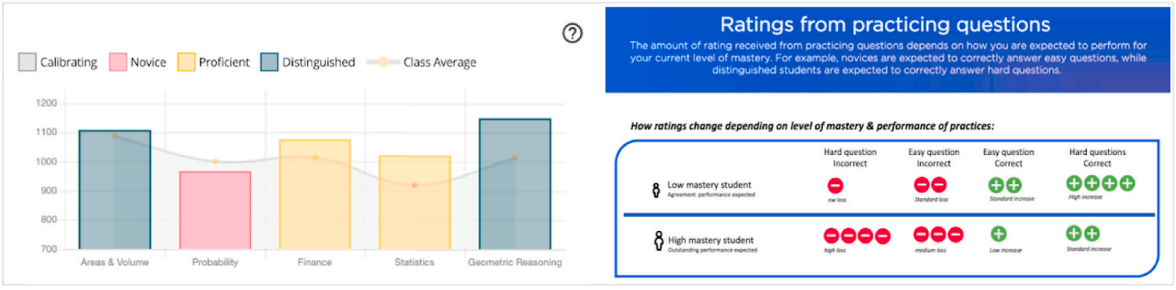


Fig. 7. Explainable learner modelling.

flagged due to moderator disagreement and the relative points of comparison provided to support this decision.

4. *Explainable and open learner model.* RiPPLE employs an extension of the *Elo rating system* to estimate a student's competence in each topic (knowledge component) of a course (Abdi, Khosravi, Sadiq, & Gasevic, 2019) (Other open learner models based on the Elo (Abdi, Khosravi, & Sadiq, 2020) Clicko (Abdi, Khosravi, & Sadiq, 2021) ratings have also been used in the system). The left side of Fig. 7 shows this learner model visualised as a bar chart. The colour of the bars categorises competencies into three levels: novice, proficient, and distinguished. "Calibrating" is used when less than a threshold number of resources are answered on a topic. The model uses an *example based* explanation approach to show the average competency of the entire cohort over each knowledge unit using a line graph. The right side of the figure shows RiPPLE's explanation of the model for students. The explanation can be accessed by clicking on the question mark on the top right hand side of the model.
5. *Explainable resource recommendations.* The adaptive engine of RiPPLE makes use of the learner model to recommend resources at the right level of difficulty for each student. In particular, it recommends easier questions on topics where students are developing mastery and harder questions on topics where the student has already demonstrated mastery. Fig. 8 demonstrates how RiPPLE has included both the recommendation and the learner model on the same page to help with transparency of how resources are recommended to students. Results from a randomised controlled experiment suggest that complementing the recommender system with the learner model can have a positive effect on student engagement and their perception of the effectiveness of the system (Abdi et al., 2020).

4.1.4. XAI designs

RiPPLE has taken a *co-design* approach of partnering with students and instructors across the conceptualisation, development, validation and deployment phases. The initial conceptualisation was done with eight academics with diverse disciplinary backgrounds including computer science, engineering, medicine, pharmacy and education and one member from the central IT systems and support of the university. During the development phase, the RiPPLE team worked closely with academics who trialled early pilots of the platform and a group of four paid student partners who provided advice and helped with usability studies. The validation was based on field studies in close partnership with instructors and students who have given consent. As an example, Abdi et al. (2020) explored the benefits of complementing recommender systems with explainable learner models. Finally, since deployment, frequent short surveys through the platform and interviews with students and instructors have been conducted to better understand the *user experience* and ensure that the stakeholders' concerns and feedback are heard and considered.

4.1.5. Potential pitfalls

By and large, the attempts taken to employ XAI in RiPPLE have been well received; however, there have also been pitfalls and side effects in terms of promoting misbehaviour that are worth discussing. (1) Most students have reported that the availability of the open learner model increased their engagement with RiPPLE as it motivated them to improve their competency. However, feedback from some students revealed that the model can also act as a source of disengagement. In particular, some students indicated that they tend not to attempt questions if they are unsure of the answer to avoid 'taking a hit' in their rating. (2) In the case of flagging similar feedback submissions, instead of trying to address the issue, a strategy taken by some students was to try to *game the system* by making minor changes and resubmitting with the hope of no longer being flagged.

4.2. FUMA a framework for user modeling and adaptation

4.2.1. Overview

FUMA is a framework for modelling and supporting students in learning for exploratory, open-ended learning environments (OLE), such as interactive simulations, educational games and MOOCS (Conati et al., 2021). FUMA uses a range of machine learning models to discover classes of exploratory learning behaviours. These behaviours can then be used to classify users based on their learning needs and to generate personalised hints in real-time that guide the student towards a more effective usage of the tools and affordances available in the OLE.

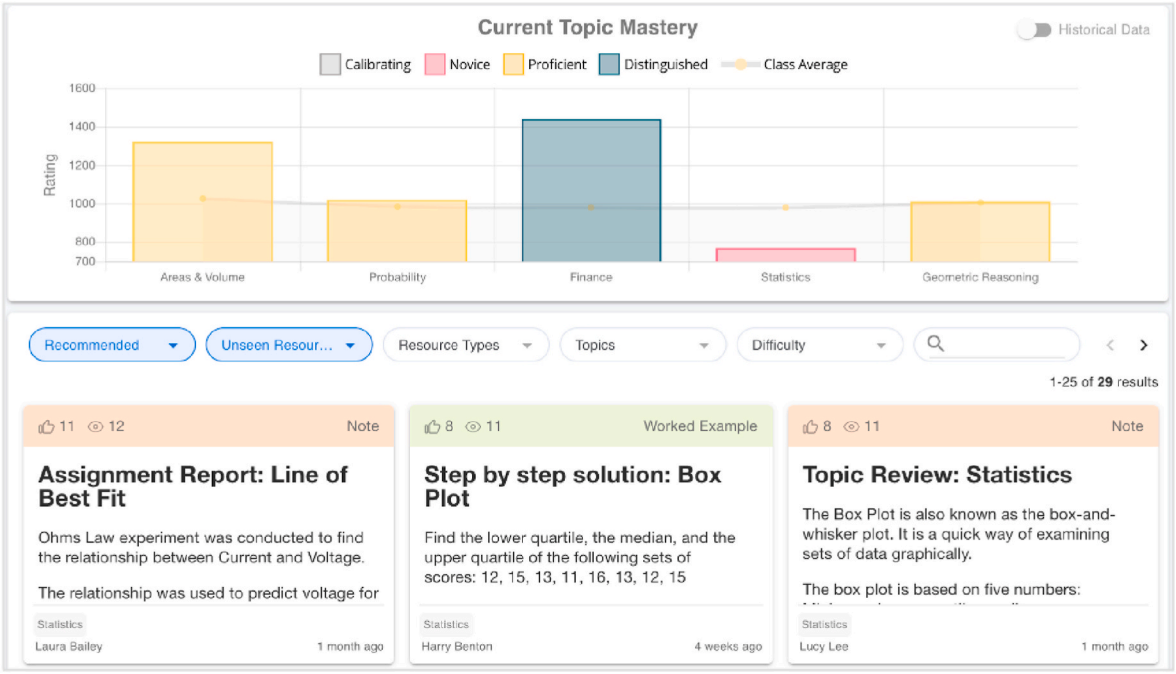


Fig. 8. Explainable recommendation.

4.2.2. Stakeholders and benefits

The primary stakeholders for this technology are *students*, especially those who need guidance during less structured educational activities. Various conducted studies (Fratamico, Conati, Kardan, & Roll, 2017; Kardan & Conati, 2015; Lallé & Conati, 2020) have shown that FUMA can identify clusters of user behaviours that map onto different learning outcomes, and predict when a new student is not learning well early on during the interaction. FUMA has been shown to foster better student learning (Kardan & Conati, 2015) and higher trust in hints (Conati et al., 2021). A second category of stakeholders are *instructors* who can use FUMA for *sensemaking* and to gain a better understanding of how the students learn (or not) from the environment (Conati et al., 2021).

4.2.3. Approaches and models

FUMA uses unsupervised *clustering* and *association rule mining* on existing data of students interacting with a target OLE to discover classes of exploratory behaviours conducive or detrimental to learning. Supervised machine learning is applied to the resulting clusters and association rules to build *classifiers* that predict in real-time whether a student is learning from the exploratory interaction and, if not, what behaviours are responsible for this outcome. These behaviours can then be used to generate personalised hints that guide the student towards a more effective usage of the tools and affordances available in the OLE. Fig. 9 includes detailed steps for each of these phases, to give a sense of the depth of the AI mechanisms embedded in FUMA.

To ascertain if the complex AI underlying the FUMA-driven hints can be externalised to students to further increase the hints uptake and effectiveness, an explanation interface was implemented to convey to the students the motivations (why) and processes used (how) for each of the hints they receive. These explanations aim to help users gain a *globally self-explaining* understanding of the AI driving the hints, as well as a *locally self-explaining* understanding of the specific hints generated. The explanation interface is structured around three tabs, each providing a self-contained, incremental part of the explanation for a given hint, as shown in Fig. 10 (A-C). The states in Fig. 9 are used to justify specific aspects of the rationale for hint computation (why explanations) and the processes to explain how some of the relevant algorithm components work. These tabs display the following three why explanations: 1) "Why am I delivered this hint?" 2) "Why am I predicted to be lower learning?", and 3) "Why are the rules used for classification?". In addition, for the second of these why explanations (Fig. 10 (B)), the user can access more details on how three specific aspects were computed (Fig. 10 (D–F)): "How was this score computed?", "How was this specific hint chosen?", and "How was my hint's rank calculated?" (see Fig. 11).

4.2.4. XAI designs

FUMA takes a user-centred design approach for implementing the explanations of the hints, and follows three guiding principles from (Kulesza et al., 2015), aiming to make the explanations be iterative, namely accessible at different levels of detail, sound, and not overwhelming. Determining how to convey coherent clear and non-overwhelming information on the elements of the FUMA model was supported by a rigorous process of iterative design and pilot evaluations, which led to the development of the interfaces presented in Fig. 10. Essentially, there is a trade-off between having complete explanations and explanations that are not overwhelming. Enabling iterative access to the explanations, as shown in the figure, is an important means to achieve this trade-off, and it is the criterion we used for the explanation functionality.

