# Viewpoint: The Future of Human-Centric Explainable Artificial Intelligence is not Post-Hoc Explanations

VINITRA SWAMY, EPFL, Switzerland
JIBRIL FREJ, EPFL, Switzerland
TANJA KÄSER, EPFL, Switzerland

Explainable Artificial Intelligence (XAI) plays a crucial role in enabling human understanding and trust in deep learning systems. As models get larger, more ubiquitous, and pervasive in aspects of daily life, explainability is necessary to minimize adverse effects of model mistakes. Unfortunately, current approaches in human-centric XAI (e.g. predictive tasks in healthcare, education, or personalized ads) tend to rely on a single post-hoc explainer, whereas recent work has identified systematic disagreement between post-hoc explainers when applied to the same instances of underlying black-box models. In this viewpoint paper, we therefore present a call for action to address the limitations of current state-of-the-art explainers. We propose a shift from post-hoc explainability to designing interpretable neural network architectures. We identify five needs of human-centric XAI (real-time, accurate, actionable, human-interpretable, and consistent) and propose two possible routes forward for interpretable-by-design neural network workflows (adaptive routing and temporal diagnostics). We postulate that the future of human-centric XAI is neither in explaining black-boxes nor in reverting to traditional, interpretable models, but in neural networks that are intrinsically interpretable.

**JAIR Track:** Hybrid Human-Artificial Intelligence

**JAIR Associate Editor:** Myrthe Tielman

**JAIR Reference Format:**
Vinitra Swamy, Jibril Frej, and Tanja Käser. 2025. Viewpoint: The Future of Human-Centric Explainable Artificial Intelligence is not Post-Hoc Explanations. *Journal of Artificial Intelligence Research* 84, Article 2 (September 2025), 7 pages. DOI: 10.1613/jair.1.17970

## 1 Introduction

The rise of neural networks is accompanied by a severe disadvantage: the lack of transparency of their decisions. Deep models are often considered black-boxes, producing highly accurate results while providing little insight into how they arrive at those conclusions. This disadvantage is especially relevant in human-centric domains where model decisions have large, real-world impacts [46, 8]. Human-centric applications refer to scenarios where a human directly relies on model predictions to inform their decision-making process [34].

The goal of eXplainable AI (XAI) is to circumvent this failing by either producing interpretations for black-box model decisions or making the model's decision-making process transparent. As illustrated in Figure 1, model explanations range from local (single point) to global granularity (entire sample). Moreover, explainability can be integrated into the modeling pipeline at three stages:

(1) **Intrinsic explainability**: traditional ML models (e.g., decision trees) or gated model architectures (e.g. concept bottleneck models) that explicitly define the decision pathway [22].

Authors' Contact Information: Vinitra Swamy, ORCID: 0000-0002-6840-5923, vinitra.swamy@epfl.ch, EPFL, Lausanne, Switzerland; Jibril Frej, ORCID: 0009-0009-0631-0636, jibril.frej@x28.ch, EPFL, Lausanne, Switzerland; Tanja Käser, ORCID: 0000-0003-0672-0415, tanja.kaeser@epfl.ch, EPFL, Lausanne, Switzerland.


