Towards a future with robust explainable AI in education

Trustworthy AI Lab for Education Summer Online Symposium 2024

by Juan D. Pinto

on June 12, 2024

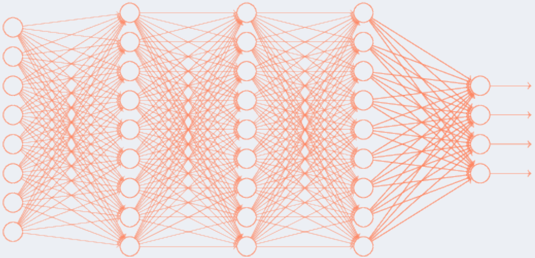
» Overview

- The challenge of interpretability
  - *Intrinsic* vs *post-hoc* explainability
  - Our takeaways for XAI in education (from *TALE Summit 2023*)
- Our **preliminary work** on interpretable neural networks for learner modeling
  - A proposed **unified framework** for evaluating explanations
- **Upcoming workshop** on XAI in education @ *EDM 2024*

The challenge of interpretability

» Model interpretability

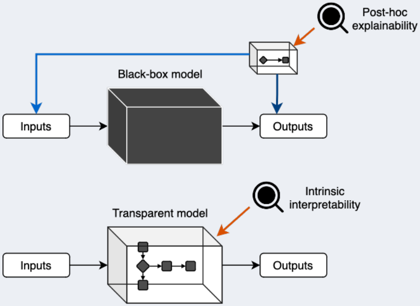
- A key component of algorithmic TRANSPARENCY
  - Important for issues of **fairness**, **accountability**, **trustworthiness**, **regulatory compliance**, and **improvability**
- "Black-box" models are now more common than ever



Deep neural network—a complex "black-box" model.

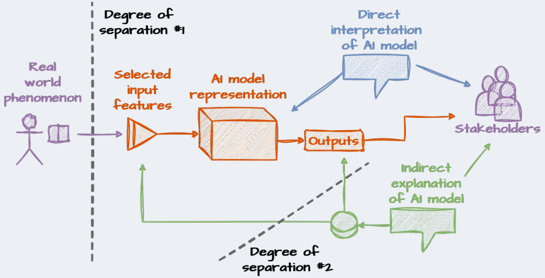
» Intrinsic interpretability and post-hoc explainability

- **Intrinsic interpretability** *(model property)*: how easy is it to understand the model's inner workings by observing its parameters?
- **Post-hoc explainability** *(analysis methods)*: without peering directly into the inner workings (parameters) of the model, what can we learn about how it works?



» Some problems with post-hoc explainability

- *Different post-hoc methods* often **lead to different conclusions** (Krishna et al., 2022; Swamy et al., 2023)[a][b]
- These methods make "blind" assumptions precisely because they treat the model as a *literal black box* (Rudin, 2019)[c]



The problem of additional separation from ground truth.

[a] Krishna, S., Han, T., Gu, A., Pombra, J., Jabbari, S., Wu, S., & Lakkaraju, H. (2022).The disagreement problem inexplainable machine learning: A practitioner's perspective.

[b] Swamy, V., Du, S., Marras, M., & Kaser, T. (2023). Trusting the explainers: Teacher validation of explainable artificialintelligence for course design.

[c] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable modelsinstead.

» Takeaways for the field

1. The establishment of a unified vision for explainable AI (XAI) in education
2. Greater awareness of the **complexities of XAI**, including the problematic limitations of post-hoc methods
3. Research into **possible approaches** for increasing model interpretability
4. The development of explainability evaluation methods

Addressing takeaway 3:
One approach for increasing the interpretability of a black-box learner model

» Data

- 6,057 students using the *Cognitive Tutor Algebra* ITS.
- **Gaming the system** (GTS) behavior: "*Attempting to succeed in an educational environment by exploiting properties of the system rather than by learning the material and trying to use that knowledge to answer correctly*." (Baker & de Carvalho, 2008)[1]
- Features from Paquette et al. (2014)[2], who used **cognitive task analysis** (expert think-aloud and training) interpret student actions.

