The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations

Vinitra Swamy   VINITRA.SWAMY@EPFL.CH
Jibril Frej   JIBRIL.FREJ@EPFL.CH
Tanja Käser   TANJA.KAESER@EPFL.CH
EPFL, Switzerland

Abstract

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 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 schemes for interpretable-by-design neural network workflows (adaptive routing with InterpretCC and temporal diagnostics with I2MD). 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.

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 (Webb et al., 2021; Conati et al., 2018).

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) explicitly define the decision pathway.

2. **In-hoc explainability**: interpreting the model gradients at inference or customizing training protocols for additional information; e.g., Grad-CAM uses backpropagation to highlight important regions of an input image (Selvaraju et al., 2017).
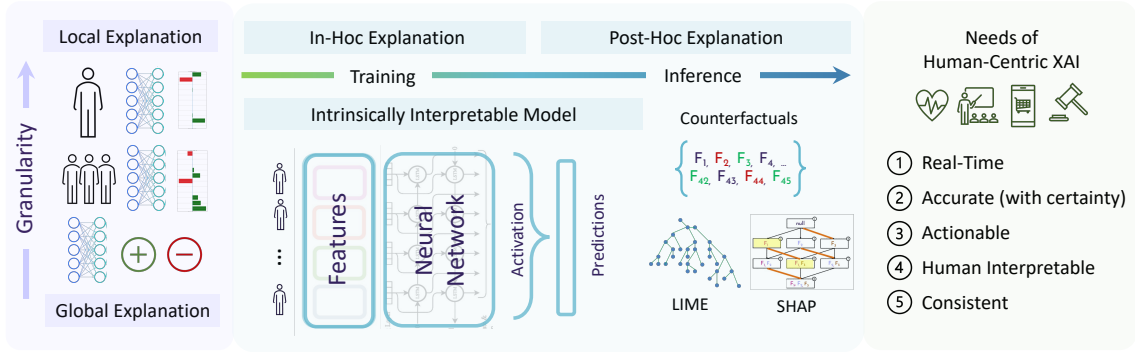
Figure 1: Explainability can be intrinsic (by design), in-hoc (e.g., gradient methods), or post-hoc (e.g., LIME, SHAP). Furthermore, the granularity of model explanations ranges from local (single user, a group of users) to global (entire sample).

3. **Post-hoc explainability**: after the decision is made, an explainer 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 (Jovanovic et al., 2016; Vultureanu-Albisi & Badica, 2021) or apply a single post-hoc explainer (Adadi & Berrada, 2018; Došilović et al., 2018). Unfortunately, recent research shows that post-hoc explanations might be unfaithful to the true model (Rudin, 2019), inconsistent (Slack et al., 2020), or method-dependent (Swamy et al., 2022; Krishna et al., 2022; Brughmans et al., 2023). Furthermore, evaluating the quality of the provided explanations is a challenge, since there is often no ground truth (Swamy et al., 2023; Dai et al., 2022).

In this paper, we therefore present a call-to-action to address the limitations of current state-of-the-art explainability methods. While previous work (Rudin, 2019) 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 ideas towards inherently interpretable deep learning workflows. We hope this paper will serve as a guideline for achieving consistency and reliability in human-centric XAI systems.

2. Requirements for 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. We define human-centric as any application where a human directly uses model predictions in decision-

making. In light of the specific challenges in human-centric domains (NASEM, 2021), we define five requirements 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., Xu et al. (2017).

2. **Accurate explanations with certainty**: Explanations need to be accurate, reflecting the neural network's decision-making process or at least accompanied by a level of confidence (Marx et al., 2023; Leichtmann et al., 2023).

3. **Actionable**: Explanations should provide actionable insights, empowering model deployers to take appropriate actions or make informed interventions (Joshi et al., 2019).

4. **Human interpretable**: Explanations should be understandable to a broad audience beyond computer scientists (Hudon et al., 2021; Haque et al., 2023). We believe that LLMs will be crucial in improving the understandability of explanations.

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 (Li et al., 2021).

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 (Lu et al., 2020) or concept-activation vectors (Kim et al., 2018) have shown success in student success prediction (Asadi et al., 2023) or identifying skin conditions (Lucieri et al., 2020), but require specific model architectures or access to model weights. Intrinsic explainers like neural additive models (Agarwal et al., 2021) 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 (Ribeiro et al., 2016), SHAP (Lundberg & Lee, 2017), or counterfactuals (Mothilal et al., 2020), have been effectively utilized for tasks like predicting ICU mortality (Katuwal & Chen, 2016), non-invasive ventilation for ALS patients (Ferreira et al., 2021), credit risk (Gramegna & Giudici, 2021), or loan repayment (Pawelczyk et al., 2020).

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. In most cases, there is **no measurement of confidence** in a generated post-hoc explanation. The **actionability** and **human-interpretability** of the explanation are solely based on the input format. 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 (Karran, Demazure, Hudon, Senecal, & Léger, 2022). 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 (Haque et al., 2023), 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 (Slack et al., 2020).

