Trustworthy AI requires algorithmic interpretability:
Some takeaways from recent uses of eXplainable AI (XAI) in education

Juan D. Pinto*, **Qianhui (Sophie) Liu** | University of Illinois Urbana-Champaign
*Corresponding author: jdpinto2@illinois.edu

WHAT IS AI INTERPRETABILITY?

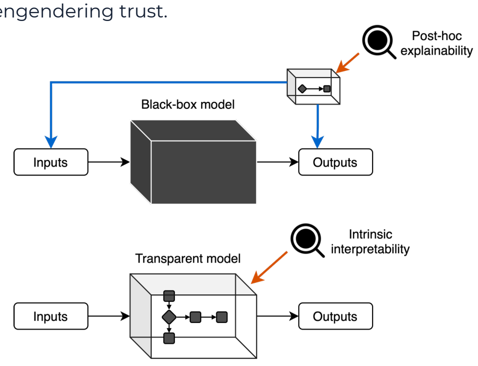
Interpretability/explainability refers to our ability to understand why a machine learning model performs a specific prediction (*local explainability*) and which inputs play the biggest role overall (*global explainability*). The field of research that studies this ability to interpret models is known as **eXplainable AI (XAI)**.

WHY IS IT IMPORTANT?

AI transparency (or the lack thereof) has important ramifications for issues of **fairness**, **accountability**, and **trust**:

- It serves as a prerequisite for addressing issues of algorithmic bias, which disproportionally and negatively affect students from historically disadvantaged populations (Kizilcec & Lee, 2022).
- Without interpretability, there is no clear path to hold parties accountable for the decisions of AI models.
- Instilling trust in AI models is a two-way process that involves providing clear and accurate explanations of decisions to students and teachers, while simultaneously ensuring that such automated decision-making aligns with human perception.

Despite this, we very often use models that are entirely opaque to interpretation—aptly called "black-box" models. This "**challenge of interpretability**" is considered one of the main challenges currently faced by AIEd researchers (Baker, 2019). While interpretability may not be a requirement for creating powerful AI, it serves as a foundational step towards engendering trust.



ISSUES WE IDENTIFIED FROM REVIEWING THE LITERATURE

Limitations of post-hoc explainability

Within XAI, there are two general approaches to model transparency: intrinsic and post-hoc explainability.

- **Intrinsic interpretability** — when a human can directly inspect and understand the *inner workings* of a model. This label typically defines the nature of the model itself.
- **Post-hoc explainability** — methods that are applied *after* a model has been created and used, which try to extract insights from indirect observations of model predictions. The most common post-hoc methods are model agnostic and rely exclusively on inputs and outputs (Carvalho et al., 2019).

Unfortunately, post-hoc explainability methods have been shown to have *serious inherent limitations*. These include:

- **Lack of agreement** between techniques (Krishna et al., 2022).
- The **risk of generating unjustified examples** for counterfactual explanations (Laugel et al., 2019).
- The **"blind" assumptions** that must be made when treating a model as a literal black box (Rudin, 2019).

Some researchers have concluded that high-stakes domains (such as education) should rely on intrinsically interpretable models instead (Rudin, 2019; Swamy, Frej, et al., 2023).

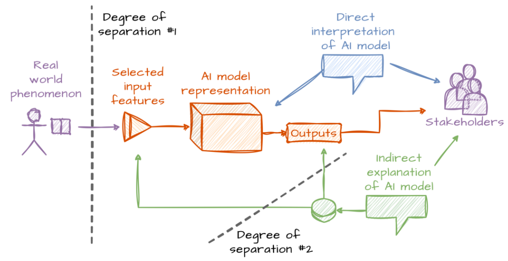
Lack of reporting

While AIEd researchers have devised robust methods for measuring different aspects of model accuracy, **evaluation methods for interpretability** have yet to be agreed upon. Standard metrics to evaluate either specific explanations or the intrinsic interpretability of a model could go a long way in motivating changes to address the *challenge of interpretability*.

Ultimately, each potential use for interpretability—such as informing learning theory, debugging models, or auditing decisions for accountability—**may require different types of evaluation approaches**.

