Human-Centric eXplainable AI in Education (HEXED) Workshop

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

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

Vinitra Swamy
EPFL
vinitra.swamy@epfl.ch

Tanja Käser
EPFL
tanja.kaeser@epfl.ch

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

Lea Cohausz
University of Mannheim
lea.cohausz@uni-mannheim.de

ABSTRACT

The educational data mining community has long acknowledged the "challenge of interpretability" [1] that has grown alongside the adoption of complex machine learning algorithms for educational purposes. Researchers have focused on a variety of approaches for addressing this concern, often turning to methods borrowed from the broader eXplainable AI (XAI) community. However, serious limitations with existing methods have led to calls for a re-imagining of what explainability should look like. The HEXED workshop aims to bring together a community of researchers who can work together to (1) develop a shared vision and common vocabulary for XAI in education, (2) share and disseminate work, (3) create robust methods for increasing interpretability, and (4) develop evaluation metrics for assessing explanations and model interpretability. We propose to achieve this through collaborative sense-making, research poster presentations, and lively discussions surrounding the current and future needs of the community.

Keywords

Explainable AI, interpretability, model transparency

1. INTRODUCTION

Advances in machine learning have led to complex models that are difficult or impossible to interpret. This lack of explainability often has multiple causes, such as a large number of parameters, the complexity of the architecture, or the abstract nature of the features used. As education research continues to increasingly rely on such models, this lack of interpretability can further obscure issues pertaining to fairness, accountability, and actionable insights [4]. This in turn can lead to a lack of trust among stakeholders, such as students, teachers, and administrators.

J. Pinto, L. Paquette, V. Swamy, T. Käser, Q. Liu, and L. Cohausz. Human-centric explainable ai in education (hexed) workshop. In B. Paaßen and C. D. Epp, editors, *Proceedings of the 17th International Conference on Educational Data Mining*, pages 1030–1033, Atlanta, Georgia, USA, July 2024. International Educational Data Mining Society.

https://doi.org/10.5281/zenodo.12730045

For these reasons, some researchers in education have turned to methods from the field of eXplainable AI (XAI) to demystify such "black-box" models. Tools such as LIME and SHAP have been used to provide explanations for a model's predictions, evaluating the influence of individual features on the model's output [9, 7]. In some cases, these explanations have even been used to drive student interventions [3]. Unfortunately, these commonly used post-hoc techniques have shown inherent limitations, including a lack of agreement between techniques [5], the risk of generating unjustified examples for counterfactual explanations [6], and the "blind" assumptions that must be made when treating a model as a literal black box [10]. Recent researchers have begun to bring awareness of the disagreement problem to education [11], some are seeking ways to assimilate techniques for causal modeling from the social sciences [2], while others are emphasizing the need for more intrinsically interpretable models that don't rely on post-hoc explanations [12, 8].

The growing need for interpretable AI in education, along with the increasing awareness of its challenges and implications, calls for a community of experts to work together to (1) develop a shared vision and common vocabulary, (2) disseminate work to raise awareness of the need for interpretable AI in education, (3) create robust methods for increasing interpretability, and (4) develop evaluation metrics for assessing explanations and model interpretability. The Human-Centric eXplainable AI in Education (HEXED) Workshop aims to bring together researchers and practitioners from the fields of machine learning and education to discuss the challenges and opportunities of using interpretable machine learning models in education research.

2. WORKSHOP THEMES

The workshop will cover a wide range of topics related to interpretable AI in education. The organizers will distribute a call for papers covering these themes, which include but are not limited to:

- The need for greater explainability in education.
- The case for intrinsic vs. post-hoc explainability.
- Ensuring explanation fidelity to the model.
- Designing evaluation metrics and methods for assessing explanations and/or models.

- Aligning explanations with teachers' and students' needs.
- Generating actionable explanations as a basis for classroom interventions and personalized learning.

The workshop aims to attract participants and attendees who may be interested in any of these and other related themes. We plan to have roughly thirty participants, though this number is likely to change based on interest and other logistics.

We plan to publish the papers submitted to and presented at HEXED through an established proceedings publication platform, such as CEUR-WS. Depending on the requirements of the publication platform, authors may have the option of including only an abstract or their full paper.