DOI: 10.1613/jair.1.17970

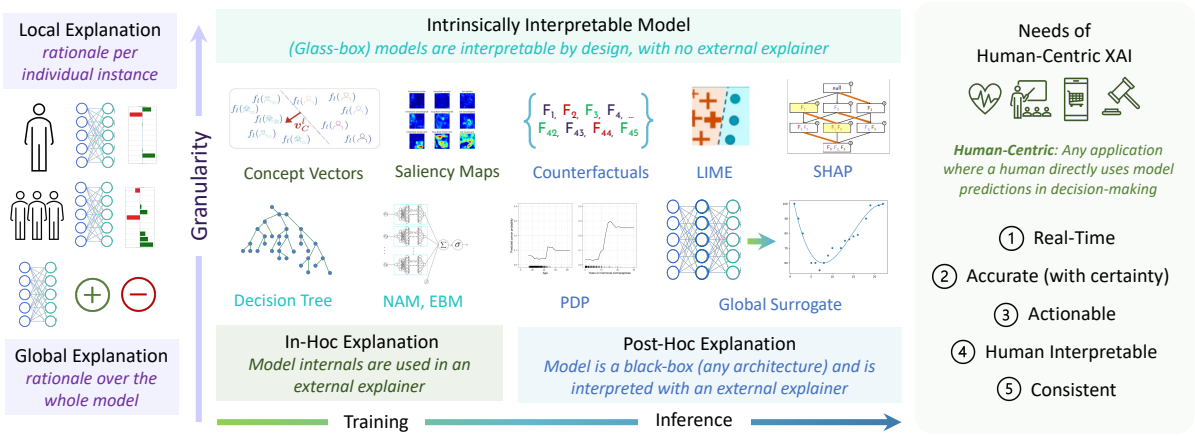


Fig. 1. Explainability can be intrinsic (by design), in-hoc (e.g., saliency methods), or post-hoc (e.g., LIME, SHAP). Furthermore, the granularity of model explanations ranges from local (single user, a group of users) to global (entire sample).

(2) **In-hoc explainability**: interpreting the model weights at inference time (e.g. saliency maps, concept activation vectors) or customizing training protocols for additional information. For example, Grad-CAM uses backpropagation to highlight important regions of an input image [39].
(3) **Post-hoc explainability**: after the decision is made, an explainer (e.g. LIME, SHAP) is fit on top of the black-box model to interpret the results.

In human-centric domains, researchers and practitioners tend to use either intrinsically interpretable traditional ML models [18, 45] or apply a single post-hoc explainer [1, 11]. Unfortunately, recent research shows that post-hoc explanations might be unfaithful to the true model [38], inconsistent [40], or method-dependent [43, 23, 6]. Furthermore, evaluating the quality of the provided explanations is a challenge, since there is often no ground truth [42, 9].

In this viewpoint paper, we therefore present a call-to-action to address the limitations of current state-of-the-art explainability methods. While previous work [38] has made a strong argument for moving away from black-box models and using inherent interpretability (i.e. traditional ML models) for impactful decisions, we suggest exploring strategies to make *deep learning* approaches intrinsically interpretable, guaranteeing transparency, robustness, and trustworthiness. We believe that human-centric domains should profit from both explainability and the recent advances in state-of-the-art machine learning methods, including large language models (LLMs).

In the following, we define five needs of human-centric XAI: real-time, accurate, actionable, human-interpretable, and consistent. We discuss the limitations of current XAI methods, their inability to meet the requirements for human-centric XAI, and two possible routes towards inherently interpretable deep learning workflows. We hope this viewpoint will serve as a call for achieving consistency and reliability in human-centric XAI systems.

## 2 Needs of Human-Centric Explainable AI

Neural networks have an enormous potential for impacting human life, from areas like personalized healthcare or educational tutoring to smart farming and finance. In light of the specific challenges in human-centric domains [34], we define five design targets that explanations should fulfill.

(1) **Real-Time**: Explanations should be provided in real-time or with minimal delay to support timely decision-making (in the scale of seconds, not tens of minutes), e.g., [47].

(2) **Accurate explanations with certainty**: Explanations need to be accurate, correctly reflecting the model's decision-making process, and accompanied by a level of confidence [31, 25]. This trait is also referred to as having high explanation *fidelity* [48].
(3) **Actionable**: Explanations should provide actionable insights, empowering model deployers to take appropriate actions or make informed interventions [17].
(4) **Human interpretable**: Explanations should be understandable to a broad audience beyond computer scientists [16, 14]. LLMs have strong potential to enhance human understanding of explanations. [49].
(5) **Consistent**: Explanations should be consistent across similar instances or contexts, ensuring reliability and predictability in the decision-making process. In a time series of interactive predictions, the explanations should not drastically differ [26].

## 3 Explainers of Today: State-of-the-Art and Limitations

Research and adoption of neural network explainability in human-centric areas has surged over the last eight years. In-hoc methods like layer relevance propagation [28] or concept-activation vectors [21] have shown success in student success prediction [4] or identifying skin conditions [29], but require specific model architectures or access to model weights. Intrinsic explainers like neural additive models [3] have shown aptitude for personalized treatments of COVID-19 patients, but require developer effort and could affect model performance. Post-hoc approaches are most commonly favored, as there is no impact on model accuracy and no additional effort required during training. Local, instance-specific post-hoc techniques such as LIME [37], SHAP [30], or counterfactuals [33], have been effectively utilized for tasks like predicting ICU mortality [20], non-invasive ventilation for ALS patients [12], credit risk [13], or loan repayment [36].