4.2.5. Potential pitfalls

Although the association rules learned by the FUMA student models have been shown to have a degree of inherent interpretability, they can

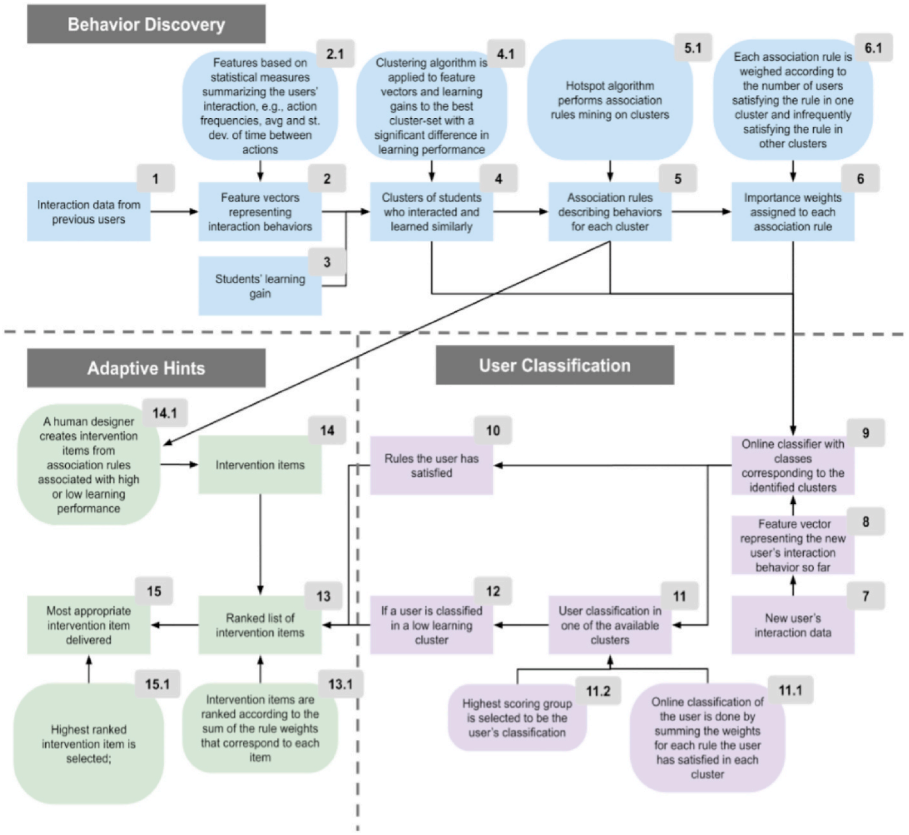


Fig. 9. User Modeling Framework broken down into three phases: Behavior Discovery, User Classification, and Adaptive Hints; rectangular nodes represent inputs and states, oval nodes represent processes.

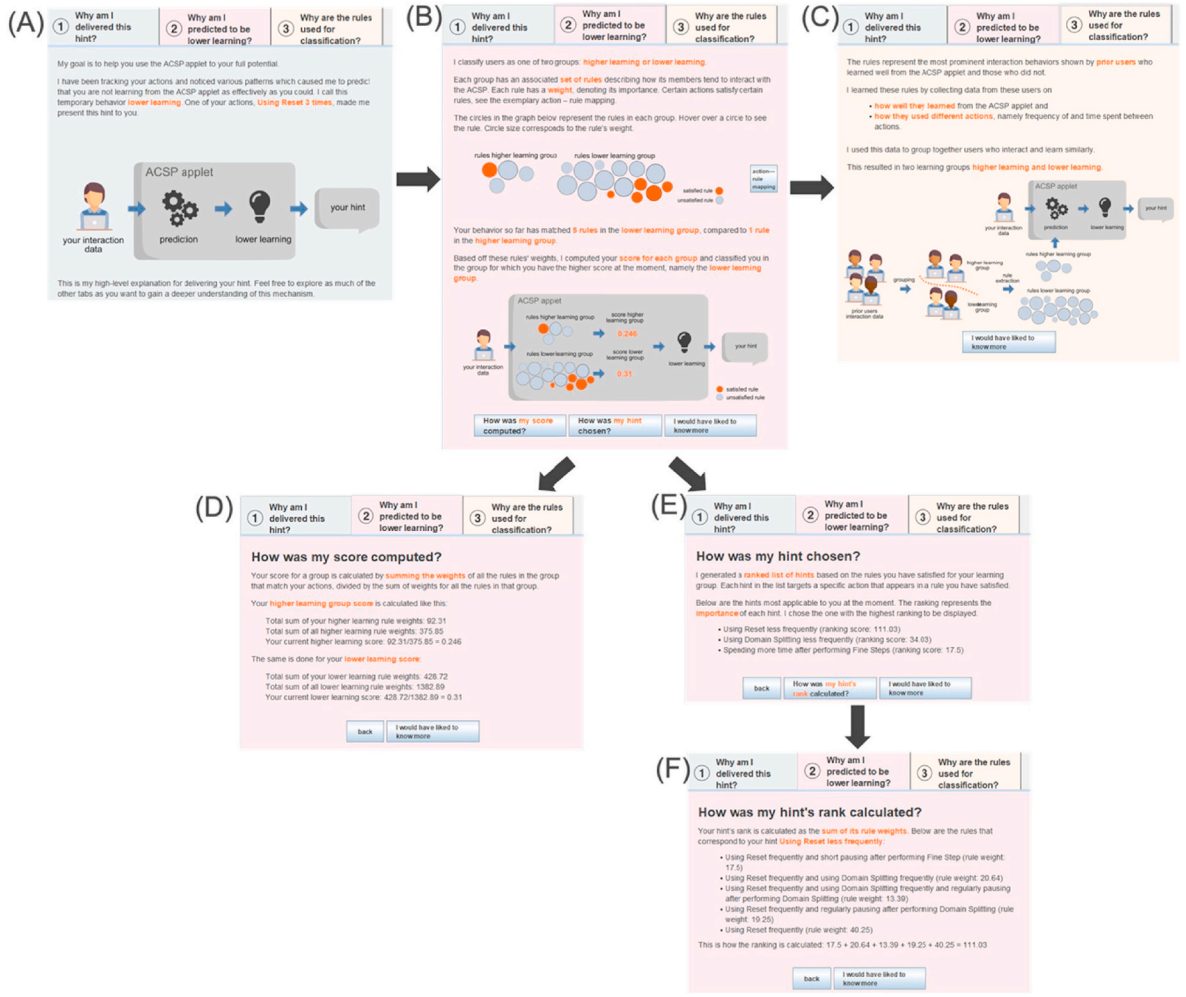


Fig. 10. Flow chart of explanation navigation.

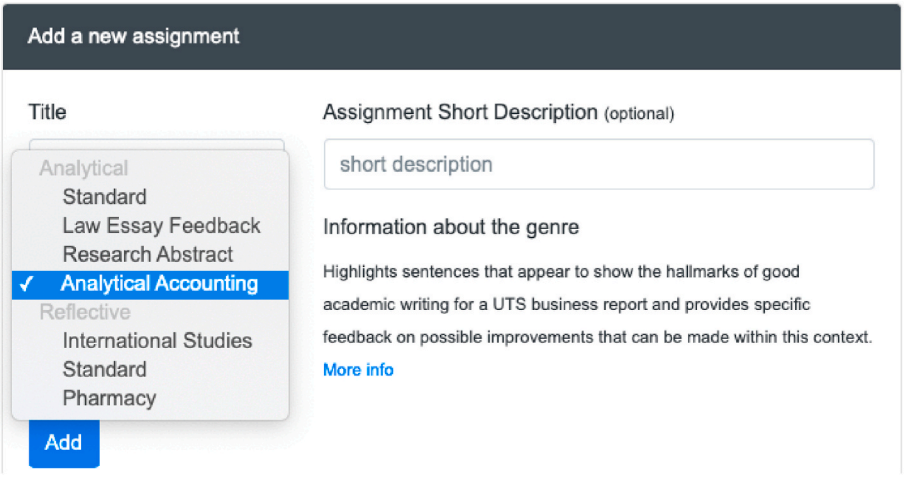


Fig. 11. Sample AcaWriter 'add a new assignment' screenshot; the tool is designed to facilitate instructors in aligning their tasks with genre-based feedback. Figure reproduced from Knight, Shibani, et al. (2020).

become exceedingly complex when dealing with sophisticated OLE that support a large set of actions. Thus, it is important to investigate, for each FUMA application, what the minimal set of features is that can capture relevant student behaviours and provide acceptable model accuracy without hindering rule interpretability. Additionally, a formal evaluation of the explanation interface shows that whether the explanations improve student learning depends to some extent on a student's level of conscientiousness, a personality trait defining one's tendency to follow rules and instructions, where students with low conscientiousness benefit from having the explanations whereas their high-level

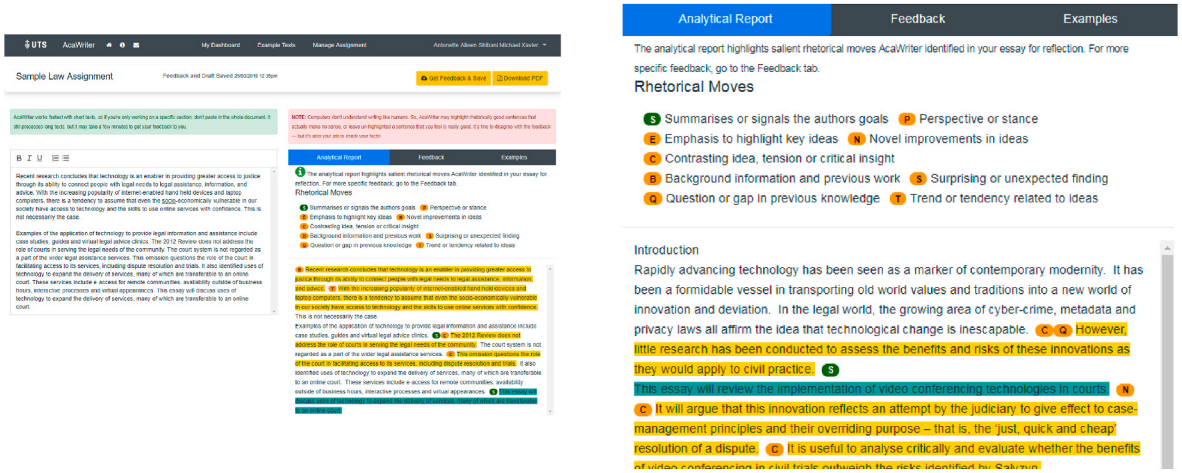


Fig. 12. Sample analytical report on a law essay in AcaWriter highlighting rhetorical moves in the writing. Left image shows the whole screen layout with the editor and feedback; right image shows the detailed feedback. Figures reproduced from Knight, Shibani, et al. (2020).

counterparts do not (Conati et al., 2021). These findings suggest personalizing the explanation functionality so that it proactively encourages users who are known to have low contentiousness or high reading proficiency to access explanations. They also uncover the need to investigate how to reduce the possibly negative effects of explanations on users with high contentiousness and low reading proficiency, for instance by discouraging explanation access or by understanding how to make them useful for these users.