We will also invite submissions of encore papers—i.e. recently published research relevant to the themes of the workshop. These will not be published in the workshop proceedings, but links to the original papers will be posted on the workshop website.

3. FORMAT OF WORKSHOP

This will be a full-day hybrid workshop and will feature a mix of poster presentations, a lively panel discussion, and interactive sessions to facilitate collaboration. We hope to attract interested researchers who may not be able to attend in person, so we plan to provide ways for remote attendees to participate and interact with in-person attendees (details on this below). However, due to the nature of the poster session, presenters must be able to attend in person.

The following proposed schedule is tentative and may be adjusted based on the number of submissions and the availability of invited speakers:

9:00–9:30am	Welcome and opening remarks
9:30–10:30am	Invited keynote talk
10:30–10:45am	Break
10:45am–12:00pm	Poster session
12:00–1:00pm	Lunch
1:00–1:45pm	Panel debate
1:45–2:30pm	Working session 1: Framing problems and needs
2:30–2:45pm	Break
2:45–3:15pm	Working session 2: Breakout group brainstorming
3:15–4:30pm	Working session 3: Creating a shared vision
4:30–4:45pm	Closing thoughts

The workshop will begin with a welcome and opening remarks from members of the organizing committee. During this welcome, we will present a digital diagram that includes the current areas of research within XAI in education and their relationship to each other. Throughout the workshop, this diagram will be displayed and periodically updated by placing presented work in its appropriate location to help contextualize it. Attendees will also have the chance to extend this diagram or add notes and questions during the day. Since the diagram will be in digital format (using an interactive collaborative platform such as Miro), remote attendees will also be able to make contributions.

The opening remarks will be followed by an invited keynote talk by a respected researcher in the field who has done work in explainable AI in education. A poster session will follow, allowing participants to discuss early and work-in-progress work that aligns with the themes of the workshop. At the beginning of the poster session, each presenter will have 20 seconds to pitch their work while their poster is digitally displayed for all to see. This will give attendees a sense of the poster presenters they would like to interact with during the rest of the session. We will also have all posters available on a digital platform that allows comments, making it possible for remote attendees to participate and provide feedback.

We will then break for lunch. In-person participants will be encouraged to have lunch together so as to continue the conversation and have the chance to build a greater sense of community.

After lunch, we will hold a panel discussion in which panelists will discuss their views on important questions in the field and will provide their outlook on potential areas of future development. Following this important discussion, members of the organizing committee will lead working sessions and breakout groups with the goal of turning the day's presentations and discussions into a shared vision for the future of explainable AI in education. This culminating vision, in the form of a document published on the HEXED website alongside the collaborative diagram, will be one of the key outcomes of the workshop, and it will serve the purpose of identifying the most pressing challenges and opportunities in the field. Finally, some closing thoughts will serve to wrap up the workshop.

4. EXPECTED OUTCOMES

The main outcomes of the workshop will be the papers published in the proceedings, the collaboratively created diagram outlining the different areas of research and their interconnectedness, and a document defining the shared vision created during the panel and working session. We will also discuss the possibility of drafting a proposal for a special issue of the Journal of Educational Data Mining on the topic of XAI in education. Additionally, we may discuss the possibility of hosting a data competition for research focused on XAI.

5. WORKSHOP ORGANIZERS

The organizing committee chairs for the HEXED Workshop, along with their bios, are listed below.

Juan D. Pinto is a PhD student at the University of Illinois Urbana-Champaign. His research involves the development of learner models using machine learning methods and tackling issues of AI interpretability in education. He is currently conducting work as a member of the Human-centered Educational Data Science (HEDS) Lab and the NSF AI Institute for Inclusive Intelligent Technologies for Education (INVITE).

Luc Paquette is an associate professor in the department of curriculum & instruction at the University of Illinois Urbana-Champaign. His research focuses on the usage of machine learning, data mining and knowledge engineering approaches to analyze and build predictive models of the behavior of students as they interact with digital learning environments such as MOOCs, intelligent tutoring systems, and educational games. He is interested in studying how those behaviors are related to learning outcomes and how predictive models of those behaviors can be used to better support the students' learning experience.

