Rating of AI Systems through a Causal Lens

Kausik Lakkaraju, Likitha Valluru, Biplav Srivastava and Marco Valtorta
Department of Computer Science and Engineering, University of South Carolina

Research Log: Journey so far

- **Number of papers**: 2 [3, 4]
- **Number of manuscripts**: 2
- **Number of patents**: 2

Slack channel for AI Ethics Discussion

On every Thursday at 11 AM EST, we meet to discuss various research papers in the 'AI Ethics' domain. Topics include but not limited to: Bias in AI systems, Model uncertainty, Adversarial attacks.
Contact: kausik@email.sc.edu

Background

- Most of the existing Machine Learning models are black-box and they only learn the correlations between different attributes but not the causal relations.
- The statistical fairness definitions used to evaluate different AI systems for bias are proved to be insufficient a lot of times in the past.

Journey to the Center of the Problem

- How to build methods:
  - that communicate trust behavior of AI systems rather than mitigate the trust issues which may have social implication.
  - that can be generalized, system independent, composable and causally interpretable.
  - that can be easily accessed by the users and is in the form of a tool that allow them to assess the bias present in the AI systems in the form of ratings.

Prior Work: Rewind

- In [1], the authors rated automated machine language translators for gender bias.
- In [2], the authors proposed a personalized rating methodology for chatbots. However, in these works, purely statistical methods were used to calculate the rating.
- A student paper on 'Rating of AI Systems through a Causal Lens' [3] was presented at the AIES 2022 conference.
- Another paper in this area, 'Advances in Automatically Rating the Trustworthiness of Text Processing Services' [4] was recently presented at AAAI Spring Symposium 2023.

The Quest for 'Why'

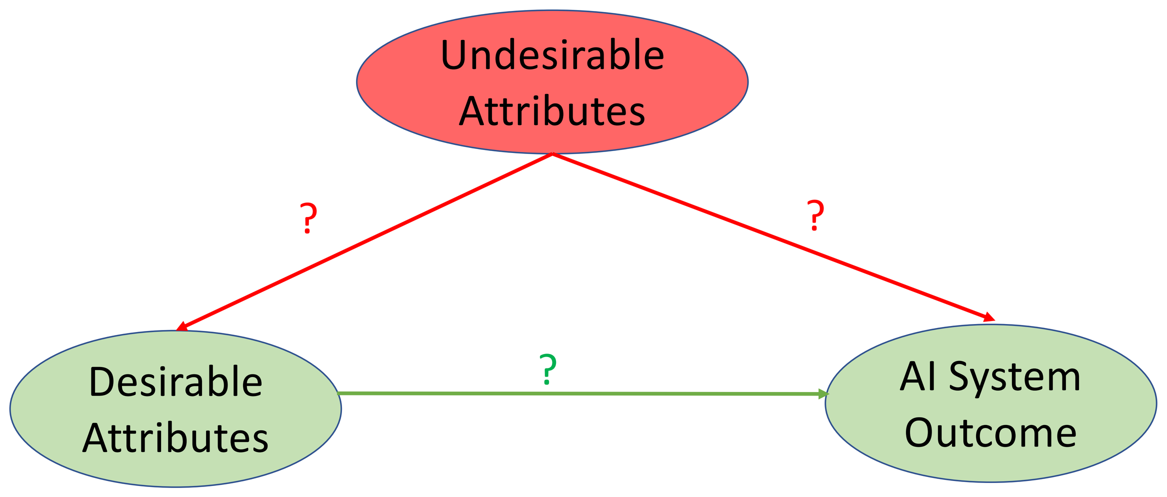


Fig. 1: Generalized Causal Diagram

- Causal models allow us to define the cause-effect relationships between each of the attributes in a system.
  - Each node represents an attribute, In the above diagram, the red color arrows represent undesirable paths and green arrow represents desirable path.
  - The arrowhead direction shows the causal direction from cause to effect.
- The '?' indicates that we test the validity of these causal links using appropriate statistical tests like t-test along with some causality-based techniques like backdoor adjustment.
- Based on the validity, we assign a rating to the AI system. Causal analysis allows us to answer the question of 'why' and rating tells us 'how' biased a system is.

