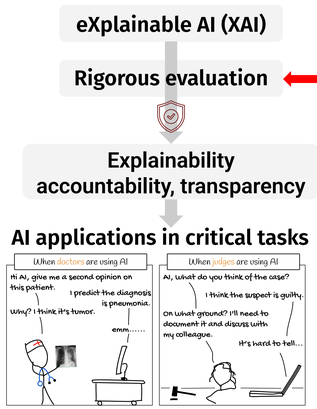
Stop Using Plausibility as the Criterion for Explainable AI

Weina Jin[1] **Xiaoxiao Li**[2] **Ghassan Hamarneh**[1] 1. Medical Image Analysis Lab, School of Computing Science, Simon Fraser University 2. Department of Electrical and Computer Engineering, University of British Columbia. Vector Institute

Motivation: Responsible evaluation of XAI



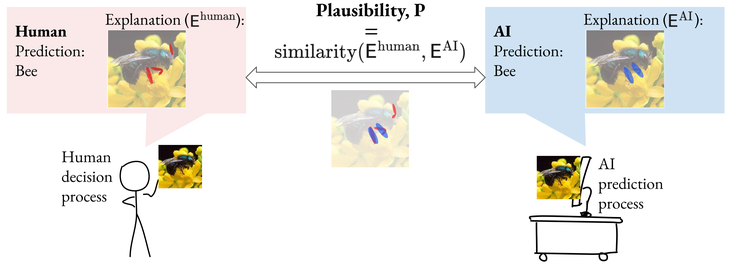
provide rationales for AI predictions in human-understandable ways [1]

Critical examination of a main evaluation criterion: plausibility

- **Explainability** is an AI ethics principle, acting as a check and balance to hold technical sectors **accountable** [2].
- **Explainability or other benefits of XAI** do not automatically hold by design or by theoretical guarantee, and they must **undergo rigorous evaluation**.
- The **evaluation methods** are vital to safeguard the scientific and responsible development of XAI, and they **should withstand critical examinations**.

Plausibility: the similarity between AI and human explanations

- **Original definition:** Plausibility assesses the reasonableness of an AI explanation by comparing it with human prior knowledge [3].
  - Can be assessed with human subjects.
- **Computational definition**: Plausibility is the content similarity or agreement between AI and human explanations expressed in the same explanation form (illustrated below).
  - Can be assessed computationally.
  - Metrics: feature correlation, feature localization, pointing game rate, etc.



Plausibility is a common criterion in XAI evaluation

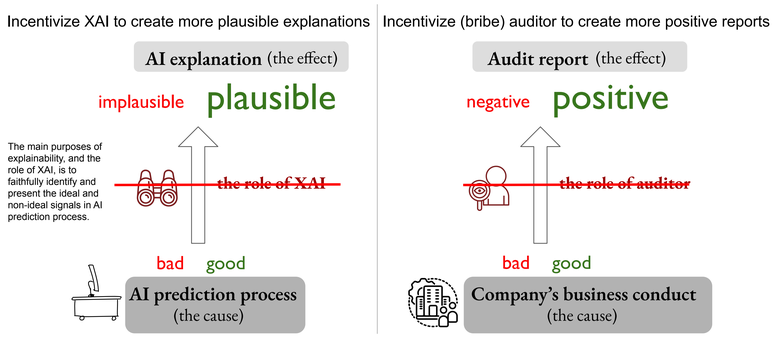
- 1/3 of XAI papers published in prominent AI conferences used the plausibility criterion [4].
- The plausibility criterion is also common in XAI literature reviews and programming libraries.

Open question

What are we missing in our community to prevent prevalent mistakes like this? [5]

Plausibility is an invalid measure of explainability
i.e., human explanations are not ground truth of AI explanations

Using the plausibility criterion is unscientific: it lacks construct validity



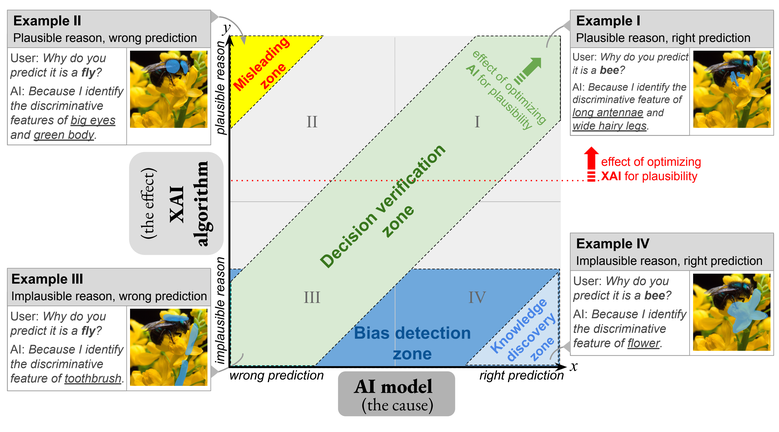
The plausibility criterion does not specify how the effect (explanation) will relate to the underlying cause, encouraging a spurious positive semblance.