Vinitra Swamy is a PhD student at EPFL. Her research with the ML4ED lab involves explainable AI for education, especially through the lens of reducing adoption barriers for neural networks. Her recent work focuses on uncovering disagreement in post-hoc explainers, using learning science experts to validate explainer accuracy and actionability, and proposing interpretable-by-design neural network architectures.

Tanja Käser is an assistant professor at the EPFL School of Computer and Communication Sciences (IC) and head of the Machine Learning for Education (ML4ED) laboratory. Her research lies at the intersection of machine learning, data mining, and education. She is particularly interested in creating accurate models of human behavior and learning, with a focus on building models that are generalizable, interpretable, and fair.

Qianhui (Sophie) Liu is a PhD student at the University of Illinois Urbana-Champaign. Her research in the HEDS lab focuses on applying data mining methods in combination with learning science theories to help improve the efficiency of teaching and learning in various educational settings. She is interested in closing the loop of machine learning to humans for actionable insights through explainable models and techniques.

Lea Cohausz is a PhD student at the University of Mannheim. Her recent work includes research on how demographic variables influence predictions in EDM and the consequences for fairness (EDM 2023) as well as identifying causal structures in educational data and their relationship to algorithmic bias (LAK 2024). She is interested in advancing our understanding of the complex relationships of factors that influence students' learning outcomes.

We will gather a group of program committee members from a diverse range of institutions. These will be researchers with experience in and publications on issues of explainable AI in education, trust, and fairness. The program committee will be tasked with reviewing submissions for inclusion in the workshop. The following committee members have accepted an invitation to participate:

- **Giora Alexandron**, Assistant Professor, Weizmann Institute of Science
- **Ryan Baker**, Professor, University of Pennsylvania
- **Anthony Botelho**, Assistant Professor, University of Florida
- **Nigel Bosch**, Assistant Professor, University of Illinois Urbana-Champaign
- **Cristina Conati**, Professor, The University of British Columbia
- **Jibril Frej**, Postdoctoral Researcher, EPFL
- **Ashish Gurung**, Postdoctoral Researcher, Carnegie Mellon University
- **Paul Hur**, PhD Student, University of Illinois Urbana-Champaign
- **HaeJin Lee**, PhD Student, University of Illinois Urbana-Champaign
- **Mirko Marras**, Assistant Professor, University of Cagliari
- **Anna Rafferty**, Associate Professor of Computer Science, Carleton College
- **Yang Shi**, PhD Student, North Carolina State University
- **Diego Zapata-Rivera**, Director of LAFI research center, Educational Testing Service

6. REFERENCES

[1] R. S. Baker. Challenges for the future of educational data mining: The baker learning analytics prizes. *Journal of Educational Data Mining*, 11(1), 2019.

[2] L. Cohausz. When probabilities are not enough: A framework for causal explanations of student success models. *Journal of Educational Data Mining*, 14(3):52–75, Dec. 2022.

[3] P. Hur, H. Lee, S. Bhat, and N. Bosch. Using machine learning explainability methods to personalize interventions for students. In *Proceedings of the 15th International Conference on Educational Data Mining*, pages 438–445. Zenodo, July 2022.

[4] H. Khosravi, S. B. Shum, G. Chen, C. Conati, Y.-S. Tsai, J. Kay, S. Knight, R. Martinez-Maldonado, S. Sadiq, and D. Gašević. Explainable artificial intelligence in education. *Computers and Education: Artificial Intelligence*, 3:100074, Jan. 2022.

[5] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. The disagreement problem in explainable machine learning: A practitioner's perspective, Feb. 2022.

[6] T. Laugel, M.-J. Lesot, C. Marsala, X. Renard, and M. Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations, July 2019.

[7] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 2017.

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

[9] M. T. Ribeiro, S. Singh, and C. Guestrin. "Why should I trust you?": Explaining the predictions of any classifier, Feb. 2016.

[10] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, May 2019.

[11] V. Swamy, S. Du, M. Marras, and T. Kaser. Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pages 345–356, Arlington TX USA, Mar. 2023. ACM.

[12] V. Swamy, J. Frej, and T. Käser. The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations, July 2023.