Lack of awareness

There appears to be a general lack of awareness among AIEd researchers regarding these issues. Studies often use the explanations created by a single post-hoc approach without questioning their fidelity to the model's inner workings. Moreover, there is a tendency to use XAI as a mechanism for designing interventions under the assumption that the model captures causal relationships in the real world with fidelity. This can produce conclusions with two degrees of separation from reality.



Lack of consistent vocabulary

Rather than being a minor semantic technicality, the inconsistency of terms and definitions may be one reason for the lack of awareness among researchers. In many cases, the qualifiers **post-hoc/intrinsic** are dropped. Some examples:

- Exclusive use of **interpretability** (eg. Doshi-Velez & Kim, 2017) or **explainability** (eg. Putnam & Conati, 2019).
- Use of both terms interchangeably (eg. Yeung, 2019).
- Use of both terms for distinction (eg. Mathrani et al., 2021).
- Same as above but with altered meanings (eg. Cohausz, 2022, who uses *interpretability* to refer to an understanding of real-world causal phenomena external to the mode).

One final note: researchers often use the term "**ante hoc**" to stand in contrast with post hoc, but this is nonsensical in the context of AI model interpretability—"ante hoc" means "before this", but *before what*? A more appropriate term may be "**intra hoc**," ("within this") but the closest we've seen is "**in hoc**" ("in this"; Swamy, Frej, et al. (2023)).

WHERE DO WE GO FROM HERE?

Ultimately, if the goal is to create **explanations that can be faithful to the model and simultaneously instill trust**, then researchers may need to forgo post-hoc approaches in favor of the more challenging task of designing AI models for education that are intrinsically interpretable. To this end, the field should seek to:

1. Establish a consistent vocabulary and vision.
2. Instill greater awareness of the need for interpretability and the complexities of XAI.
3. Create robust approaches for achieving intrinsic interpretability
4. Develop evaluation methods that can lead to more regular reporting of model transparency.

REFERENCES

Alamri, R., & Alharbi, B. (2021). Explainable student performance prediction models: A systematic review. *IEEE Access*, *9*, 33132–33143. https://doi.org/10.1109/ACCESS.2021.3061368

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

Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine learning interpretability: A survey on methods and metrics. *Electronics*, *8*(8). https://doi.org/10.3390/electronics8080832

Cohausz, L. (2022). When probabilities are not enough: A framework for causal explanations of student success models. *Journal of Educational Data Mining*, *14*(3), 52–75. https://doi.org/10.5281/zenodo.7304800

Doshi-Velez, F., & Kim, B. (2017). *Towards a rigorous science of interpretable machine learning* (No. arXiv:1702.08608). arXiv. https://arxiv.org/abs/1702.08608

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The Ethics of Artificial Intelligence in Education*. Routledge.

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* (No. arXiv:2202.01602). arXiv. https://doi.org/10.48550/arXiv.2202.01602

Laugel, T., Lesot, M.-J., Marsala, C., Renard, X., & Detyniecki, M. (2019). *The dangers of post-hoc interpretability: Unjustified counterfactual explanations* (No. arXiv:1907.09294). arXiv. https://arxiv.org/abs/1907.09294

Mathrani, A., Susnjak, T., Ramaswami, G., & Barczak, A. (2021). Perspectives on the challenges of generalizability, transparency and ethics in predictive learning analytics. *Computers and Education Open*, *2*, 100060. https://doi.org/10.1016/j.caeo.2021.100060

Putnam, V., & Conati, C. (2019). Exploring the need for explainable artificial intelligence (XAI) in intelligent tutoring systems (ITS). *IUI Workshops'19*.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: Teacher validation of explainable artificial intelligence for course design. *LAK23: 13th International Learning Analytics and Knowledge Conference*, 345–356. https://doi.org/10.1145/3576050.3576147

Swamy, V., Frej, J., & Käser, T. (2023). *The future of human-centric eXplainable Artificial Intelligence (XAI) is not post-hoc explanations* (No. arXiv:2307.00364). arXiv. https://arxiv.org/abs/2307.00364

Yeung, C.-K. (2019). Deep-IRT: Make deep learning based knowledge tracing explainable using item response theory. *Proceedings of The 12th International Conference on Educational Data Mining (EDM 2019)*.

ILLINOIS