The 2nd Human-Centric eXplainable AI in Education (HEXED) Workshop

Vinitra Swamy
EPFL
vinitra.swamy@epfl.ch

Jakub Kuzilek
Humboldt-University of Berlin
jakub.kuzilek@hu-berlin.de

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

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

Tanja Käser
EPFL
tanja.kaeser@epfl.ch

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

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

ABSTRACT

As machine learning models and their larger language model counterparts become more prevalent in education, the challenge of interpretability has grown alongside their adoption [1]. While researchers have explored various approaches, often drawing from the broader eXplainable AI (XAI) community, current methods remain limited [5]. We highlight the need to rethink explainability in educational settings. The 2nd iteration of the Human-Centric eXplainable AI in Education (HEXED) workshop brings together researchers dedicated to advancing interpretability in AI-driven education. Our goals are to (1) establish a shared vision and common vocabulary for XAI in education, (2) facilitate the exchange of recent research and best practices, (3) brainstorm practical methods to enhance model transparency that are specific to the education domain, and (4) define evaluation metrics for assessing explanations and interpretability with teachers, students, and parents. We also aim to discuss what the rise of LLMs means for the field of eXplainable AI and seek synergies to enhance LLMs with methods in XAI towards increasing student, parent, and teacher trust. Through research presentations and structured discussions, we aim to address key challenges and shape the future of explainable AI in education.

Keywords

Explainable AI, interpretability, model transparency

1. INTRODUCTION

Since the beginning of the field, machine learning models have grown increasingly complex, often making them difficult to interpret. This lack of transparency stems from factors such as large parameter spaces, intricate architectures, and abstract feature representations. In education research, reliance on opaque models can obscure critical issues related to fairness, accountability, and actionable insights [5], ultimately eroding trust among students, educators, and administrators.

To address this, researchers have adopted eXplainable AI (XAI) techniques like LIME and SHAP to clarify model predictions by identifying influential features [10, 8]. These methods have supported student interventions [4]. However, post-hoc explanations have significant drawbacks, including inconsistencies between techniques [6, 14], misleading counterfactual examples [7], and fundamental limitations of black-box models [11].

Recent efforts in education research highlight these limitations [12] and explore alternatives. Some researchers integrate causal modeling techniques from the social sciences [2, 15], while others advocate for intrinsically interpretable models that do not rely on post-hoc explanations [13, 9]. These approaches aim to enhance trust, usability, and real-world impact in educational AI [3].

The growing demand for interpretable AI in education, alongside increasing awareness of its challenges, highlights the need for a dedicated research community. This community must work together to (1) establish a shared vision and vocabulary, (2) raise awareness of the importance of interpretability in education, (3) develop robust methods to enhance interpretability, and (4) create evaluation metrics for assessing explanations and model transparency. They must also build consensus around how eXplainable AI can be utilized for educational good in the age of large language models (LLMs) and multimodal models.

The Human-Centric eXplainable AI in Education (HEXED) Workshop brings together researchers and practitioners from machine learning and education to explore the challenges and opportunities of interpretable AI in education research.

Vinitra Swamy, Jakub Kuzilek, Juan Pinto, Luc Paquette, Tanja Kaser, Qianhui Sophie Liu, and Lea Cohausz. The 2nd Human-Centric eXplainable AI in Education (HEXED) Workshop. In Caitlin Mills, Giora Alexandron, Davide Taibi, Giosuè Lo Bosco, and Luc Paquette (eds.) Proceedings of the 18th International Conference on Educational Data Mining, Palermo, Italy, July, 2025, pp. 716–719. International Educational Data Mining Society (2025).

https://doi.org/10.5281/zenodo.15870312

This second edition builds on the success of HEXED at EDM 2024 [1] and follows in the spirit of the FATED workshop [2] on fairness and transparency in education at EDM 2022. During HEXED 2024, we had over 50 participants registered with both online and in-person participation, with 8 papers accepted to the workshop, and proceedings published in CEUR-WS in conjunction with the L3MNGET Workshop[3]. As an outcome of the previous workshop, we now also have an active HEXED slack community with over 30 participants[4].

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 faithfulness.
- 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 role of LLMs in Explainable AI.

The workshop aims to attract participants and attendees who may be interested in any of these and other related themes. 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 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

HEXED 2025 is a full-day hybrid workshop featuring 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).

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: XAI in the age of LLMs
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 the outcomes from the first HEXED workshop, summarize them and evaluate the goals set. For the second workshop, 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. We are also inviting researchers from the field of explainable AI broadly to present approaches from different fields to enable better synergies within the whole explainable AI research field. 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 one minute 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 the intersection of LLMs and XAI. 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

[1] https://hexed-workshop.github.io/
[2] https://fated2022.github.io/
[3] https://sites.google.com/view/llmworkshopedm/
[4] http://hexed-workspace.slack.com

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. The document will also set goals for broadening and supporting the community in the next year. 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.

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 deep models. Her work focuses on uncovering disagreement in post-hoc explainers (EDM 2022), using learning science experts to validate explainer accuracy and actionability (LAK 2023), and proposing interpretable-by-design neural network architectures (ICLR 2025, NeurIPS 2023). She is optimistic about the use of LLMs for communicating XAI insights to students, teachers, and parents.

Jakub Kuzilek is affiliated as a postdoctoral researcher with the Computer Science Education / Computer Science and Society research group at Humboldt-Universität zu Berlin. His research focuses on learning environments powered by machine learning systems development, explainable machine learning in education and predictive analysis of student behaviour in online environments.

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.

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.

HEXED 2025 has gathered a group of program committee members from a diverse range of institutions. These researchers have 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 already accepted an invitation to participate:

- **Giora Alexandron**, Assistant Professor, Weizmann Institute of Science
- **Ryan Baker**, Professor, University of Pennsylvania
- **Przemyslaw Biecek**, Professor, Warsaw University of Technology
- **Nigel Bosch**, Assistant Professor, University of Illinois Urbana-Champaign
- **Anthony Botelho**, Assistant Professor, University of Florida
- **Mustafa Cavus**, Assistant Professor, Eskisehir Technical University
- **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**, Assistant Professor, Utah State University
- **Diego Zapata-Rivera**, Director of LAFI research center, Educational Testing Service
- **Shradha Sehgal**, Machine Learning Engineer, Netflix

6. ACKNOWLEDGMENTS

The work in this workshop is supported by the Swiss State Secretariat for Education, Research and Innovation (SERI), the German Federal Ministry of Education and Research (BMBF), grant no. 16DHBKI045, and The United State's National Science Foundation, grant no. 1942962.

7. 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):1–17, 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] H. Drachsler. *Trusted learning analytics*. Universität Hamburg, 2018.

[4] 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.

[5] 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, Jan. 2022.

[6] 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.

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

[8] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions, 2017.

[9] 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.

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

[11] 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.

[12] 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.

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

[14] V. Swamy, B. Radmehr, N. Krco, M. Marras, and T. Käser. Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In *Proceedings of the 15th International Conference on Educational Data Mining (EDM 2022)*, 2022.

[15] V. Swamy, D. Romano, B. S. Desikan, O.-M. Camburu, and T. Käser. illuminate: An llm-xai framework leveraging social science explanation theories towards actionable student performance feedback. *AAAI*, 2025.