4.3. AcaWriter

4.3.1. Overview

Writing Analytics tools apply analytic techniques drawing on process and sequence mining, natural language processing, and machine learning approaches to understand and provide feedback on features of the writing (process or product) to support that writing (see, e.g. Buckingham Shum, Knight, McNamara, Allen, Betik and Crossley, 2016). The focus of these tools has tended to be on formative feedback in tasks that do not have pre-created (and validated) questions, building on a longer body of work in developing Automated Essay Scoring (AES) systems for automated summative assessment and Automated Writing Evaluation (AWE) using similar approaches for formative purposes (Warschauer & Grimes, 2008) and intelligent tutoring systems that guide students through strategies to address well-defined constructed response writing tasks (Roscoe & McNamara, 2013). AcaWriter is a *theory driven* writing analytics tool that provides feedback primarily to *university students*, on the rhetorical structures in either their academic scholarly, or reflective, writing (Knight, Shibani, et al., 2020).

4.3.2. Stakeholders and benefits

The AcaWriter tool has primarily been targeted at university level writing, through both integration into specific courses (e.g., Law and Accounting for the analytical parser, and Pharmacy for the reflective parser), wider rollout to all students and supporting materials for academics to implement the tool in their own context, and a higher-degree research (HDR) version both as a standalone and integrated into an online course for learning how to write an abstract (Abel, Kitto, Knight, & Buckingham Shum, 2018). The AcaWriter tool has been designed to provide low stakes *formative feedback*, in which the user is engaged in tasks to develop their understanding of their subject, including through their writing and the incorporation of argumentative structures – or rhetorical moves – into this writing.

4.3.3. Approaches and models

Feedback is provided using a rule-based natural language processing system, which identifies syntactic relations between rhetorical concepts ('in contast', 'previous research', etc.) and key terms or topics in the writing. The feedback provided is coupled with a close alignment to the task's *learning design*, in order to *augment* wider assessment structures with this formative feedback (Knight, 2020; Knight, Shibani, & Buckingham Shum, 2018). It is through this alignment that *explanation* is drawn. Elements of the feedback provided are editable for the particular learning context (Shibani, Knight, & Buckingham Shum, 2019), with materials available to guide instructors in aligning the feedback to their assignments. The intent is to provide methods to align feedback (or explanations) to learning contexts through design that encodes expectations of what is to be learned and structures - including XAI - to support this learning (Knight, Gibson, & Shibani, 2020).

The implementation approach for AcaWriter has aligned learning design and analytics. This implementation is evaluated by operationalising what it would mean for the feedback to have impact, or to put it another way, for the explanation to achieve understanding in the student. For example, we have investigated whether revisions to the text introduce particular target features. At its simplest, this has involved analysing drafts and redrafts of the same text. In other contexts a guided process was implemented that builds up the students' knowledge of the constituent parts of the feedback and its relation to their assessment tasks, targeting those assessment tasks and the students evaluative judgement capacities (Shibani, Knight, Buckingham Shum, & Ryan, 2017).

4.3.4. XAI designs

The feedback information provided through the tool takes several forms:

1. *Annotation at the sentence level*, with iconic labels and highlighting indicating the type of sentence identified by the tool (and by extension, unlabelled sentences show that no rhetorical move is detected);
2. Through a 'feedback' tab that provides *document level feedback commentary* regarding issues such as a typical sequence of moves, or the presence of particular moves in introductory sections;
3. A third tab offers reminders of *how the moves map to the writing assignment rubric*, plus *examples* of the kinds of moves students can make in their writing; these can be customised to the writing task, although there is potential to automate selection of examples that match the genre of the text provided (Knight, Abel, et al., 2020).

The tool includes a number of student guides regarding its effective use, as well as a notice that users should "Remember, AcaWriter does not

really understand your writing the way people do […] If you think it got it wrong, that's fine – now you're thinking about more than spelling, grammar, and plagiarism.", or, as the webpage introduction says "UTS isn't here to tell you what to think, but to help you learn how to think. Similarly, AcaWriter won't tell you what to write, but will help you learn how to say it in the most rigorous, effective way." These aim to build *trust*, *agency* and *AI literacy* in understanding the uses and pitfalls of the tool.

In summary, the AcaWriter tool instantiates XAI through implementing feedback on units of analysis — sentences and documents — in the context of design for learning that couples feedback and formative learning tasks, and resources that seek to explain the features on which feedback is provided (while also acknowledging the inherent limitations of such automated analyses).

AcaWriter feedback is *theory driven*, underpinned by literature in writing theory, which has been instantiated in a number of broadly comparable tools (see Knight, Abel, et al., 2020) in this case into a rule-based NLP system. Early versions of a user interface onto the rhetorical move detection were evaluated through both student and instructor *user experience evaluation* and *co-design*, with subsequent evaluation involving more formal testing with quasi-experimental designs (Knight, Buckingham Shum, Ryan, Sándor, & Wang, 2018; Knight, Shibani, et al., 2020). The tool was scaled through piloting in one discipline context, and adapting these learning designs to other discipline contexts. In these evaluations, we demonstrate that the tool adds value to perceived 'usefulness' of *formative writing tasks*, and that students with access to the tool do indeed make more revisions and include more rhetorical moves (Knight, Buckingham Shum, Ryan, Sándor, & Wang, 2018; 2020c).

User-centred co-design with academics has been key in this implementation and disciplinary contextualisation (Knight, Gibson, & Shibani, 2020; Shibani, Knight, & Buckingham Shum, 2020). The tool is now deployed at scale, with students and instructors able to access the tool and in the academic case create their own 'assignment' code for students to receive one of the pre-created feedback options. This work points to the potential of these tools, with students who receive the feedback more likely to develop draft texts, and to incorporate rhetorical moves (Knight, Shibani, et al., 2020).

4.3.5. Potential pitfalls

While the tool is designed to augment effective design for learning, its implementation raises a number of pitfalls. Chief among these is that for *explanation* to be achieved, users (primarily students) must of course understand the characteristics of the features or moves for which feedback is received. As a result, the tool and effectiveness of explanation cannot be considered outside the context of use. In those contexts, the tool also provides feedback only on a narrow range of features (rhetorical structures) for particular types of writing. In addition, it does so imperfectly, particularly where other features (such as spelling and grammar) may interfere in the text processing. These imperfections may provide some learning opportunity (Kitto, Buckingham Shum, & Gibson, 2018; Knight, Gibson, & Shibani, 2020) and the interface itself does encourage users to reflect on the feedback to build their own judgements. Nevertheless, opportunity costs (time spent on x, rather than y) may be present in any system. These pitfalls highlight the importance of human-in-the-loop decision making, and task design for learning in implementing learning analytics tools. The importance of empirically studying students' use of automated feedback cannot be over-stated; it is only by seeing their engagement (or lack thereof) that we can tell if the feedback and explanations are actioned appropriately. The observation that students needed scaffolding in their use of AcaWriter has led, for instance, to the creation of online tutorials that prepare them (Abel, 2022), and activities that promote deeper critical reflection on the feedback (Shibani, Knight, & Buckingham Shum, 2022).

4.4. Data storytelling through teamwork analytics in healthcare

4.4.1. Overview

Immersive, high-fidelity simulations are a common pedagogical approach used in healthcare to support students' and practitioners' learning. In these, participants are posed with an authentic challenge for them to decide on the potential course of action under time pressure. Simulations are commonly conducted in learning spaces instrumented with various physical and digital devices that resemble those available in real hospital rooms. The simulations are followed by a debrief in which participants engage in deep reflection about errors they may have made, areas of improvement and how stressed they may have felt during the simulation (Palominos, Levett-Jones, Power, & Martinez-Maldonado, 2019). In the debriefs, an educator leads the reflection based on direct observations or video-recordings of the simulation. However, the educator and learners commonly do not have objective evidence to discuss as replaying the video-recordings can be time consuming and unpractical. This has motivated the use of multimodal learning analytics to augment the data capture capabilities of the simulation rooms to identify salient aspects of team activity and make these available during the debrief (Martinez-Maldonado et al., 2017). The data is collected using indoor positioning sensors, microphones, physiological wristbands and an observation console (Martinez-Maldonado, Echeverria, Fernandez Nieto, & Buckingham Shum, 2020).

4.4.2. Stakeholders and benefits

The multimodal analytics are aimed at supporting *educators* and *undergraduate nursing students* enrolled in clinical units. These students do not have the time to delve into the analysis of their own data because they also need to reflect on the clinical case they had to address. Based on co-design sessions (Prieto-Alvarez, Martinez-Maldonado, & Shum, 2018), it was identified that both kinds of users would benefit from explanatory visualisations by enabling them to reflect on evidence automatically captured during the group activities. This can increase *accountability* of students who can demonstrate how their actions reflect the application of their clinical knowledge. Exposing the algorithms used to generate the visual interfaces aims at enhancing *trust* in the system and ensure the final users have the *agency* to make their own teaching and learning decisions.

4.4.3. Approaches and models

The concept of data storytelling was adopted to use narrative to communicate insights extracted from the multimodal data. The idea of data storytelling has been suggested, both by the general XAI research community (El-Assady et al., 2019) and the emerging XAI sub-community within learning analytics (De Laet et al., 2018), as a way to automatically structure explanations from data analysis and to emphasise some data points that are relevant to particular stakeholders or users. Here we discuss two ways used to incorporate data storytelling in the multimodal learning analytics interfaces for healthcare simulation.

1. *Explaining key insights from multimodal data.* A number of data storytelling interfaces were created to support *sensemaking* during the debrief. The multimodal data captured during the teamwork sessions can be complex. For example, actions can be logged by the digital devices located in the room and also by an observer. Even though these logged actions are discrete and easy to recognise by an educator or learners (e.g., checking vital signs or providing medication to the patient), sensemaking challenges emerge if they try to focus on assessing such actions for reflection. For example, it is not enough for learners just to 'see' their actions but they need to assess these actions in light of domain knowledge, guidelines or regulations, to understand if and how they can improve their practice. In other words, learners can get lost in the interpretation of such actions (see Fig. 13).