However, post-hoc approaches, while popular, are accompanied by weaknesses in real-world settings. The computational time is often in the tens of minutes; not **real-time** enough for users, students, or patients to make a decision based on the explanation alongside a prediction [32]. In most cases, there is **no measurement of confidence** in a generated post-hoc explanation [24]. The **actionability** and **human-interpretability** of the explanation are solely based on the input format [49]. As human-centric tasks often use tabular or time series data, the subsequent explanations are often not concise, actionable or interpretable easily beyond the scope of a data scientist's knowledge [19]. Recent research on explanation user design has shown that humans across healthcare, law, finance, education, and e-commerce, among others, prefer hybrid text and visual explanations [14], a format not easily provided by current post-hoc libraries. Lastly, the **consistency** of the explanations is not inherently measured. Several explainability methods could produce vastly different explanations with different random seeds or at different time steps [40].

Furthermore, post-hoc explanations are difficult to evaluate. Current metrics (e.g. saliency, faithfulness) aim to quantify the quality of an explanation in comparison to expert-generated ground truth [2]. However, accurate explanations need to be true to the model internals, not human perceptions. In this light, the most trustworthy metrics measure the prediction gap (e.g. PIU, PGU), removing features that are considered important by the explanation and seeing how the prediction changes [9]. This approach is still time-consuming and imperfect, as it fails to account for cross-feature dependencies. Recent literature [23, 6, 43] has examined the results of over 50 explainability methods with diverse datasets ranging from criminal justice to healthcare to education through a variety of metrics (rank agreement, Jenson-Shannon distance) and demonstrated strong, systematic disagreement across methods. Validating explanations through human experts can also be difficult: explanations are subjective, and most can be justified. [23], [42], and [10] have conducted user studies to examine trust in explainers, measuring data scientist and human expert preference of explanations. Results indicate that while humans generally find explanations helpful, no method is recognized as most trustworthy. As further shown by [42], most preferred explanations align with the prior beliefs of validators.

We anticipate that the state-of-the-art in AI will continue to prefer large, pretrained deep models over traditional interpretable models for the foreseeable future; the capabilities and ease-of-use of neural networks outweigh any black-box drawbacks. Our goal is therefore to identify a way to use deep learning in an interpretable workflow.

## 4 Intrinsically Interpretable Deep Learning Design

In human-centric applications, there is no margin for error; It is crucial to prioritize designs that are intrinsically interpretable as opposed to imperfect approximations of importance. We present two possible approaches towards intrinsically interpretable deep learning workflows, targeting both local and global explainability.

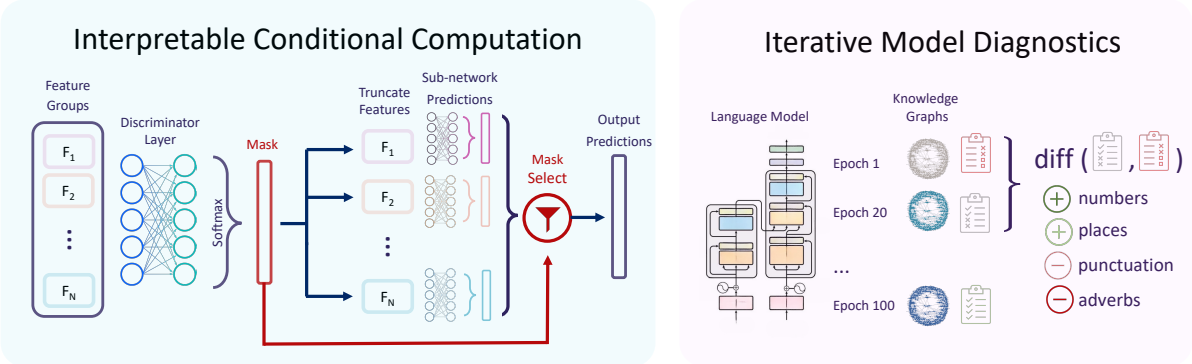


Fig. 2. Proposed architecture of adaptive routing with Interpretable Conditional Computation (left). A discriminator layer adaptively selects feature groups as important, then sends truncated feature sets to expert sub-networks. Proposed architecture of global model benchmarks with Interpretable Iterative Model Diagnostics (right). Knowledge graphs are extracted from a language model at iterative stages of training and compared over time with diagnostic benchmarks.

