Journal of Artificial Intelligence Research 1 (2023) 1-15 Submitted 2023; published Y

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, often defined as determining which features are most important to a model's prediction. As models get larger, more ubiquitous, and pervasive in aspects of daily life, explainability is necessary to avoid or 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 explainer. This is a particularly concerning trend when considering that recent work has identified systematic disagreement in explainability methods when applied to the same points and 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 to shift from post-hoc explainability to designing interpretable neural network architectures; moving away from approximation techniques in human-centric and high impact applications. 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 for interpretable conditional computation and diagnostic benchmarks for iterative model learning). 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 one severe disadvantage: the lack of transparency of their decisions. Deep models are often considered black-boxes because they can produce highly accurate results at the cost of 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 impact (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 model). Moreover, explainability can be integrated into the modeling pipeline at three stages: **Intrinsic explainability**: traditional ML models such as decision trees explicitly define the decision pathway. **In-hoc explainability**: interpreting the model gradients at inference or customizing training protocols for additional information; for example, Grad-CAM uses backpropagation to highlight important regions of an input image (Selvaraju et al., 2017). **Post-hoc explainability**: after the decision is made, an explainer is fit on top of the black-box model to interpret the results.

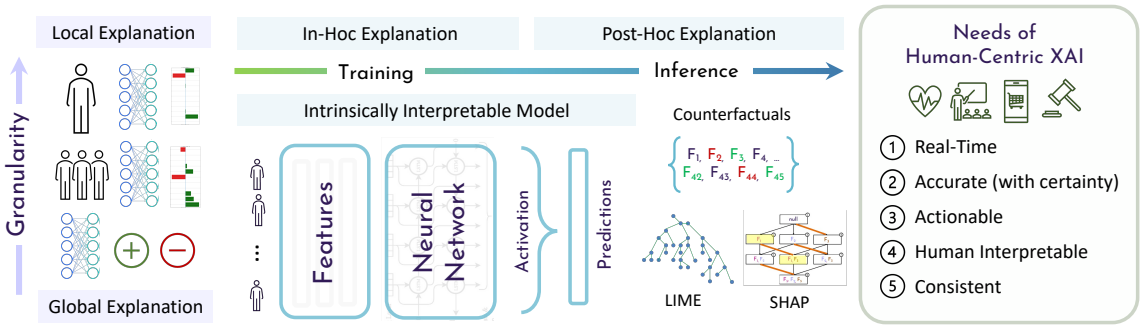

Figure 1: Explainability can be intrinsic (by design), in-hoc (e.g. gradient methods), or post-hoc (e.g. LIME, SHAP). Furthermore, granularity of model explanation is a scale from local (single user, a group of users) to global (the entire model).

In human-centric domains, researchers and practitioners tend to use either traditional ML models yielding intrinsic interpretability (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 problematic trends with post-hoc explainers. Explanations might not be faithful to the true model (Rudin, 2019) or inconsistent (Slack et al., 2020) and the explanations of different post-hoc explainers for the same model and data point have been shown to vary considerably across different methods (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. Previous work (Rudin, 2019) has made a strong argument for moving away from black-box models and for using inherent interpretability (i.e. traditional ML models) for decisions that matter. While we also propose to move away from black-box models and post-hoc explainers, we suggest exploring strategies to make *deep learning* approaches intrinsically interpretable, guaranteeing transparency, robustness, and trustworthiness of our current AI systems by design. 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).

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, and their inability to meet the requirements for human-centric XAI. We propose to focus on interpretation by design and present two ideas for inherently interpretable deep learning workflows. We hope this paper will serve as a call-to-action and 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 that has a human in the loop, where a human will directly

use the results of the model prediction as a basis of their decision-making process. In light of the specific challenges in these human-centric domains (NASEM, 2021), we have defined 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, and if not, should at least be 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, presenting information in a concise and decipherable manner. This often has to do with the interpretability of the input data and how the explanation is conveyed (visuals, text, etc.) (Hudon et al., 2021; Haque et al., 2023).

