Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence (IJCAI-25)
Doctoral Consortium Track

Rating AI Models for Robustness Through a Causal Lens

Kausik Lakkaraju
AI Institute, University of South Carolina
kausik@email.sc.edu

Abstract

AI models are increasingly accessible through chatbots and other applications, but their black-box nature and sensitivity to small changes in the input make them hard to interpret and trust. Existing correlation-based robustness metrics fail to explain model errors or isolate causal effects. To address this, I propose ARC (AI Rating through Causality), a causally-grounded framework for rating AI models based on their robustness. ARC evaluates robustness by quantifying statistical and confounding biases, as well as the impact of perturbations on model performance across diverse tasks. ARC produces interpretable raw scores and ratings, helping developers and users make informed decisions about model robustness. Two future directions include: (1) deriving raw scores for composite models from their component scores, and (2) combining ratings with traditional explainable AI approaches to provide a more holistic view of model behavior.

1 Problem

1.1 Motivation

AI models have advanced rapidly and are now widely accessible through chatbots and other applications. However, their growing complexity has introduced opacity, making them harder to interpret, especially in critical domains like healthcare and education. These models often learn correlations rather than causal relationships, limiting their trustworthiness. Robustness refers to a model's ability to maintain performance under varying conditions, such as input perturbations or the presence of protected attributes. My work measures robustness by quantifying statistical bias, confounding bias, and the impact of perturbations on model performance. Existing methods often overlook confounders and fail to isolate the true effects of perturbations. ARC addresses these gaps and generalizes easily to other AI models. In contrast to existing correlation-based robustness or fairness metrics, ARC provides a causally grounded approach that supports direct comparison across systems.

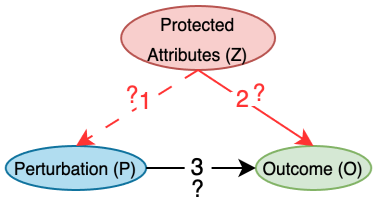

Figure 1: Proposed generalized causal model. The validity of link '1' depends on the conditional distribution $(P|Z)$, while the validity of links '2' and '3' can be tested using ARC.

1.2 Research Questions

The causal model $\mathcal{M}$ used by ARC, is shown in Figure 1. Arrowheads indicate the causal direction from cause to effect. If *Protected Attribute* (Z) is a common cause of both *Perturbation* (P) and *Outcome* (O), it introduces a spurious correlation between P and O, a confounding effect, making Z the confounder. This is undesirable as it obscures the true causal impact of P on O, leading to potentially misleading conclusions about model behavior. ARC aims to address the following research questions:

RQ1: Can we identify and quantify the effect of protected attributes on the model's outcome in the absence of confounders, indicating statistical bias?

RQ2: Can we identify and quantify the effect of protected attributes on the relationship between perturbations and model outcomes, indicating confounding bias?

RQ3: Can we quantify the causal impact of perturbations on the model's outcome?

RQ4: Can the robustness of a composite system be inferred from the raw scores of its individual components (e.g., a sentiment model combined with a translator)?

RQ5: Can existing eXplainable AI (XAI) approaches complement ARC and work in tandem with it?

1.3 Contributions

Through our previous works on causality-based rating, we addressed RQ1–RQ3 for AI models across different tasks, including sentiment analysis [Lakkaraju *et al.*, 2024b], time-series forecasting [Lakkaraju *et al.*, 2024a; Lakkaraju *et al.*, 2025], binary classification, and group recommendation. The latter two are integrated into the ARC tool along with the former two, available here: http://casy.cse.sc.edu/causal_rating.

2 Approach

ARC takes input data and applies domain-relevant perturbations (for e.g., introducing null values in numerical time-series data which can happen due to incomplete transmission) to obtain predictions, which are then evaluated using causality-based metrics we introduced in [Lakkaraju *et al.*, 2023; Lakkaraju *et al.*, 2024b] to compute raw scores. The raw scores are used to rate the AI models under each perturbation, and ratings are assigned by grouping scores into discrete levels.