### 4.1 Interpretable Conditional Computation

One proposed approach to interpretable-by-design models integrates a mixture of experts with global adaptivity, ensuring explanations accurately reflect model behavior with full certainty while preserving performance. Inspired by conditional computation in neural networks [5] to enhance computational efficiency, this method distinguishes itself from feature grouping interpretability approaches [35, 7] by emphasizing human-specified feature groupings and adaptive expert subnetworks.

A simple implementation towards this idea is a *feature gating* model that dynamically learns a sparse feature mask, enforcing sparsity regularization to select minimal predictive feature sets for each data point. While this approach might appear to compromise accuracy, its adaptivity can enhance performance by reducing noise. This concept extends to a *mixture-of-experts* model, where expert sub-networks are dynamically activated based on grouped features (Figure 2). Feature groupings can be defined through humans or LLMs specifying a one-to-one assignment of features to human-understandable concepts, with each expert trained only on its subset of features and selectively activated for each data point. Initial experiments towards this approach in education (time-series), news (text), and healthcare (tabular) are presented in [41].

Interpretable Conditional Computation can optimize the interpretability-accuracy trade-off: easy-to-classify instances use less features and therefore have high interpretability while difficult-to-classify points use more features and do not trade accuracy for interpretability. The advantages of this approach are multifold, as explanations could be 1) **real-time** (a prediction is provided simultaneously with the explanation) 2) **accurate** (the

model only uses specific features or feature groups), 3) **consistent** (the same learned experts will be activated for each point), and 4) **human interpretable** (sparse explanations with human-specified groupings of features). InterpretCC's **actionability** depends on the actionability of the user-specified features, and could therefore be at a disadvantage if the specification (via human or LLM) is not meaningful.

### 4.2 Interpretable Iterative Model Diagnostics

Current deep learning performance metrics, such as accuracy and F1 score, provide an incomplete global view of model strengths and weaknesses. Addressing this, differential diagnostics of iterative model snapshots during training can offer a detailed understanding of model abilities. Experimental evaluations [44, 27] illustrate this approach by interpreting language models through comparisons of knowledge graphs extracted at various training stages (Figure 2), revealing when specific skills are learned. Tailored datasets targeting identified weaknesses can be created during training or fine-tuning, yielding a more performant model earlier and integrating XAI insights directly into the modeling pipeline. Although iterative temporal diagnostics have been discussed for usability [15], their potential for interpretability remains underexplored.

This approach suggests explanations that are **consistent** (a model snapshot will extract the same diagnostic explanations every time) and **actionable** (granular benchmarking allows developers to correct their models with custom datasets). However, it would not be **real-time**, as extracting diagnostics from model snapshots is time-consuming in the training process. **Human interpretability** depends on the choice and granularity of diagnostics. Likewise, **accuracy** depends on the breadth of the diagnostics chosen and does not have a measure of certainty. A narrow iterative benchmark might not fully capture model weaknesses, while an overly broad iterative benchmark might not be easily understandable, illustrating the interpretability-accuracy tradeoff.

## 5 Conclusion

The evolving landscape of machine learning models, characterized by the ubiquity of LLMs, transformers, and other advanced techniques, necessitates a departure from the traditional approach of explaining black-box models. Instead, there is a growing need to incorporate interpretability as an inherent feature of model design. In this viewpoint, we have discussed five needs of human-centric XAI and have shown that the current state-of-the-art is not meeting these needs. We have also presented two possible directions towards intrinsic interpretable design for neural networks and discussed their applications towards the five needs of human-centric XAI. As researchers, model developers, and practitioners, we must move away from imperfect, post-hoc XAI estimation and towards guaranteed interpretability with less friction and higher adoption in deep learning workflows.

## Acknowledgments

This project was supported by the Swiss State Secretariat for Education, Research and Innovation (SERI). We thank Tanya Nazaretsky and Martin Jaggi for helpful discussions surrounding this viewpoint paper.

## References

[1] A. Adadi and M. Berrada. 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). In *IEEE access*. Vol. 6. IEEE, 52138–52160.

[2] C. Agarwal, S. Krishna, E. Saxena, M. Pawelczyk, N. Johnson, I. Puri, M. Zitnik, and H. Lakkaraju. 2022. OpenXAI: towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems*. Vol. 35, 15784–15799.