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 has surged over the last eight years, particularly in human-centric areas. 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) and SHAP (Lundberg & Lee, 2017) have been effectively utilized in a variety of models, including those predicting ICU mortality (Katuwal & Chen, 2016), non-invasive ventilation for ALS patients (Ferreira et al., 2021), and credit risk (Gramegna & Giudici, 2021). Counterfactual explanations (Mothilal et al., 2020; Klaise et al., 2021; Dhurandhar et al., 2018) have been used in numerous tasks including document classification (Martens & Provost, 2014), loan repayment (Pawelczyk et al., 2020), and image classification (Goyal et al., 2019). Less research has focused on in-hoc methods. For instance, layer relevance propagation (Lu et al., 2020) has been employed for student knowledge tracing. In addition, concept-activation vectors (Kim et al., 2018), have been proven successful in predicting student success (Asadi et al., 2023) or identifying skin conditions (Lucieri et al., 2020).

Each of the post-hoc XAI solutions presented above, among many others not mentioned, have weaknesses for deployment in a real-world setting. The computational time, especially with SHAP, LIME, or counterfactual generation, is in the tens of minutes; not real time enough for users, students, or patients to make a decision based on the explanation in the time the prediction is made (often in the scale of milliseconds). In most cases, there is no measurement of trust or confidence in a generated explanation. The actionability and human-understandability of the explanation is based on the input format. As human-centric tasks often use tabular or time series data, their 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 intrinsically measured; generating an explanation for the next step in the time series could vary greatly from the previous step. Several explainability methods could produce vastly different explanations with different random seeds (Slack et al., 2020).

Furthermore, explanations are difficult to evaluate. Current metrics (e.g. saliency, faithfulness, stability, fairness, measured in the recent OpenXAI metric suite (Agarwal et al., 2022)) have aimed to quantify the quality of an explanation, requiring a ground truth which is usually not available. However, to create a metric for a quality explanation, we need to de-bias what humans consider rational and optimize for what the model is actually doing. 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, while the best way to evaluate explanations currently, is still time-consuming and imperfect, as it fails to account for cross-feature dependencies.

Recent literature (Swamy et al., 2022; Krishna et al., 2022; Brughmans et al., 2023) has examined the results of over 50 explainability methods (e.g. LIME, SHAP, Counterfactuals, gradient-based) with diverse real-world datasets ranging from criminal justice to healthcare to education through a variety of metrics (rank agreement, Jenson-Shannon distance). These works demonstrate 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 one method is recognized as most trustworthy. As further shown by Swamy et al. (2023), most explanations align with the prior beliefs of validators, and therefore can never be a unbiased solution.

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, and it is crucial to prioritize design that is intrinsically interpretable as opposed to imperfect approximations of importance. In this section, we present two ideas towards intrinsically interpretable deep learning workflows. The first, interpretable conditional computation, is interpretable at both the local (single point) and global (entire model) level. The second approach is a global explainability approach.