Identifier	Description
[did not think before help request]	Pause smaller or equal to 5 seconds before a help request
[thought before help request]	Pause greater or equal to 6 seconds before a help request
[read help messages]	Pause greater or equal to 9 seconds per help message after a help request
[scanning help messages]	Pause between 4 and 8 seconds per help message after a help request
[searching for bottom-out hint]	Pause smaller or equal to 3 seconds per help message after a help request
[thought before attempt]	Pause greater or equal to 6 seconds before step attempt
[planned ahead]	Last action was a correct step attempt with a pause greater or equal to 11 seconds
[guess]	Pause smaller or equal to 5 seconds before step attempt
[unsuccessful but sincere attempt]	Pause greater than or equal to 6 seconds before a bug
[guessing with values from problem]	Pause smaller than or equal to 5 seconds before a bug
[read error message]	Pause greater than or equal to 9 seconds after a bug
[did not read error message]	Pause smaller than or equal to 8 seconds after a bug
[thought about error]	Pause greater than or equal to 6 seconds after an incorrect step attempt
[same answer/diff. context]	Answer was the same as the previous action, but in a different context
[similar answer]	Answer was similar to the previous action (Levenshtein distance of 1 or 2)
[switched context before right]	Context of the current action is not the same as the context for the previous (incorrect) action (referred to as "soft underbelly" in Baker, Mitrovic, & Mathews 2010)
[same context]	Context of the current action is the same as the previous action
[repeated step]	Answer and context are the same as the previous action
~~[diff. answer AND/OR diff. context]~~	~~Answer or context is not the same as the previous action~~

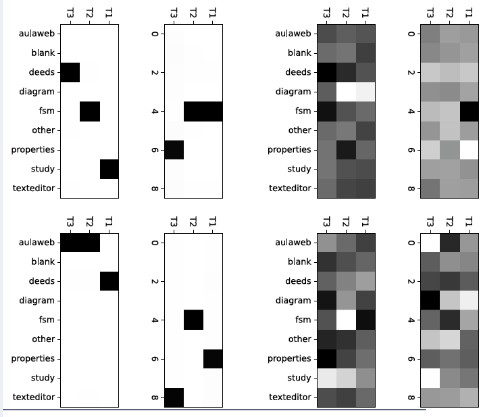
Pattern
incorrect → [guess] & [same answer/diff. context] & **incorrect**
incorrect → [similar answer] [same context] & **incorrect** → [similar answer] & [same context] & **attempt**
incorrect → [similar answer] & **incorrect** → [same answer/diff. context] & **attempt**
[guess] & **incorrect** → [guess] & [diff. answer AND/OR diff. context] & **incorrect** → [guess] & [diff. answer AND/OR diff. context & **attempt**
incorrect → [similar answer] & **incorrect** → [guess] & **attempt**
help & [searching for bottom-out hint] → **incorrect** → [similar answer] & **incorrect**
incorrect → [same answer/diff. context] & **incorrect** → [switched context before correct] & **attempt/help**
bug → [same answer/diff. context] & **correct** → **bug**
incorrect → [similar answer] & **incorrect** → [switched context before correct] & **incorrect**
incorrect → [switched context before correct] & **incorrect** → [similar answer] & **incorrect**
incorrect → [similar answer] & **incorrect** → [did not think before help] & **help** → **incorrect** (with first or second answer similar to the last one)
help → **incorrect** → **incorrect** → **incorrect** (with at least one similar answer between steps)
incorrect → **incorrect** → **incorrect** → [did not think before help request] & **help** (at least one similar answer between steps)

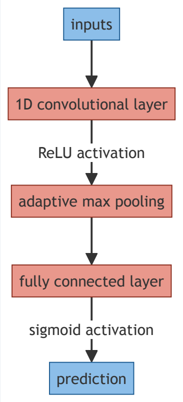
[1] Baker, R. S., & de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays.

[2] Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014). Towards understanding expert coding of studentdisengagement in online learning.

» Constraints-based interpretability: binary filters (Pinto et al., 2023)[4]

- Constrained a CNN by using a *penalty term* in its loss function designed to encourage it to learn *interpretable filters*, as first described by Jiang & Bosch (2021).[3]





[3] Jiang, L., & Bosch, N. (2021). Predictive sequential pattern mining via interpretable convolutional neural networks.

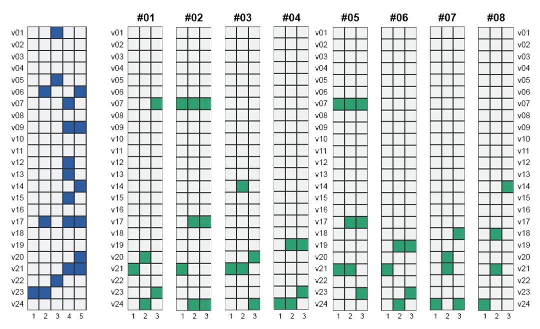
[4] Pinto, J. D., Paquette, L., & Bosch, N. (2023). Interpretable neural networks vs. expert-defined models for learner behavior detection.