In response, the data storytellling approach adopted in this case consisted in providing layered *local explanations* based on the learning goals of the simulation. A data storytelling layer is created for each learning goal, which consists in combination of data visualisations and textual narratives. Fig. 14 (left) a presents an example of a data storytelling layer that emphasises only certain data points that are relevant to a learning goal (i.e., see Fig. 14, A). In this case, one of the learning goals is for students to reflect whether they checked the vital signs of the patient every 10 min after complaining from chest pain and clearly deteriorating (B). The interface is then improved by adding narrative that *explains* why certain data points are relevant (A) or to indicate the absence of critical data points (C). An explanatory title describes, in lay language, the main take-away message from that data layer. If the goal is to reflect on stress or arousal as automatically captured by the physiological wristband, the students or the educator can press a button (E) and labels appear on top of the visualisation to indicate in text if mild, high or very high arousal was detected by the electrodermal activity sensor (F), which can be indicative of high stress or cognitive load. At the top of the visualisation, a short summary (D) is provided, indicating which roles in the team experienced more arousal during the simulation. The purpose is to spark discussion by providing brief but clear explanations from the data analysis in ways that users without multimodal data analysis training can make sense of. Inroads are currently being made to also create layers that communicate insights from the analysis of x-y indoor positioning traces (Fernandez-Nieto, Martinez-Maldonado, et al., 2021).

2. *Explaining the algorithm used for narrative visualisation. Rule-based algorithms* were used to generate the storytelling layers presented above. These can receive parameters as input or analyse sensor data and compare it across the whole dataset. For example, to assess whether an action was performed in the right time frame, a parameter can reflect official indicators according to healthcare guidelines. For the case of physiological data, the data streams from a student are compared across the arousal levels experienced by other students in the past and therefore determine the level of arousal for the present student. The data stories hide the underlying algorithm during the reflection.

Fig. 15 illustrates how these rules have been opened for participants to provide them with *global explanations* about how the different visual and narrative elements appear on the interface. For example, Fig. 15 (A) show how the narrative that appears at the top of the data storytelling depends on a parameter that reflect one healthcare guideline ("Students should check the vital signs of the patient every 2 min"). Similar parameters are used to automatically add the orange shaded areas to indicate that certain actions were missing during the simulation (B). A different rule assessed the order of certain actions (e.g., indicating that calling to the doctor should happen before a key event).

An evaluation of manipulability and transparency of the teamwork analytics dashboard showed that educators wanted to gain control over the way the data stories are programmed to be able to trust in them (Martinez-Maldonado et al., 2020). They want to be aware of the parameters used to generate the stories for the purpose of compliance with

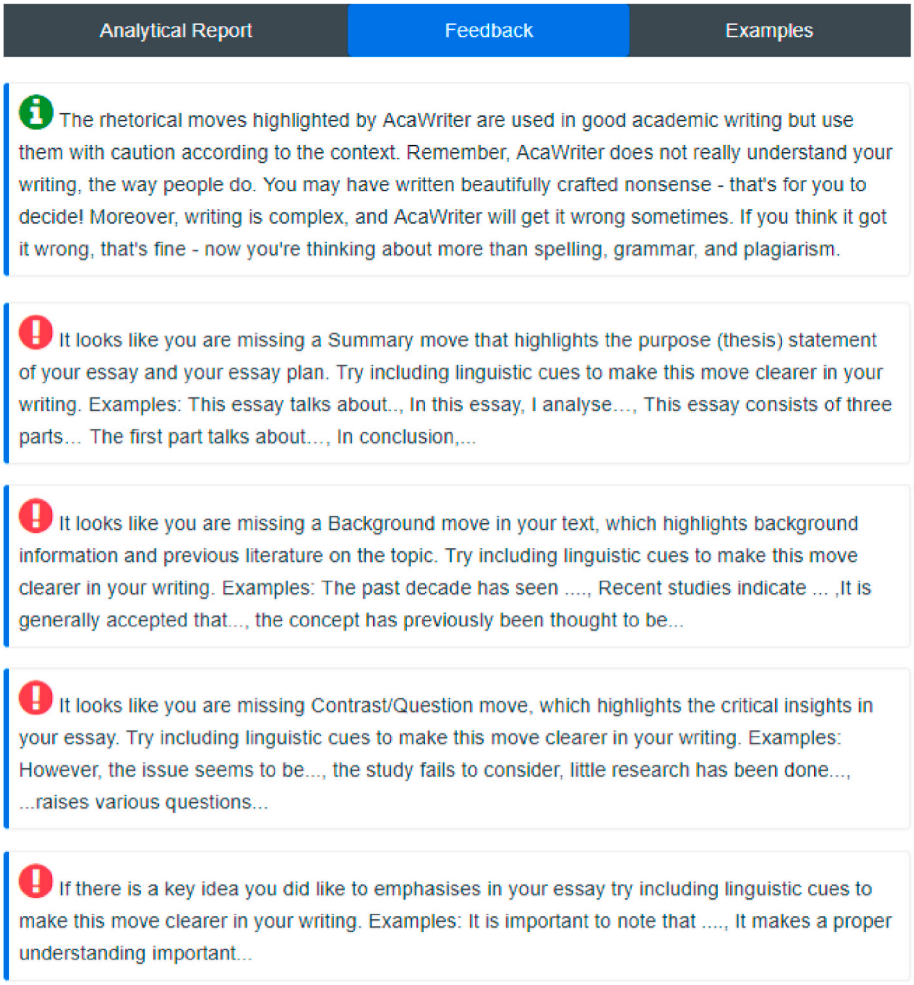


Fig. 13. Sample feedback messages for law from AcaWriter on missing rhetorical moves. Figures reproduced from Knight, Shibani, et al. (2020).

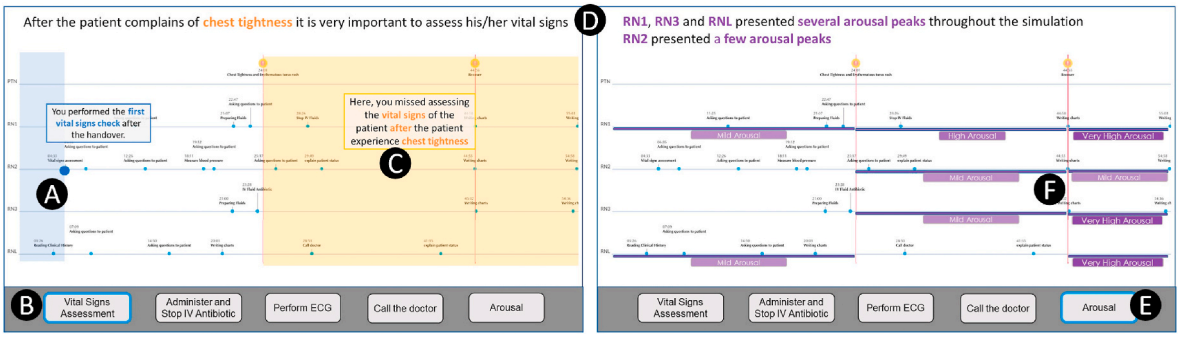


Fig. 14. Two example data storytelling layers explaining activity logs (left) and physiological data (right), adapted from (Fernandez-Nieto, Echeverria-Barzola, et al., 2021, pp. 1–15).

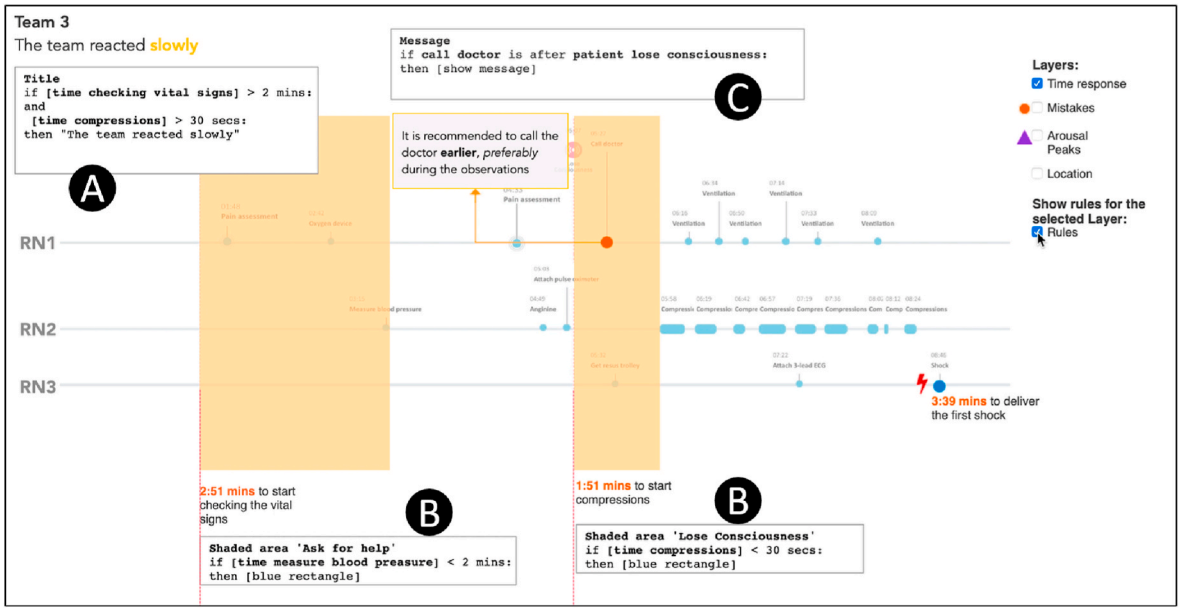


Fig. 15. Explaining the algorithm used for narrative visualisation, adapted from (Echeverria, Martinez-Maldonado, & Buckingham Shum, 2019).

accreditation goals, and be able to adapt/change these parameters according to the cohort of students (for example, by being more relaxed with first year students compared to final year students). By contrast, nursing students did not feel in the position to understand or change the parameters, but they appreciated being able to understand the intended goals of the task based on those parameters.