[3] R. Agarwal, L. Melnick, N. Frosst, X. Zhang, B. Lengerich, R. Caruana, and G. E. Hinton. 2021. Neural additive models: interpretable machine learning with neural nets. 34, 4699–4711.

[4] M. Asadi, V. Swamy, J. Frej, J. Vignoud, M. Marras, and T. Käser. 2023. Ripple: concept-based interpretation for raw time series models in education. In *The 37th AAAI Conference on Artificial Intelligence (EAAI)*.

[5] Y. Bengio, N. Léonard, and A. Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. In *arXiv preprint arXiv:1308.3432*.

[6] D. Brughmans, L. Melis, and D. Martens. 2023. Disagreement amongst counterfactual explanations: how transparency can be deceptive. In arXiv: 2304.12667 [cs.AI].

[7] C. Chen, O. Li, C. Tao, A. Barnett, J. Su, and C. Rudin. 2018. This looks like that: deep learning for interpretable image recognition. In *NeurIPS*.

[8] C. Conati, K. Porayska-Pomsta, and M. Mavrikis. 2018. AI in education needs interpretable machine learning: lessons from open learner modelling. In *International Conference on Machine Learning*.

[9] J. Dai, S. Upadhyay, U. Aivodji, S. H. Bach, and H. Lakkaraju. 2022. Fairness via explanation quality: evaluating disparities in the quality of post hoc explanations. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 203–214.

[10] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das. 2018. Explanations based on the missing: towards contrastive explanations with pertinent negatives. In *Neural Information Processing Systems*.

[11] F. K. Došilović, M. Brčić, and N. Hlupić. 2018. Explainable artificial intelligence: a survey. In *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)*. IEEE, 0210–0215.

[12] A. Ferreira, S. C. Madeira, M. Gromicho, M. d. Carvalho, S. Vinga, and A. M. Carvalho. 2021. Predictive medicine using interpretable recurrent neural networks. In *International Conference on Pattern Recognition*.

[13] A. Gramegna and P. Giudici. 2021. Shap and lime: an evaluation of discriminative power in credit risk. In *Frontiers in Artificial Intelligence*.

[14] A. B. Haque, A. N. Islam, and P. Mikalef. 2023. Explainable artificial intelligence (XAI) from a user perspective: a synthesis of prior literature and problematizing avenues for future research. In *Technological Forecasting and Social Change*. Vol. 186. Elsevier, 122120.

[15] T. T. Hewett. 1986. The role of iterative evaluation in designing systems for usability. In *People and Computers II: Designing for Usability*. Cambridge University Press, Cambridge, 196–214.

[16] A. Hudon, T. Demazure, A. Karran, P.-M. Léger, and S. Sénécal. 2021. Explainable artificial intelligence (XAI): how the visualization of AI predictions affects user cognitive load and confidence. In *Information Systems and Neuroscience: NeuroIS Retreat 2021*. Springer, 237–246.

[17] S. Joshi, O. Koyejo, W. Vijitbenjaronk, B. Kim, and J. Ghosh. 2019. Towards realistic individual recourse and actionable explanations in black-box decision making systems. In *arXiv preprint arXiv:1907.09615*.

[18] M. Jovanovic, S. Radovanovic, M. Vukicevic, S. Van Poucke, and B. Delibasic. 2016. Building interpretable predictive models for pediatric hospital readmission using tree-lasso logistic regression. In *Artificial intelligence in medicine*. Vol. 72. Elsevier, 12–21.

[19] A. J. Karran, T. Demazure, A. Hudon, S. Senecal, and P.-M. Léger. 2022. Designing for confidence: the impact of visualizing artificial intelligence decisions. In *Frontiers in Neuroscience*. Vol. 16. Frontiers Media SA.

[20] G. J. Katuwal and R. Chen. 2016. Machine learning model interpretability for precision medicine. In *arXiv preprint arXiv:1610.09045*.

[21] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. Viegas, and R. A. Sayres. 2018. Interpretability beyond feature attribution: quantitative testing with concept activation vectors (TCAV). In *ICML*. http://proceedings.mlr.press/v80/kim18d/kim18d.pdf.

[22] P. W. Koh, T. Nguyen, Y. S. Tang, S. Mussmann, E. Pierson, B. Kim, and P. Liang. 2020. Concept bottleneck models. In *International conference on machine learning*. PMLR, 5338–5348.