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 (Agarwal, Krishna, Saxena, Pawelczyk, Johnson, Puri, Zitnik, & Lakkaraju, 2022). 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 (Dai et al., 2022). This approach is still time-consuming and imperfect, as it fails to account for cross-feature dependencies. Recent literature (Krishna et al., 2022; Brughmans et al., 2023; Swamy et al., 2022) 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. Krishna et al. (2022), Swamy et al. (2023), and Dhurandhar et al. (2018) 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 Swamy et al. (2023), 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 ideas towards intrinsically interpretable deep learning workflows (InterpretCC and I2MD), targeting both local and global explanability.

4.1 InterpretCC: Interpretable Conditional Computation

InterpretCC aims to guarantee an explanation's accuracy to model behavior with 100% certainty, while maintaining performance by input point adaptivity. This approach is inspired by conditional computation in neural networks (Bengio et al., 2013) to speed up neural network computation. While it is similar to feature grouping interpretability approaches like Nauta et al. (2023), Chen et al. (2018), it differentiates by its focus on human-specified feature groupings and an adaptive choice of expert subnetworks. Extensive experiments of InterpretCC have been performed in Swamy et al. (2024).

The simplest implementation of InterpretCC is a feature gating model that learns a dynamic feature mask and enforces sparsity regularization on the input features. For each point, the goal is to choose a minimal predictive feature set. While it might seem that model accuracy will be compromised by this approach, the adaptivity has potential to improve
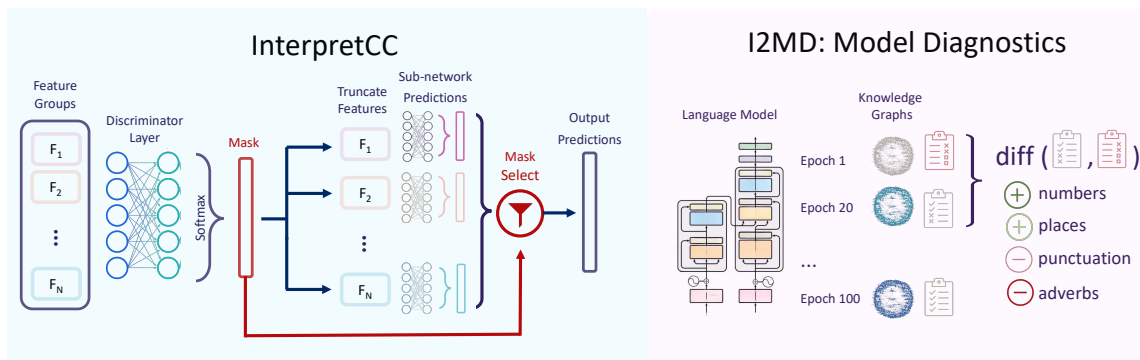
Figure 2: Proposed architecture of adaptive routing with Intepretable Conditional Computation (InterpretCC, left). A discriminator layer adaptively selects feature groupings as important, then sends truncated feature sets to expert sub-networks. Example of global model benchmarks with Interpretable Iterative Model Diagnostics (I2MD, right). Knowledge graphs are extracted from a language model at iterative stages of training and compared over time with diagnostic benchmarks.

performance by reducing noise. This idea can be expanded to an interpretable mixture-of-experts model by dynamically activating expert sub-networks (Figure 2). Instead of restricting the features individually, we can group features together meaningfully (either by human specification or automated approaches). Expert sub-networks are trained only using their subset of features and either activated or ignored for each point.

InterpretCC optimizes 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 InterpretCC are multifold, as explanations are 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 (e.g., in the breast cancer setting of Swamy et al. (2024) with grouped cell nuclei, explanations are not actionable).

4.2 I2MD: Interpretable Iterative Model Diagnostics

Current deep learning performance metrics (accuracy, F1 score) paint a starkly incomplete global picture of model strengths and weaknesses. I2MD seeks to address this gap by examining the differential diagnostics of iterative model snapshots during training to build a detailed understanding of model abilities. Experimental evaluation of this approach in Swamy et al. (2021), Liu et al. (2021) demonstrates how language models can be interpreted by comparing knowledge graphs extracted during various stages of training (Figure 2). This enables understanding of which skills the model learns at what time. During training or fine-tuning, tailored I2MD datasets can be created to target extracted model weaknesses; this results in a more performant model earlier in the training process and closes the loop, integrating XAI results back into the modeling pipeline. While the idea of iterative tem-

poral diagnostics has been extensively discussed for usability (Hewett, 1986), its role in interpretability has not been explored yet.

The I2MD approach provides 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 is not **real-time**, as extracting diagnostics from model snapshots is time-consuming in the training process. I2MD's **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 work, 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 initial ideas 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.

References

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

Agarwal, C., Krishna, S., Saxena, E., Pawelczyk, M., Johnson, N., Puri, I., Zitnik, M., & Lakkaraju, H. (2022). OpenXAI: Towards a transparent evaluation of model explanations. In *Advances in Neural Information Processing Systems*, Vol. 35, pp. 15784–15799.

Agarwal, R., Melnick, L., Frosst, N., Zhang, X., Lengerich, B., Caruana, R., & Hinton, G. E. (2021). Neural additive models: Interpretable machine learning with neural nets. *Advances in neural information processing systems*, *34*, 4699–4711.