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

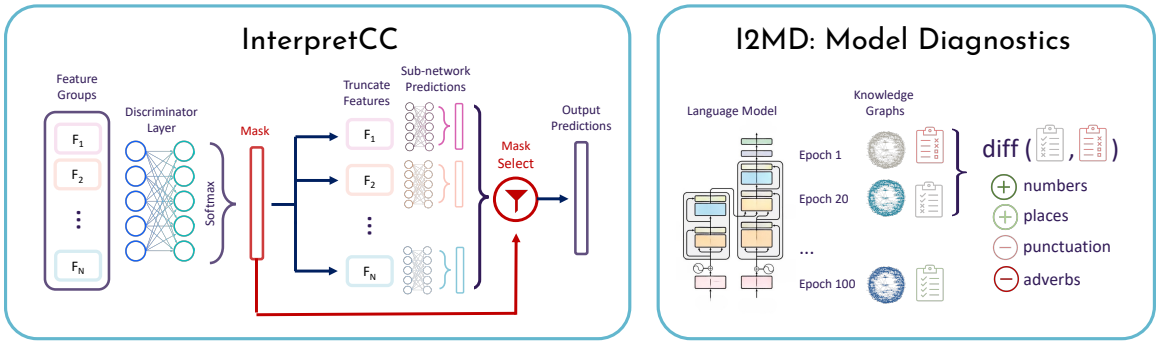

Figure 2: Proposed architecture of adaptive routing with Intepretable Conditional Computation (InterpretCC, left), detailed in Section 4.1. A discriminator layer selects adaptively selects feature groupings as important, then sends truncated feature sets to expert sub-networks. On the right, we present an example of global model benchmarks with Interpretable Iterative Model Diagnostics (I2MD), detailed in Section 4.2. Knowledge graphs are extracted from a language model at iterative stages of training and compared over time with diagnostic benchmarks.

by conditional computation in neural networks, initially introduced by Bengio et al. (2013) to speed up neural network computation.

The simplest implementation of InterpretCC is to learn a dynamic feature mask and enforce sparsity regularization on the number of input features. For each point, the goal is to choose as minimal a feature set as possible. While it might seem that some model accuracy will be compromised by this approach, this adaptivity has potential to improve performance by reducing noise.

This idea can be expanded to use expert sub-networks (Figure 2) which are dynamically conditioned on the input data to mimic a tree-like network with decision pathways. Instead of restricting the features independently, we can group features meaningfully together (either by human-selected clusterings or automated approaches). Expert sub-networks are trained only using their feature group subset and we can decide to either consider a single sub-network or several with different weightings based on a confidence threshold. We train the network to dynamically choose which route(s) to use based on the input data.

The advantages of InterpretCC are multifold: it has the potential to not comprise accuracy, it has guaranteed interpretability (as the model only uses specific features or feature groups), and it has no additional cost to the traditional development workflow. Additionally, InterpretCC optimizes the interpretability-accuracy trade-off with a customizable sparsity criterion; easy-to-classify points have high interpretability and more difficult-to-classify points do not trade accuracy for interpretability.

4.2 I2MD: Interpretable Iterative Model Diagnostics

Current deep learning performance metrics (accuracy, F1 score) paint a starkly incomplete picture of model strengths and weaknesses. I2MD seeks to address this gap by examining the differential diagnostics of iterative snapshots of model training to build a detailed understanding of model abilities. For example, language models can be interpreted by extracting

knowledge graphs during various stages of training (Swamy et al., 2021) and comparing the iterative knowledge graphs to understand which skills the model learns at what time (illustrated in Figure 2). Snapshots of pre-trained models (i.e. BERT, GPT3, T5) can provide knowledge graphs that can be directly compared to one another, allowing practitioners to make an informed choice of which model strengths best fits their downstream use case.

Another use case is in iterative self-learning, showcased recently by the DeepMind Alpha models (Jumper et al., 2021). Each model generates data to train on, improves itself with that data, and is placed in direct competition with its previous iteration. Using granular diagnostic benchmarks between each iteration of model improvement to track model ability development, the training process can be transparent to the developer. During training or fine-tuning, tailored 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.

5. Conclusion

The evolving landscape of machine learning models, characterized by the ubiquity of large language models (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 the models themselves. In this work, we have discussed five needs of human-centric XAI. We have shown how the current state-of-the-art is not meeting those needs and how post-hoc explainers will always be imperfect measurements of model behavior. We have also presented two initial ideas towards intrinsic interpretable design for neural networks. 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.

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..

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*.

Goyal, Y., Wu, Z., Ernst, J., Batra, D., Parikh, D., & Lee, S. (2019). Counterfactual visual explanations. In *PMLR*.

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.

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.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. In *Nature*, Vol. 596, pp. 583–589. Nature Publishing Group UK London.

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*.

Klaise, J., Van Looveren, A., Vacanti, G., & Coca, A. (2021). Alibi explain: algorithms for explaining machine learning models. In *JMLR*.

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.

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*.

Martens, D., & Provost, F. (2014). Explaining data-driven document classifications. In *Management Information Systems Quarterly*.

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*.

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., 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.