[23] S. Krishna, T. Han, A. Gu, J. Pombra, S. Jabbari, S. Wu, and H. Lakkaraju. 2022. The disagreement problem in explainable machine learning: a practitioner's perspective. In *arXiv preprint arXiv:2202.01602*.

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

[25] B. Leichtmann, C. Humer, A. Hinterreiter, M. Streit, and M. Mara. 2023. Effects of explainable artificial intelligence on trust and human behavior in a high-risk decision task. In *Computers in Human Behavior*. Vol. 139. Elsevier, 107539.

[26] L. Li, T. Lassiter, J. Oh, and M. K. Lee. 2021. Algorithmic hiring in practice: recruiter and hr professional's perspectives on AI use in hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 166–176.

[27] L. Z. Liu, Y. Wang, J. Kasai, H. Hajishirzi, and N. A. Smith. 2021. Probing across time: what does roberta know and when? In *EMNLP Findings*.

[28] Y. Lu, D. Wang, Q. Meng, and P. Chen. 2020. Towards interpretable deep learning models for knowledge tracing. In *Artificial Intelligence in Education*.

[29] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel, and S. Ahmed. 2020. On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *2020 international joint conference on neural networks (IJCNN)*. IEEE, 1–10.

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

[31] C. Marx, Y. Park, H. Hasson, Y. Wang, S. Ermon, and L. Huan. 2023. But are you sure? an uncertainty-aware perspective on explainable AI. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 7375–7391.

[32] C. Molnar. 2020. *Interpretable machine learning*. Lulu.com.

[33] R. K. Mothilal, A. Sharma, and C. Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Conference on Fairness, Accountability, and Transparency*, 607–617.

[34] NASEM. 2021. Human-AI teaming: state-of-the-art and research needs. In *National Academy of Sciences, Engineering, and Medicine*.

[35] M. Nauta, J. Schlötterer, M. van Keulen, and C. Seifert. 2023. Pip-net: patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
[36] M. Pawelczyk, K. Broelemann, and G. Kasneci. 2020. Learning model-agnostic counterfactual explanations for tabular data. In *The Web Conference*.
[37] M. T. Ribeiro, S. Singh, and C. Guestrin. 2016. Why should I trust you?: explaining the predictions of any classifier. In *KDD*.
[38] C. Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nature Machine Intelligence* number 5. Vol. 1. Nature Publishing Group UK London, 206–215.
[39] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. 2017. Grad-cam: visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
[40] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju. 2020. Fooling lime and shap: adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 180–186.
[41] V. Swamy, J. Blackwell, J. Frej, M. Jaggi, and T. Käser. 2025. Interpretcc: intrinsic user-centric interpretability through global mixture of experts. In *International Conference on Learning Representations (ICLR)*.
[42] V. Swamy, S. Du, M. Marras, and T. Käser. 2023. Trusting the explainers: teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, 345–356.
[43] V. Swamy, B. Radhmehr, N. Krco, M. Marras, and T. Käser. 2022. Evaluating the explainers: black-box explainable machine learning for student success prediction in MOOCs. In *Educational Data Mining*.
[44] V. Swamy, A. Romanou, and M. Jaggi. 2021. Interpreting language models through knowledge graph extraction. In *NeurIPS Explainable AI Workshop*.
[45] A. Vultureanu-Albisi and C. Badica. 2021. Improving students' performance by interpretable explanations using ensemble tree-based approaches. In *IEEE International Symposium on Applied Computational Intelligence and Informatics*.
[46] M. E. Webb, A. Fluck, J. Magenheim, J. Malyn-Smith, J. Waters, M. Deschênes, and J. Zagami. 2021. Machine learning for human learners: opportunities, issues, tensions and threats. In *Education Tech Research and Development*.
[47] J. Xu, R. Rahmatizadeh, L. Bölöni, and D. Turgut. 2017. Real-time prediction of taxi demand using recurrent neural networks. In *IEEE Transactions on Intelligent Transportation Systems* number 8. Vol. 19. IEEE, 2572–2581.
[48] C.-K. Yeh, C.-Y. Hsieh, A. Suggala, D. I. Inouye, and P. K. Ravikumar. 2019. On the (in) fidelity and sensitivity of explanations. 32.
[49] A. Zytek, S. Pidò, and K. Veeramachaneni. 2024. Llms for xai: future directions for explaining explanations.

Received 27 December 2024; accepted 28 May 2025