4.4.4. *XAI designs*

A bottom-up, human centred approach was followed in this case to co-create data stories with relevant stakeholders. A number of *co-design* techniques such as persona identification, user journeys and rapid prototyping were conducted in various cycles of design with educators and nursing students (Prieto-Alvarez, Martinez-Maldonado, & Shum, 2018). In fact, the intention behind this case is that the characteristics of the learning design drive the kinds of explanations that are offered by each data story and, therefore, the **UX design**. The narratives added to the visualisations are a reflection of the kind of feedback a teacher would provide to students and can be edited by them in design time. A new simulation would require the adaptation of the interface according to the learning goals, to emphasise the data points that are relevant to the task at hand. With effective authoring interfaces, the tool can become an explanatory learning analytics interface that augments the kind of feedback that can be provided by the educator during the debrief.

4.4.5. *Potential pitfalls*

The tool has been well received by educators and learners. They can see the potential of augmenting the evidence they can reflect upon and also how the same evidence could be used more widely to improve clinical practice in authentic scenarios. However, the data storytelling approach can open up an interesting debate around *incomplete explanations*. This approach addressed the problem of data abundance by extracting what is important. However, what if important data and insights get unintentionally hidden? Since teamwork simulation is a very complex activity, many other salient aspects may not be covered by the current data capture capabilities. The data stories would emphasise the data relevant to the initial pedagogical goals. A derived potential issue is that *misbehaviour* could be inadvertently promoted if feedback and reflection is only focused on the data that can be automatically captured. For example, sensors may be able to capture if certain medical procedure was performed but not necessarily how it was performed. This could even encourage students to game the system. Keeping the educator in the loop is therefore critical to guide and complement the use of evidence for promoting reflective practices. This naturally occurs for the case of teacher-led debriefs in health care. However, if data storytelling is applied in interfaces used by students alone, these potential pitfalls should be carefully considered and mitigated.

5. Opportunities, challenges and future research needs

This section discusses opportunities, challenges and future research needs for advancing the effective incorporation of XAI in education.

5.1. Actionable explanations

Explainable AI has often been associated with the aim to provide actionable explanations for stakeholders with the growing importance of big data in education and learning analytics. While actionable explanations are commonly related to data-informed decision making in education, a formal definition of actionable explanations as *actionable insights* has been proposed: "should be interpreted as data that allows a corrective procedure, or feedback loop, to be established for a set of actions" (Jørnø & Gynther, 2018, p. 198). For explainable AI in education this means that the emphasis should not only be on explaining the inner workings of an algorithm and how certain results are computed, but that there should be a purposeful design consideration of an AI system that can guide the user to take a certain action (Winne, 2021). As shown in the current paper and the literature, certain classes of AIED system seek to provide *actionable* explanations to promote learning, such as provision of formative feedback (Knight, 2020), triggering reflection and metacognitive monitoring, explainable recommendation of learning resources to engage with (Abdi et al., 2020; Barria-Pineda et al., 2021), (Palominos et al., 2019), revision of course content (Ali, Hatala, Gašević, & Jovanović, 2012), assess algorithmic bias and fairness (Baker & Hawn, 2021; Kizilcec and Lee, In Press), optimal use of instructors time to review student work (Darvishi et al., 2021) or provide support for those who need it the most (Khosravi, Shabaninejad, et al., 2021), or more generally taking a wide range of pedagogical actions (Pardo et al., 2016).

Future research on XAI to provide actionable insights should address several critical challenges. First, a recent systematic review of dashboards (Matcha, Gašević, Pardo et al., 2020), a form of explainable AI systems in learning analytics, showed significant limitations related to data used along with challenges related to design and evaluation of systems. No study on dashboards had data about the key component of self-regulated learning —- cognitive and metacognitive tactics and strategies. As Winne (2021) argues, the absence of data about cognitive and metacognitive tactics and strategies does not allow for providing guidance to learners about how to approach their learning. That is, even if the algorithms that are used can offer a great deal of explainability, we should also work on mechanisms for purposeful collection of data and measurement of constructs on which actionable insights are to be provided (Gašević, Dawson, & Siemens, 2015; Winne, 2020).

5.2. Personalised explanations

Existing research on XAI suggests that having AI systems explain their inner workings to their end users can help foster transparency, interpretability, and trust. However, there are also results suggesting that such explanations are not always valued by or beneficial for all users. For instance, in Section 4.2 we discussed findings showing how some user characteristics modulate the effect of explanations on both student's learning and perception of the adaptive hints available in the ACSP tutoring system, providing insights on how personalisation might bolster explanation effectiveness in this context. These results are in line with previous work that showed the impact of user differences on explanation effectiveness in domains different to education (e.g. (Millecamp, Htun, Conati and Verbert, 2020, 2019; Kleinerman, Rosenfeld, & Kraus, 2018; Kouki, Schaffer, Pujara, O'Donovan, & Getoor, 2019)), and calls for research further exploring the value of personalised XAI in education. The vision is that of a personalised XAI, endowing AI-driven pedagogical systems with the ability to understand when, and how to provide explanations to their end users (e.g., students, teachers, parents) (Conati et al., 2021). Specifically, it is important to identify: *what types of explanations different end users need* (e.g., why or how the system generated a specific outcome); *how the explanations should be delivered* in order to be informative and non-intrusive; and whether these factors might depend on *individual differences* (e.g., long-term abilities and traits, short-term cognitive and affective states, preferences) in order to enable delivery of personalised explanations that accounts for user diversity and bolster inclusiveness of outcomes.

5.3. Human-centred AI design in education

The process for the design of AI and learning analytics systems is underexplored in spite of promising results that have been demonstrated in the recent literature. A recent systematic literature review (Bodily & Verbert, 2017) showed that only about one in ten studies reported a needs analysis before designing a learning analytic dashboard or recommender system. An important research direction is to work on methods that will build design partnerships between developers, researchers and the stakeholders who are typically the people who are ultimately expected to use the tools. Particular attention should be paid to stakeholders from underrepresented groups to ensure they have a meaningful voice (Buckingham Shum et al., 2019; Dollinger et al., 2019). Design considerations should also focus on the way we envision interaction of stakeholders with explainable AI as another form of actionability. Opening learner models can also be a form of invitation for learners to 'push back' against some of information that is incomplete or doubtful (Bull, 2016), and thus contribute to validation of an AI system, while promoting learning engagement.

5.4. Evaluating explanations

We have already noted the importance of studying XAI in use, to see whether in fact the explanations are *understood*, and *appropriately actioned*. Without such evidence, no claim to have designed effective XAI can be made. HCI offers diverse range of techniques can be used to gather such evidence, each of which brings its own strengths and weaknesses (Olson & Kellogg, 2014). These include usability laboratory studies to gain insight from learner reflections in combination with high definition interaction data; activity log analysis to test at scale whether learner actions appear to have addressed feedback and explanations; retrospective cued recall that uses images/video to prompt users to share their thinking; and qualitative self-report data on the level of trustworthiness learners/educators have in the tool/infrastructure.

5.5. Towards trustworthy AI educational systems

Ultimately, we might state our goal to be the creation of *trustworthy*, *AI-augmented*, *sociotechnical* systems. Firstly, the emphasis on *trustworthiness* points to the fact that this is not the same as creating AIED that people trust —- after all, people place misguided trust in all sorts of technology. Secondly, the emphasis on *AI-augmented* draws attention to our belief that while it is certainly possible to learn some things as an isolated individual in front of a computer, the most effective, engaging forms of learning are relational, human activities, involving emotions, other learners, and a relationship in which you trust your teacher/coach or peers to both support and challenge you. AI augments this system. Connected to this, thirdly, *sociotechnical* systems reminds us that we need to frame AIED design not solely as the creation of a digital tool, but as the design of an overall educational system involving other people with different roles, working in an organisational contexts (e.g. a high school in an economically struggling suburb; an elite university; a training unit in a large multinational), in a society that may introduce further constraints (e.g. a national curriculum, certain kinds of teachers, assessment expectations, parental expectations). These contexts and constraints define the 'life-chances' of a given AIED tool. As so many educational technology cases have demonstrated over the years (Scanlon et al., 2013), failure to take these into account simply adds to the

graveyard of promising tools that never achieved sustained adoption.

This paper has begun to demonstrate that what counts as an 'adequate, trustworthy' *explanation* will vary drastically, depending on whether that person is an AIED researcher, a learning scientist, a school teacher or principal, a parent or a student. Possibly only the researcher would know what they were even looking at if permitted to look 'inside the algorithmic black box'. An important line of empirical work is therefore opening up, to clarify what stakeholders consider to be trustworthy explanations, and beyond that of course, what it takes to trust the software as a whole, which will turn on all its behaviour, not just its explanations.

However, as noted above, through no fault of their own, people may be misguided in placing trust in a system they only partially understand. We need as a field, therefore, to move beyond just the empirical question, and ask a more normative question: By what criteria *should* we (as educational and AI experts) declare an AIED system to be trustworthy? By analogy, this is the difference between answering the question, *Are people willing to fly on this plane?* and *Is this plane fit for purpose, as judged by engineering and other civil aviation standards?* The difficulty that AIED currently faces as a field, is that the societal 'trust infrastructure' that we have around mature engineering fields is simply not in place yet. While one key element to building trust is the software's capacity to explain its behaviour adequately to a given stakeholder, in fact, a chain of diverse arguments underpins the claim that the software is trustworthy; developing criteria to test the integrity of each link in that chain is the focus of one or more disciplines, which together establish a notion of overall 'system integrity' (Buckingham Shum, 2019).

6. Conclusion

Advancements in AI are impacting on most if not all sectors and education is no exception. Given that the development of AI tools and technologies is outpacing the social and even legal aspects of the implications of wide scale adoption, it is understandable that there is a degree of public mistrust. Explainable AI is a growing area of research wherein the concerns of fairness, accountability, transparency and ethics can be mitigated. In the educational setting, the importance and need for Explainable AI is further heightened due to the issues related to learner autonomy and agency, support for learner metacognitive processes and reflective processes and broader issues relating to authentic assessment, credentialing and academic integrity.

In this paper we have examined the multiple and complex aspects of Explainable AI in education. We have outlined a framework called XAI-ED with which educational AI tools can be studied, designed and developed. XAI-ED covers six fundamental aspects of XAI in education relating to stakeholders, benefits, approaches, models, designs and pitfalls. The comprehensive discussion of these six aspects allows for a detailed understanding of current state-of-art and open challenges. We have further anchored these six aspects into four diverse case studies relating to leanersourced adaptive systems, open-ended learning environments, writing analytics tools, team based learning to support knowledge transfer. There is a wealth of experiential knowledge embedded in these case studies, which we have deliberated upon to expose a number of XAI opportunities and challenges. Finally, we have used the analysis of the six fundamental aspects and experiences from the case studies to synthesise an agenda for future research for XAI in education. We highlight five areas of particular importance, namely the need for development of actionable and personalised explanations, further incorporation of human-centred design in development of educational tools, rigorous evaluation of the impact of incorporating XAI in education and ultimately advancing towards development of trustworthy AI educational systems.