2.1 Evaluation of Approach

In [Lakkaraju *et al.*, 2025], we conducted a user study presenting time-series model prediction errors alongside our computed ratings, which helped users more easily compare model robustness. In future work, another user study will be necessary to evaluate the effectiveness of combining our rating framework with XAI approaches. To assess the reliability of our causal analyses, we plan to use Rosenbaum sensitivity analysis [Rosenbaum, 1987] to estimate the magnitude of hidden bias needed to invalidate our findings.

2.2 Impact of Research

My research could have the following impact:
a) Enabling robustness-based comparisons of AI models, including both bias and sensitivity to perturbations, across diverse tasks, which could encourage benchmarking platforms like Hugging Face to adopt more meaningful robustness metrics.
b) Providing a practical framework for analyzing composite models, which are common in real-world applications.
c) Translating complex causal analyses into interpretable scores and ratings, allowing non-expert users to make informed decisions and supporting model selection and auditing.

3 Future Work

We explored RQ4 stated in Section 1.2 in [Lakkaraju *et al.*, 2023] by combining machine translation systems with sentiment analysis models. However, we did not establish a formal relationship such as an inequality or functional bound between the raw scores of individual components and the composite system. Deriving such a relationship, even approximate (e.g., bounding the composite score by the min or max of component scores), could be critical for understanding and debugging systems like chatbots built from multiple AI models. RQ5 remains an open direction. ARC shares similarities with global XAI approaches but differs in a fundamental way. Unlike traditional global XAI methods [Ribeiro *et al.*, 2016; Sundararajan *et al.*, 2017; Lundberg and Lee, 2017], ARC takes a causally-grounded, hypothesis-driven approach, generating raw scores based on user-defined hypotheses. For example, a user might want to test whether a specific gender *causally* affects the sentiment predicted by a model for sentences such as "Amanda is feeling depressed." Global XAI methods help identify features that consistently influence model predictions. When many features are present, such insights can guide ARC to prioritize which variables to test for causal impact, rather than evaluating all features equally. The ultimate goal is to position ARC as a decision-support framework that aligns with the six desirable criteria for effective decision aids outlined in [Miller, 2023], which are grounded in the ten cardinal decision issues proposed by [Yates and Potworowski, 2012].

References

[Lakkaraju *et al.*, 2023] K. Lakkaraju, A. Gupta, B. Srivastava, M. Valtorta, and D. Wu. The effect of human v/s synthetic test data and round-tripping on assessment of sentiment analysis systems for bias. In *2023 5th IEEE International Conference on Trust, Privacy and Security in Intelligent Systems and Applications (TPS-ISA)*, pages 380–389, Los Alamitos, CA, USA, nov 2023. IEEE Computer Society.

[Lakkaraju *et al.*, 2024a] Kausik Lakkaraju, Rachneet Kaur, Zhen Zeng, Parisa Zehtabi, Sunandita Patra, Biplav Srivastava, and Marco Valtorta. Rating multi-modal time-series forecasting models (mm-tsfm) for robustness through a causal lens. *arXiv preprint arXiv:2406.12908*, 2024.

[Lakkaraju *et al.*, 2024b] Kausik Lakkaraju, Biplav Srivastava, and Marco Valtorta. Rating sentiment analysis systems for bias through a causal lens. *IEEE Transactions on Technology and Society*, pages 1–1, 2024.

[Lakkaraju *et al.*, 2025] Kausik Lakkaraju, Rachneet Kaur, Parisa Zehtabi, Sunandita Patra, Siva Likitha Valluru, Zhen Zeng, Biplav Srivastava, and Marco Valtorta. On creating a causally grounded usable rating method for assessing the robustness of foundation models supporting time series. *arXiv preprint arXiv:2502.12226*, 2025.

[Lundberg and Lee, 2017] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.

[Miller, 2023] Tim Miller. Explainable ai is dead, long live explainable ai! hypothesis-driven decision support, 2023.

[Ribeiro *et al.*, 2016] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[Rosenbaum, 1987] Paul R Rosenbaum. Sensitivity analysis for certain permutation inferences in matched observational studies. *Biometrika*, 74(1):13–26, 1987.

[Sundararajan *et al.*, 2017] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[Yates and Potworowski, 2012] J Frank Yates and Georges A Potworowski. Evidence-based decision management. 2012.