Blind spots in XAI
1. XAI is to hold AI accountable, not to persuade users to adopt AI predictions.

2. Assumptions for human explanations (e.g., the quality of explanation is validly associated with the quality of decision) do not automatically hold for AI explanations.

Using the plausibility criterion increases misleading explanations

Using the plausibility criterion is unethical: it manipulates users and destroys trust



Optimizing AI models to **learn** plausible

Suitable

should not be confused with

Optimizing/evaluating XAI algorithms to **present** plausible

Unsuitable

Using the plausibility criterion undermines human autonomy and cannot achieve complementary human-AI performance

Theorem 1 (Case of Impossible Complementarity for XAI). *Let h, m, and t be the accuracies of the human, AI, and human-AI team, respectively; and $f(P_i)$ be a function of the explanation plausibility P_i, denoting the probability from a Bernoulli distribution of human acceptance of the AI suggestion for the instance $x_i \in \mathcal{D}$, then:*

If plausibility is independent of the AI prediction correctness, then the human-AI team can never achieve complementary accuracy, i.e.: $t \leq \max(h, m)$.

Theorem 2 (Conditions for XAI Complementarity). *Let h, m, and t be the accuracies of the human, AI, and human-AI team, respectively; $f(P_i)$ be the probability from a Bernoulli distribution of human acceptance of an AI suggestion for an instance $x_i \in \mathcal{D}$, where $f(.)$ is a monotonically non-decreasing function of the explanation plausibility P_i; and P_i^r and P_i^w be the plausibility values of an AI explanation when an instance x_i is predicted correctly or incorrectly, then:*

Complementary human-AI accuracy can be achieved, i.e., $t > \max(h, m)$, when

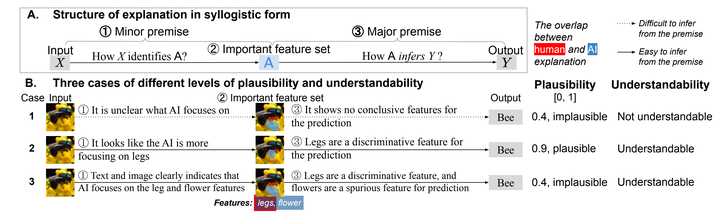
$$\begin{cases} h \geq m & \text{and} \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w]; & \text{or,} \\ m > h & \text{and} \quad \mathbb{E}[f^r] > \frac{h(1-m)}{m(1-h)}\mathbb{E}[f^w] + \frac{m-h}{m(1-h)} \end{cases}$$

where $\mathbb{E}[f^r]$ and $\mathbb{E}[f^w]$ are the expectations of $f(P_i^r)$ and $f(P_i^w)$ over the dataset $\mathcal{D}$, indicating among the correctly or incorrectly predicted instances of AI, how many are accepted by human.

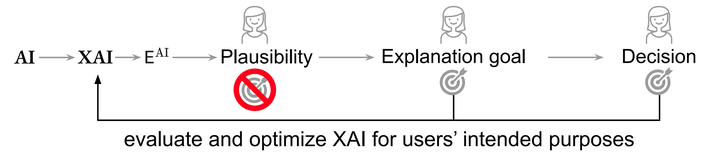
Our finding suggests an alternative objective of AI development

The AI agenda does not necessarily need to focus on outperforming and replacing humans. By making AI a good team member that can reliably indicate its weaknesses and limitations, we can achieve better performance even with an inferior AI [5].

Using the plausibility criterion can marginalize other alternatives of enhancing understandability



How to use plausibility properly? Use it as a means, not as an end (criterion)



References

[1] F. Doshi-Velez, B. Kim. Towards A Rigorous Science of Interpretable Machine Learning. 2017. arXiv: 1702.08608
[2] L. Floridi, et al. AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. 2018. Minds and Machines
[3] A. Jacovi and Y. Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? 2020. ACL
[4] M. Nauta, et al. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. 2023. ACM Computing Surveys
[5] W. Jin, E. L. Zheng, G. Hamarneh. Making Power Explicable in AI: Analyzing, Understanding, and Redirecting Power to Operationalize Ethics in AI Technical Practice. 2025. arXiv: 2510.10588