We conclude that XAI is a critical element that is indispensable to fully avail the opportunities and benefits that AIED systems present for education, human capital development and learning sciences We call upon the research and practitioner community in AIED systems to review, critique, champion and advance the opportunities, challenges and future research needs we have outlined in this paper.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported through multiple granting agencies including: the Australian Research Council's Industrial Transformation Training Centre for Information Resilience (CIRES) (IC200100022), and Discovery Projects (DP210100060 and DP220101209), Jacobs Foundation Center for Learning and Living with AI (CELLA) and Connecting the EdTech Research Ecosystem (CERES), the National Science and Engineering Research Council of Canada (NSERC) and the Economic and Social Research Council of the United Kingdom (ES/S015701/1).

References

Abdi, S., Khosravi, H., & Sadiq, S. (2020). *Modelling learners in crowdsourcing educational systems, in: International Conference on ArtificialIntelligence in Education* (pp. 3–9). Springer.

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2019). A multivariate elo-based learner model for adaptive educational systems. In *Proceedings of the educational data mining conference* (pp. 462–467).

Abdi, S., Khosravi, H., & Sadiq, S. (2021). Modelling learners in adaptive educational systems: A multivariate glicko-based approach. In *LAK21:11th International Learning Analytics and Knowledge Conference* (pp. 497–503). New York, NY, USA: Association for Computing Machinery.

Abdi, S., Khosravi, H., Sadiq, S., & Gasevic, D. (2020). Complementing educational recommender systems with open learner models. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 360–365).

Abel, S. (2022). *Writing an abstract (online open access course*. University of Technology Sydney. URL: https://open.uts.edu.au/uts-open/study-area/communication-media/writing-an-abstract.

Abel, S., Kitto, K., Knight, S., & Buckingham Shum, S. (2018). Designing personalised, automated feedback to develop students' research writing skills. In *35th international conference on innovation, practice and research in the use of educational technologies in tertiary education (ASCILITE)* (pp. 15–24). ASCILITE, 00000.

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. *Handbook Research on Learning and Instruction*, 522–560.

Ali, L., Hatala, M., Gašević, D., & Jovanović, J. (2012). A qualitative evaluation of evolution of a learning analytics tool. *Computers & Education, 58*, 470–489.

Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E., & Berthouze, N. (2020). Evaluating saliency map explanations for convolutional neural networks: A user study. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 275–285).

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences, 4*, 167–207.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

Baker, R. S., De Carvalho, A., Raspat, J., Aleven, V., Corbett, A. T., & Koedinger, K. R. (2009). Educational software features that encourage and discourage "gaming the system". In *Proceedings of the 14th international conference on artificial intelligence in education* (pp. 475–482).

Baker, R. S., & Hawn, A. (2021). Algorithmic bias in education. *International Journal of Artificial Intelligence in Education*. https://doi.org/10.1007/s40593-021-00285-9

Barria-Pineda, J., Akhuseyinoglu, K., Želem-Ćelap, S., Brusilovsky, P., Milicevic, A. K., & Ivanovic, M. (2021). Explainable recommendations in a personalized programming practice system. In *International conference on artificial intelligence in education* (pp. 64–76). Springer.

Batane, T. (2010). Turning to turnitin to fight plagiarism among university students. *Journal of Educational Technology & Society, 13*, 1–12.

Berger, R., Rugen, L., Woodfin, L., & Education, E. (2014). *Leaders of their own learning: Transforming schools through student-engaged assessment*. John Wiley & Sons.

Biggs, J. B., & Collis, K. F. (2014). *Evaluating the quality of learning: The SOLO taxonomy (structure of the observed learning outcome)*. Academic Press.

Bloom, B. S. (1984). The 2 sigma problem: The search for methods of group instruction as effective as one-to-one tutoring. *Educational Researcher, 13*, 4–16 (Publisher: Sage Publications Sage CA: Thousand Oaks, CA).

Bloom, B. S., et al. (1956). *Taxonomy of educational objectives, 1* pp. 20–24). New York: McKay: Cognitive domain.

Bodily, R., & Verbert, K. (2017). Review of research on student-facing learning analytics dashboards and educational recommender systems. *IEEE Transactions on Learning Technologies, 10*, 405–418.

du Boulay, B. (2016). Recent meta-reviews and meta–analyses of AIED systems. *International Journal of Artificial Intelligence in Education, 26*, 536–537.

Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. ERIC.

Brusilovsky, P., Somyürek, S., Guerra, J., Hosseini, R., & Zadorozhny, V. (2015). The value of social: Comparing open student modeling and open social student modeling. In *International conference on user modeling, adaptation, and personalization* (pp. 44–55). Springer.

Buckingham Shum, S. (2019). Black box learning analytics? Beyond algorithmic transparency (video tutorial, learning analytics summer school, society for learning analytics research). URL: https://youtu.be/VJBt1qpPfXw.

Buckingham Shum, S., Ferguson, R., & Martinez-Maldonado, R. (2019). Human-centred learning analytics. *Journal of Learning Analytics, 6*, 1–9.

Buckingham Shum, S., Knight, S., McNamara, D., Allen Laura, K., Betik, D., & Crossley, S. (2016). Critical perspectives on writing analytics. In *6th ACM learning analytics and knowledge conference* (pp. 481–483). ACM. http://dl.acm.org/citation.cfm?id=2883854.

Bull, S. (2016). Negotiated learner modelling to maintain today's learner models. *Research and Practice in Technology Enhanced Learning, 11*, 1–29.

Bull, S. (2020). There are open learner models about. *IEEE Transactions on Learning Technologies*.

Bull, S., & Kay, J. (2013). Open learner models as drivers for metacognitive processes. In *International handbook of metacognition and learning technologies* (pp. 349–365). Springer.

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

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.

Chounta, I. A., Bardone, E., Raudsep, A., & Pedaste, M. (2021). Exploring teachers' perceptions of artificial intelligence as a tool to support their practice in Estonian k-12 education. *International Journal of Artificial Intelligence in Education*, 1–31.

Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. *Artificial Intelligence, 298*, Article 103503.

Cukurova, M., Zhou, Q., Spikol, D., & Landolfi, L. (2020). Modelling collaborative problem-solving competence with transparent learning analytics: Is video data enough?. In *Proceedings of the tenth international conference on learning analytics & knowledge* (pp. 270–275). New York, NY, USA: Association for Computing Machinery. URL: https://doi-org.ezproxy.lib.monash.edu.au/10.1145/3375462.3375484.

Darvishi, A., Khosravi, H., Abdi, S., Sadiq, S., & Gasevic, D. (2022a). Incorporating training, self-monitoring and ai-assistance to improve peerfeedback quality. In *9th International Conference on Learning at Scale, ACM* https://doi.org/10.1145/3491140.3528265

Darvishi, A., Khosravi, H., & Sadiq, S. (2020). Utilising learnersourcing to inform design loop adaptivity. In C. Alario-Hoyos, M. J. Rodríguez-Triana, M. Scheffel, I. Arnedillo-Sánchez, & S. M. Dennerlein (Eds.), *Addressing global challenges and quality education* (pp. 332–346). Cham: Springer International Publishing.

Darvishi, A., Khosravi, H., & Sadiq, S. (2021). Employing peer review to evaluate the quality of student generated content at scale: A trust propagation approach. In *Proceedings of the eighth ACM conference on learning@ scale* (pp. 139–150).

Darvishi, A., Khosravi, H., Sadiq, S., & Gasevic, D. (2022b). Incorporating ai and learning analytics to build trustworthy peer assessment systems.British. *Journal of Educational Technology*. https://doi.org/10.1111/bjet.13233

De Laet, T., Broos, T., Duorado, R., De Croon, R., Millecamp, M., & Verbert, K. (2018). Explainable learning analytics: Challenges and opportunities of this emerging research line. In *Adjunt proceedings of the international learning analytics & knowledge conference* (pp. 1–7).

Dikaya, L. A., Avanesian, G., Dikiy, I. S., Kirik, V. A., & Egorova, V. A. (2021). How personality traits are related to the attitudes toward forced remote learning during covid-19: Predictive analysis using generalized additive modeling. *Frontiers in Education, 6*, 108. URL: https://www.frontiersin.org/article/10.3389/feduc.2021.629213.

Dollinger, M., Liu, D., Arthars, N., & Lodge, J. (2019). Working together in learning analytics towards the co-creation of value. *J. Learn. Analy., 6*, 10–26.

Dollinger, M., & Lodge, J. M. (2018). Co-creation strategies for learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 97–101).

Dove, G., Halskov, K., Forlizzi, J., & Zimmerman, J. (2017). Ux design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 chi conference on human factors in computing systems* (pp. 278–288).

Drachsler, H., & Greller, W. (2016). Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 89–98).

Drake, B. M., & Walz, A. (2018). Evolving business intelligence and data analytics in higher education. *New Direct. Institut. Res.*, 39–52, 2018.

Echeverria, V., Martinez-Maldonado, R., & Buckingham Shum, S. (2019). Towards collaboration translucence: Giving meaning to multimodal group data. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–16).

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., & Weisz, J. D. (2021). *Expanding explainability: Towards social transparency in AI systems*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3411764.3445188. URL:.

El Maarry, K., Güntzer, U., & Balke, W. T. (2015). A majority of wrongs doesn't make it right-on crowdsourcing quality for skewed domain tasks. In *International conference on web information systems engineering* (pp. 293–308). Springer.

El-Assady, M., Jentner, W., Kehlbeck, R., Schlegel, U., Sevastjanova, R., Sperrle, F., et al. (2019). Towards xai: Structuring the processes of explanations. In *ACM workshop on human-centered machine learning*.

Engin, G., Aksoyer, B., Avdagic, M., Bozanlı, D., Hanay, U., Maden, D., et al. (2014). Rule-based expert systems for supporting university students. URL: *Procedia Computer Science, 31*, 22–31. https://doi.org/10.1016/j.procs.2014.05.241, 2nd International Conference on Information Technology and Quantitative Management, ITQM 2014 https://www.sciencedirect.com/science/article/pii/S1877050914004189.