Addressing takeaway 4: Evaluating interpretability

» Evaluating interpretability

- *Human-grounded evaluation*: involves human users performing simplified tasks (Doshi-Velez & Kim, 2017)[a].
- **Participants:** experts + non-experts
- **Two tasks:**
  - Forward simulation - predict the model's output given specific inputs
  - Counterfactual simulation - identify how a specific input needs to be changed to alter the model's output
- **Analysis:**
  - Accuracy rate
  - IRR between participants
  - Group comparisons

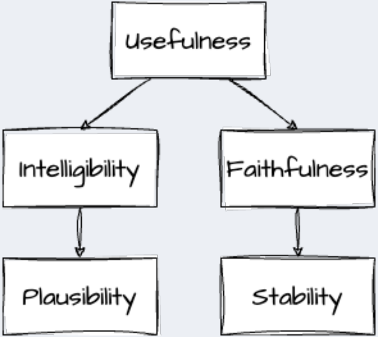
Example forward simulation task: If *GTS*, which pattern?



[a] Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

» Towards a unified framework for evaluating explanations (Pinto & Paquette, 2024)[5]

* Explanations serve as mediators between models and stakeholders.
  * Applies to both intrinsically interpretable models and black-box models with post-hoc explanations.



Evaluation criteria framework. Edges depict the direction of dependence (A -> B = A depends on B).

[5] Pinto, J. D., & Paquette, L. (2024). Towards a unified framework for evaluating explanations.

Addressing takeaway 1:
Attend our (hybrid) workshop!

» HEXED @ EDM 2024

* HEXED (**Human-Centric eXplainable AI in Education**) Workshop
* **July 14, 2024** @ EDM 2024 (Atlanta, Georgia)
    * Hybrid event
* Organizers from *University of Illinois Urbana-Champaign*, *EPFL*, and *University of Mannheim*.
* *https://hexed-workshop.github.io*

Thank you!

jdpinto.com

References

» References

Baker, R. S., & de Carvalho, A. M. J. A. (2008). Labeling student behavior faster and more precisely with text replays. *Proceedings of the 1st International Conference on Educational Data Mining (EDM)*, 38–47.

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

Jiang, L., & Bosch, N. (2021). Predictive sequential pattern mining via interpretable convolutional neural networks. *Proceedings of the 14th International Conference on Educational Data Mining (EDM 2021)*, 761–766.

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

Paquette, L., de Carvalho, A. M. J. A., & Baker, R. S. (2014). Towards understanding expert coding of student disengagement in online learning. *Proceedings of the 36th Annual Cognitive Science Conference*, 1126–1131.

Pinto, J. D., & Paquette, L. (2024). *Towards a unified framework for evaluating explanations* (No. arXiv:2405.14016). arXiv. https://arxiv.org/abs/2405.14016

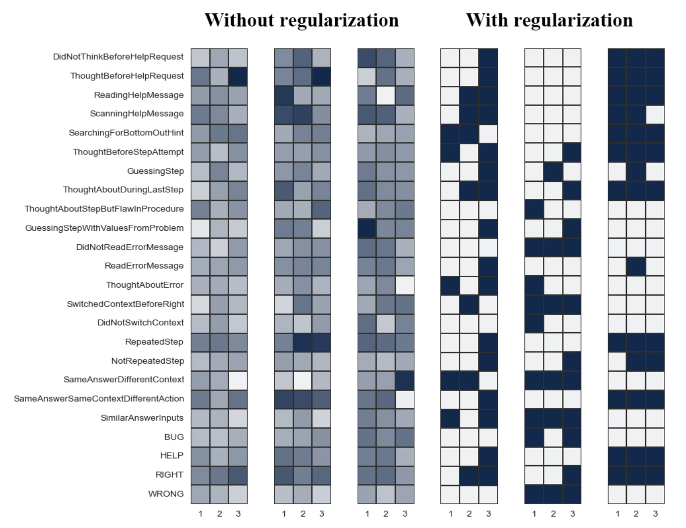
Pinto, J. D., Paquette, L., & Bosch, N. (2023). Interpretable neural networks vs. Expert-defined models for learner behavior detection. *Companion Proceedings of the 13th International Conference on Learning Analytics & Knowledge Conference (LAK23)*, 105–107.

» References (cont.)

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

» Preliminary results



Weights unregularized vs. regularized filters.