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

Bengio, Y., Léonard, N., & Courville, A. (2013). Estimating or propagating gradients through stochastic neurons for conditional computation. In *arXiv preprint arXiv:1308.3432*.

Brughmans, D., Melis, L., & Martens, D. (2023). Disagreement amongst counterfactual explanations: How transparency can be deceptive..

Chen, C., Li, O., Tao, C., Barnett, A., Su, J., & Rudin, C. (2018). This looks like that: Deep learning for interpretable image recognition.. *NeurIPS*.

Conati, C., Porayska-Pomsta, K., & Mavrikis, M. (2018). AI in education needs interpretable machine learning: Lessons from open learner modelling. In *International Conference on Machine Learning*.

Dai, J., Upadhyay, S., Aivodji, U., Bach, S. H., & Lakkaraju, H. (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*, pp. 203–214.

Dhurandhar, A., Chen, P.-Y., Luss, R., Tu, C.-C., Ting, P., Shanmugam, K., & Das, P. (2018). Explanations based on the missing: Towards contrastive explanations with pertinent negatives. In *Neural Information Processing Systems*.

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

Ferreira, A., Madeira, S. C., Gromicho, M., Carvalho, M. d., Vinga, S., & Carvalho, A. M. (2021). Predictive medicine using interpretable recurrent neural networks. In *International Conference on Pattern Recognition*.

Gramegna, A., & Giudici, P. (2021). Shap and lime: An evaluation of discriminative power in credit risk. In *Frontiers in Artificial Intelligence*.

Haque, A. B., Islam, A. N., & Mikalef, P. (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, p. 122120. Elsevier.

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

Hudon, A., Demazure, T., Karran, A., Léger, P.-M., & Sénécal, S. (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*, pp. 237–246. Springer.

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

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

Karran, A. J., Demazure, T., Hudon, A., Senecal, S., & Léger, P.-M. (2022). Designing for confidence: The impact of visualizing artificial intelligence decisions. In *Frontiers in Neuroscience*, Vol. 16. Frontiers Media SA.

Katuwal, G. J., & Chen, R. (2016). Machine learning model interpretability for precision medicine. In *arXiv preprint arXiv:1610.09045*.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C. J., Wexler, J., Viegas, F., & Sayres, R. A. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *ICML*.

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

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

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (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*, pp. 166–176.

Liu, L. Z., Wang, Y., Kasai, J., Hajishirzi, H., & Smith, N. A. (2021). Probing across time: What does roberta know and when?. *EMNLP Findings*.

Lu, Y., Wang, D., Meng, Q., & Chen, P. (2020). Towards interpretable deep learning models for knowledge tracing. In *Artificial Intelligence in Education*.

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

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Neural Information Processing Systems*.

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

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

NASEM (2021). Human-AI teaming: State-of-the-art and research needs. In *National Academy of Sciences, Engineering, and Medicine*.

Nauta, M., Schlötterer, J., van Keulen, M., & Seifert, C. (2023). Pip-net: Patch-based intuitive prototypes for interpretable image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2744–2753.

Pawelczyk, M., Broelemann, K., & Kasneci, G. (2020). Learning model-agnostic counterfactual explanations for tabular data. In *The Web Conference*.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In *KDD*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. In *Nature Machine Intelligence*, Vol. 1, pp. 206–215. Nature Publishing Group UK London.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626.

Slack, D., Hilgard, S., Jia, E., Singh, S., & Lakkaraju, H. (2020). Fooling lime and shap: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 180–186.

Swamy, V., Blackwell, J., Frej, J., Jaggi, M., & Käser, T. (2024). Interpretcc: Conditional computation for inherently interpretable neural networks. *arXiv preprint arXiv:2402.02933*.

Swamy, V., Du, S., Marras, M., & Käser, T. (2023). Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. In *LAK23: 13th International Learning Analytics and Knowledge Conference*, pp. 345–356.

Swamy, V., Radhmehr, B., Krco, N., Marras, M., & Käser, T. (2022). Evaluating the explainers: Black-box explainable machine learning for student success prediction in MOOCs. In *Educational Data Mining*.

Swamy, V., Romanou, A., & Jaggi, M. (2021). Interpreting language models through knowledge graph extraction. In *NeurIPS Explainable AI Workshop*.

Vultureanu-Albisi, A., & Badica, C. (2021). Improving students' performance by interpretable explanations using ensemble tree-based approaches. In *IEEE International Symposium on Applied Computational Intelligence and Informatics*.

Webb, M. E., Fluck, A., Magenheim, J., Malyn-Smith, J., Waters, J., Deschênes, M., & Zagami, J. (2021). Machine learning for human learners: opportunities, issues, tensions and threats. In *Education Tech Research and Development*.

Xu, J., Rahmatizadeh, R., Bölöni, L., & Turgut, D. (2017). Real-time prediction of taxi demand using recurrent neural networks. In *IEEE Transactions on Intelligent Transportation Systems*, Vol. 19, pp. 2572–2581. IEEE.