Er, E., Villa-Torrano, C., Dimitriadis, Y., Gasevic, D., Bote-Lorenzo, M. L., Asensio-Pérez, J. I., et al. (2021). *Theory-based learning analytics to explore student engagement patterns in a peer review activity* (pp. 196–206). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3448139.3448158. URL:.

Fernandez-Nieto, G., Echeverria-Barzola, V., Buckingham Shum, S., Mangaroska, K., Kitto, K., Palominos, E., et al. (2021). *Storytelling with learner data: Guiding student reflection on multimodal team data* (pp. 1–15). IEEE Transactions on learning technologies in press.

Fernandez-Nieto, G., Martinez-Maldonado, R., Echeverria, V., Kitto, K., An, P., & Buckingham Shum, S. (2021). What can analytics for teamwork proxemics reveal about positioning dynamics in clinical simulations?. In , *5. Proceedings of the ACM on Human-Computer Interaction* (pp. 1–24).

Foltỳnek, T., Meuschke, N., & Gipp, B. (2019). Academic plagiarism detection: A systematic literature review. *ACM Computing Surveys, 52*, 1–42.

Fratamico, L., Conati, C., Kardan, S., & Roll, I. (2017). Applying a framework for student modeling in exploratory learning environments: Comparing data representation granularity to handle environment complexity. *International Journal of Artificial Intelligence in Education, 27*, 320–352.

Gardner, J., Brooks, C., & Baker, R. (2019). Evaluating the fairness of predictive student models through slicing analysis. In *Proceedings of the 9th international conference on learning analytics &; knowledge* (pp. 225–234). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3303772.3303791. URL:.

Garrison, D. R. (2016). *E-Learning in the 21st century: A community of inquiry framework for research and practice*. New York and London: Routledge.

Gašević, D., Dawson, S., & Siemens, G. (2015). Let's not forget: Learning analytics are about learning. *TechTrends, 59*, 64–71.

Gašević, D., Kovanović, V., & Joksimović, S. (2017). Piecing the learning analytics puzzle: A consolidated model of a field of research and practice. *Learning: Research and Practice, 3*, 63–78.

Gervet, T., Koedinger, K., Schneider, J., Mitchell, T., et al. (2020). When is deep learning the best approach to knowledge tracing? *JEDM| Journal Educational Data Mining, 12*, 31–54.

Ghosh, A., Heffernan, N., & Lan, A. S. (2020). Context-aware attentive knowledge tracing. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining* (pp. 2330–2339). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3394486.3403282. URL:.

Goldstein, I. P. (1979). The genetic graph: A representation for the evolution of procedural knowledge. *International Journal of Man-Machine Studies, 11*, 51–77.

Gunning, D. (2017). *Explainable artificial intelligence (xai)*. Defense Advanced Research Projects Agency (DARPA). nd Web 2.

Hassenzahl, M., & Tractinsky, N. (2006). User experience-a research agenda. *Behaviour & Information Technology, 25*, 91–97.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*, 81–112.

Heckler, N. C., Rice, M., & Hobson Bryan, C. (2013). Turnitin systems: A deterrent to plagiarism in college classrooms. *Journal of Research on Technology in Education, 45*, 229–248.

Holmes, W., Porayska-Pomsta, K., (in press). Ethics in artificial intelligence in education. Taylor & Francis.

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Buckingham Shum, S., et al. (2021). Ethics of ai in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education*, 1–23.

Hutt, S., Gardner, M., Duckworth, A. L., & D'Mello, S. K. (2019). *Evaluating fairness and generalizability in models predicting on-time graduation from college applications*. International Educational Data Mining Society.

Jørnø, R. L., & Gynther, K. (2018). What constitutes an "actionable insight" in learning analytics? *Journal of Learning Analytics, 5*, 198–221.

Kahneman, D., Sibony, O., & Sunstein, C. R. (2021). *Noise: A flaw in human judgment*. Brown: Little.

Kardan, S., & Conati, C. (2015). Providing adaptive support in an interactive simulation for learning: An experimental evaluation. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems* (pp. 3671–3680).

Kay, J. (2001). Learner control. *User Modeling and User-Adapted Interaction, 11*, 111–127.

Kay, J. (2006). Scrutable adaptation: Because we can and must. In *International conference on adaptive hypermedia and adaptive web-based systems* (pp. 11–19). Springer.

Kay, J., & Kummerfeld, B. (2019). From data to personal user models for life-long, life-wide learners. *British Journal of Educational Technology, 50*, 2871–2884.

Khosravi, H., Demartini, G., Sadiq, S., & Gasevic, D. (2021). Charting the design and analytics agenda of learnersourcing systems. In *LAK21: 11th international learning analytics and knowledge conference* (pp. 32–42). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3448139.3448143.

Khosravi, H., Kitto, K., & Joseph, W. (2019). Ripple: A crowdsourced adaptive platform for recommendation of learning activities. *Jouranl of Learning Analytics, 6*, 91–105.

Khosravi, H., Shabaninejad, S., Bakharia, A., Sadiq, S., Indulska, M., & Gašević, D. (2021). Intelligent learning analytics dashboards: Automated drill-down recommendations to support teacher data exploration. *Jouranl of Learning Analytics, 8*, 133–154.

Kim, J. (2015). *Learnersourcing: Improving learning with collective learner activity*. Massachusetts Institute of Technology. Ph.D. thesis.

Kitto, K., Buckingham Shum, S., & Gibson, A. (2018). Embracing imperfection in learning analytics. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 451–460). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3170358.3170413.

Kizilcec, R., & Lee, H. (2022). *Algorithmic fairness in education*. ArXiv, 2007.05443 https://arxiv.org/abs/2007.05443.

Kleinerman, A., Rosenfeld, A., & Kraus, S. (2018). Providing explanations for recommendations in reciprocal environments. In *Proceedings of the 12th ACM conference on recommender systems* (pp. 22–30).

Knight, S. (2020). *Augmenting assessment with learning analytics. 7*. Springer. https://doi.org/10.1007/978-3-030-41956-1_10. of The Enabling Power of Assessment. URL:.

Knight, S., Abel, S., Shibani, A., Yoong Kuan, G., Conijn, R., Gibson, A., et al. (2020). Are you being rhetorical? A description of rhetorical move annotation tools and open corpus of sample machine annotated rhetorical moves. *Jouranl of Learning Analytics, 7*, 138–154. https://doi.org/10.18608/jla.2020.73.10

Knight, S., Buckingham Shum, S., Ryan, P., Sándor, A., & Wang, X. (2018). Academic writing analytics for civil law: Participatory design through academic and student engagement. *International Journal of Artificial Intelligence in Education, 28*, 1–28. https://doi.org/10.1007/s40593-016-0121-0

Knight, S., Gibson, A., & Shibani, A. (2020). Implementing learning analytics for learning impact: Taking tools to task. *Internet and Higher Education*. https://doi.org/10.1016/j.iheduc.2020.100729

Knight, S., Shibani, A., Abel, S., Gibson, A., Ryan, P., Sutton, N., et al. (2020). Acawriter: A learning analytics tool for formative feedback on academic writing. *Journal of Writing Research, 12*, 299–344.

Knight, S., Shibani, A., & Buckingham Shum, S. (2018). Augmenting formative writing assessment with learning analytics: A design abstraction approach. In J. Kay, & R. Luckin (Eds.), *13th international conference of the learning sciences: Rethinking learning in the digital age. Making the learning sciences count* (pp. 1783–1790). International Society of the Learning Sciences.

Knijnenburg, B. P., Page, X., Wisniewski, P., Lipford, H. R., Proferes, N., & Romano, J. (2022). *Modern socio-technical perspectives on privacy*.

Koedinger, K. R., Anderson, J. R., Hadley, W. H., Mark, M. A., et al. (1997). Intelligent tutoring goes to school in the big city. *International Journal of Artificial Intelligence in Education, 8*, 30–43.

Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: Bridging the science-practice chasm to enhance robust student learning. *Cognitive Science, 36*, 757–798.

Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2019). Personalized explanations for hybrid recommender systems. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 379–390).

Kreber, C. (2005). Reflection on teaching and the scholarship of teaching: Focus on science instructors. *Higher Education, 50*, 323–359.

Krishna, K., & Murty, M. N. (1999). Genetic k-means algorithm. In , *29. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)* (pp. 433–439).

Kulesza, T., Burnett, M., Wong, W. K., & Stumpf, S. (2015). Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces* (pp. 126–137).

Kulik, J. A., & Fletcher, J. (2016). Effectiveness of intelligent tutoring systems: A meta-analytic review. *Review of Educational Research, 86*, 42–78.

Lakkaraju, H., Aguiar, E., Shan, C., Miller, D., Bhanpuri, N., Ghani, R., et al. (2015). A machine learning framework to identify students at risk of adverse academic outcomes. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1909–1918).

Lallé, S., & Conati, C. (2020). A data-driven student model to provide adaptive support during video watching across moocs. In *International conference on artificial intelligence in education* (pp. 282–295). Springer.

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the ai: Informing design practices for explainable ai user experiences. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–15).

Liao, Q. V., Pribić, M., Han, J., Miller, S., & Sow, D. (2021). *Question-driven design process for explainable ai user experiences*. arXiv preprint arXiv:2104.03483.

Lipton, Z. C. (2018). The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *ACM Queue, 16*, 31–57.

Liu, H., Gegov, A., & Cocea, M. (2015). *Rule based systems for big data: A machine learning approach, 13*. Springer.

Long, Y., & Aleven, V. (2017). Enhancing learning outcomes through self-regulated learning support with an open learner model. *User Modeling and User-Adapted Interaction, 27*, 55–88.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI conference on human factors in computing systems* (pp. 1–16).

Lyle, S. (2008). Dialogic teaching: Discussing theoretical contexts and reviewing evidence from classroom practice. *Language and Education, 22*, 222–240.

Ma, W., Adesope, O. O., Nesbit, J. C., & Liu, Q. (2014). Intelligent tutoring systems and learning outcomes: A meta-analysis. *Journal of Educational Psychology, 106*, 901.

Martinez-Maldonado, R., Echeverria, V., Fernandez Nieto, G., & Buckingham Shum, S. (2020). From data to insights: A layered storytelling approach for multimodal learning analytics. In *Proceedings of the 2020 chi conference on human factors in computing systems* (pp. 1–15).

Martinez-Maldonado, R., Power, T., Hayes, C., Abdiprano, A., Vo, T., Axisa, C., et al. (2017). Analytics meet patient manikins: Challenges in an authentic small-group healthcare simulation classroom. In *Proceedings of the seventh international learning analytics & knowledge conference* (pp. 90–94).

Maskin, E. S. (2008). Mechanism design: How to implement social goals. *The American Economic Review, 98*, 567–576.

Matcha, W., Gašević, D., Pardo, A., et al. (2020). A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective. *IEEE Transactions on Learning Technologies, 13*, 226–245.

Miao, F., Holmes, W., Huang, R., & Hui, Z. (2021). *AI and education: Guidance for policy-makers*. Technical Report. UNESCO. URL: https://unesdoc.unesco.org/ark:/48223/pf0000376709.

Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2019). To explain or not to explain: The effects of personal characteristics when explaining music recommendations. In *Proceedings of the 24th international conference on intelligent user interfaces* (pp. 397–407).

Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2020). What's in a user? Towards personalising transparency for music recommender interfaces. In *Proceedings of the 28th ACM conference on user modeling* (pp. 173–182). Adaptation and Personalization.

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

Mills, C. N., Potenza, M. T., Fremer, J. J., & Ward, W. C. (2005). *Computer-based testing: Building the foundation for future assessments*. Routledge.

Mitrovic, A. (2003). An intelligent sql tutor on the web. *International Journal of Artificial Intelligence in Education, 13*, 173–197.

Mitrovic, A. (2012). Fifteen years of constraint-based tutors: What we have achieved and where we are going. *User Modeling and User-Adapted Interaction, 22*, 39–72.

Mitrovic, A., & Martin, B. (2007). Evaluating the effect of open student models on self-assessment. *International Journal of Artificial Intelligence in Education, 17*, 121–144.

Molnar, C. (2019). Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/.

Molnar, C., König, G., Herbinger, J., Freiesleben, T., Dandl, S., Scholbeck, C. A., et al. (2020). *Pitfalls to avoid when interpreting machine learning models*. arXiv preprint arXiv:2007.04131.

Moore, M. G. (2013). The theory of transactional distance. In M. G. Moore (Ed.), *Handbook of distance education* (pp. 84–103). New York and London: Routledge.

Moore, J. D., & Paris, C. L. (1992). Exploiting user feedback to compensate for the unreliability of user models. *User Modeling and User-Adapted Interaction, 2*, 287–330.

Moosavi-Dezfooli, S. M., Fawzi, A., Fawzi, O., & Frossard, P. (2017). Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1765–1773).

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education, 31*, 199–218.

Olson, J. S., & Kellogg, W. A. (2014). *Ways of knowing in HCI*. Springer Publishing Company, Incorporated.

Páez, A. (2019). The pragmatic turn in explainable artificial intelligence (XAI). *Minds and Machines, 29*, 441–459. https://doi.org/10.1007/s11023-019-09502-w. URL:.

Palominos, E., Levett-Jones, T., Power, T., & Martinez-Maldonado, R. (2019). Healthcare students' perceptions and experiences of making errors in simulation: An integrative review. *Nurse Education Today, 77*, 32–39.

Pardo, A., Mirriahi, N., Martinez-Maldonado, R., Jovanovic, J., Dawson, S., & Gašević, D. (2016). Generating actionable predictive models of academic performance. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 474–478).

Pavlik, P. I., Jr., Cen, H., & Koedinger, K. R. (2009). *Performance factors analysis–a new alternative to knowledge tracing*. Online Submission.

Piech, C., Bassen, J., Huang, J., Ganguli, S., Sahami, M., Guibas, L. J., et al. (2015). Deep knowledge tracing. In *Advances in neural information processing systems* (pp. 505–513).

Porayska-Pomsta, K., Woolf, B., Holmes, W., & Holstein, K. (2021). The FATE of AI in education: Fairness, accountability, transparency, and ethics. *International Journal of Artificial Intelligence in Education*. URL: https://link.springer.com/journal/40593/topicalCollection/AC_dcac58fbbf2e68a27dd420b8fa69ba47.

Price, M., Handley, K., Millar, J., & O'donovan, B. (2010). Feedback: All that effort, but what is the effect? *Assessment & Evaluation in Higher Education, 35*, 277–289.

Prieto-Alvarez, C. G., Martinez-Maldonado, R., & Anderson, T. D. (2018). Co-designing learning analytics tools with learners. In J. M. Lodge, J. C. Horvath, & L. Corrin (Eds.), *Learning analytics in the classroom* (1st ed., pp. 93–110). Abingdon, Oxon ; New York, NY: Routledge. https://doi.org/10.4324/9781351113038-7. Routledge, 2019.

Prieto-Alvarez, C. G., Martinez-Maldonado, R., & Shum, S. B. (2018). Mapping learner-data journeys: Evolution of a visual co-design tool. In *Proceedings of the 30th Australian conference on computer-human interaction* (pp. 205–214).

Ribeiro, M. T., 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).

Rieber, R. W., & Carton, A. S. (1987). The collected works of ls vygotsky. *Problems of General Psychology., 1*, 325–339.

Roscoe, R. D., & McNamara, D. S. (2013). Writing pal: Feasibility of an intelligent writing strategy tutor in the high school classroom. *Journal of Educational Psychology, 105*, 1010.

Sanders, E. B. N., & Stappers, P. J. (2008). Co-creation and the new landscapes of design. *CoDesign, 4*, 5–18.

Scanlon, E., Sharples, M., Fenton-O'Creevy, M., Fleck, J., Cooban, C., Ferguson, R., et al. (2013). *Beyond prototypes: Enabling innovation in technology-enhanced learning*. milton keynes: Open university.

Sejnowski, T. J. (2020). The unreasonable effectiveness of deep learning in artificial intelligence. *Proceedings of the National Academy of Sciences, 117*, 30033–30038. https://doi.org/10.1073/pnas.1907373117. arXiv:https://www.pnas.org/content/117/48/30033.full.pdf.

Self, J., et al. (1999). The defining characteristics of intelligent tutoring systems research: Itss care, precisely. *International Journal of Artificial Intelligence in Education, 10*, 350–364.

Selwyn, N. (2019). *Should robots replace teachers?: AI and the future of education*. John Wiley & Sons.

Selwyn, N. (2021). There is a danger we get too robotic": an investigation of institutional data logics within secondary schools. *Educational Review*, 1–17.

Shapley, L. S. (2016). 17. a value for n-person games. In *Contributions to the theory of games (AM-28), II* pp. 307–318). Princeton University Press.

Sha, L., Raković, M., Whitelock-Wainwright, A., Carroll, D., Yew, V. M., Gasevic, D., et al. (2021). Assessing algorithmic fairness in automatic classifiers of educational forum posts. In *International conference on artificial intelligence in education* (pp. 381–394). Springer.

Shibani, A., Knight, S., & Buckingham Shum, S. B. (2019). Contextualizable learning analytics design: A generic model and writing analytics evaluations. event-place: Tempe, AZ, USA). In *Proceedings of the 9th international conference on learning analytics & knowledge* (pp. 210–219). ACM. https://doi.org/10.1145/3303772.3303785. URL:.

Shibani, A., Knight, S., & Buckingham Shum, S. (2020). Educator perspectives on learning analytics in classroom practice. *Internet and Higher Education*. https://doi.org/10.1016/j.iheduc.2020.100730.00000

Shibani, A., Knight, S., & Buckingham Shum, S. (2022). Questioning learning analytics? Cultivating critical engagement as student automated feedback literacy. In *LAK22: 12th international learning analytics and knowledge conference* (pp. 326–335). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3506860.3506912.

Shibani, A., Knight, S., Buckingham Shum, S., & Ryan, P. (2017). Design and implementation of a pedagogic intervention using writing analytics. In W. Chen, J. C. Yang, A. F. Mohd Ayub, S. L. Wong, & A. Mitrovic (Eds.), *25th international conference on computers in education, Asia-pacific society for computers in education* (pp. 306–315), 00000.

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International Journal of Human-Computer Studies, 146*, Article 102551.

Srinivasan, R., & Chander, A. (2020). Explanation perspectives from the cognitive sciences—a survey. In *Proceedings of the twenty-ninth international joint conference on artificial intelligence, international joint conferences on artificial intelligence organization* (pp. 4812–4818). URL: https://www.ijcai.org/proceedings/2020/670.

Steenbergen-Hu, S., & Cooper, H. (2013). A meta-analysis of the effectiveness of intelligent tutoring systems on k–12 students' mathematical learning. *Journal of Educational Psychology, 105*, 970.

Steenbergen-Hu, S., & Cooper, H. (2014). A meta-analysis of the effectiveness of intelligent tutoring systems on college students' academic learning. *Journal of Educational Psychology, 106*, 331.

Tsai, Y. S., Whitelock-Wainwright, A., & Gašević, D. (2021). More than figures on your laptop: (dis)trustfulimplementation of learning analytics. *J. Learn. Analy., 8*, 81–100.

UNESCO. (2019). *Beijing consensus on artificial intelligence and education*. URL: UNESCO. Technical Report https://unesdoc.unesco.org/ark:/48223/pf0000368303.

VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. *Educational Psychologist, 46*, 197–221.

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems* (pp. 1–15).

Warschauer, M., & Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: International Journal, 3*, 22–36. https://doi.org/10.1080/15544800701771580. URL:.

Williamson, B. (2018). The hidden architecture of higher education: Building a big data infrastructure for the 'smarter university. *International Journal of Technologies in Higher Education, 15*, 1–26.

Winne, P. H. (1995). Inherent details in self-regulated learning. *Educational Psychologist, 30*, 173–187.

Winne, P. H. (2020). Construct and consequential validity for learning analytics based on trace data. *Computers in Human Behavior, 112*, Article 106457.

Winne, P. H. (2021). Open learner models working in symbiosis with self-regulating learners: A research agenda. *International Journal of Artificial Intelligence in Education, 31*, 446–459. https://doi.org/10.1007/s40593-020-00212-4. URL:.

Xu, W. (2019). Toward human-centered AI: A perspective from human-computer interaction. *Interactions, 26*, 42–46.

Yang, F., Huang, Z., Scholtz, J., & Arendt, D. L. (2020). How do visual explanations foster end users' appropriate trust in machine learning?. In *Proceedings of the 25th international conference on intelligent user interfaces* (pp. 189–201).

Yang, Q., Scuito, A., Zimmerman, J., Forlizzi, J., & Steinfeld, A. (2018). Investigating how experienced ux designers effectively work with machine learning. In *Proceedings of the 2018 designing interactive systems conference* (pp. 585–596).