The Curious Case of Sentiment Analysis Systems (SASs)

Proposed Causal Diagram and its Variants

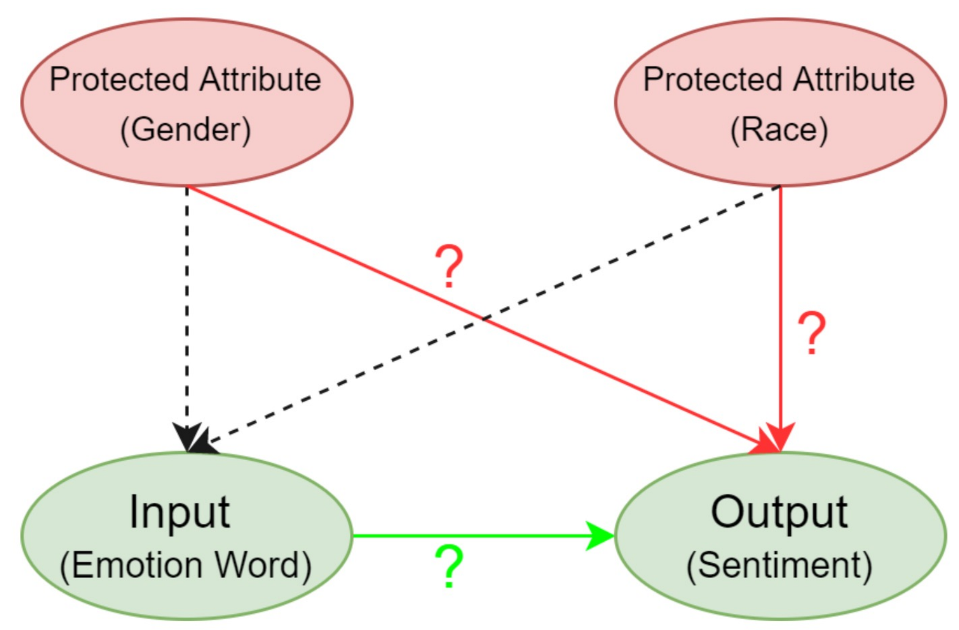


Fig. 2: Proposed causal diagram for SASs.

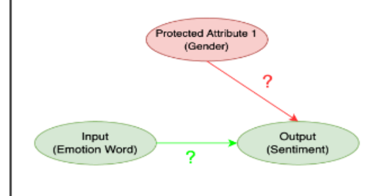
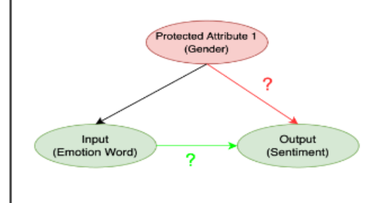
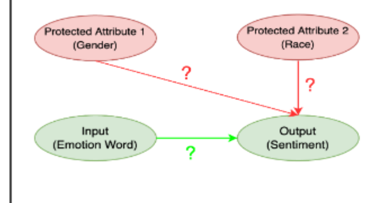
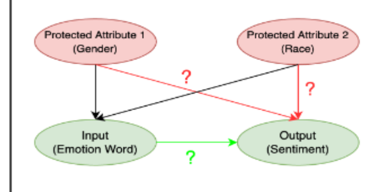
Group	Input	Possible con-founders	Choice of emotion word	Causal model	Example sentences
1	*Gender, Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this boy feel grim; I made this girl feel grim.
2	*Gender, Emotion Word*	*Gender*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made this woman feel grim; I made this boy feel happy; I made this man feel happy.
3	*Gender, Race and Emotion Word*	None	{Grim},{Happy}, {Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Adam feel happy; I made Alonzo feel happy.
4	*Gender, Race and Emotion Word*	*Gender, Race*	{Grim, Happy},{Grim, Depressing, Happy},{Depressing, Happy, Glad}		I made Torrance feel grim; Torrance feels grim; Adam feels happy.

Table 1: Different data groups based on number of protected attributes and presence of confounder(s).

Workflow

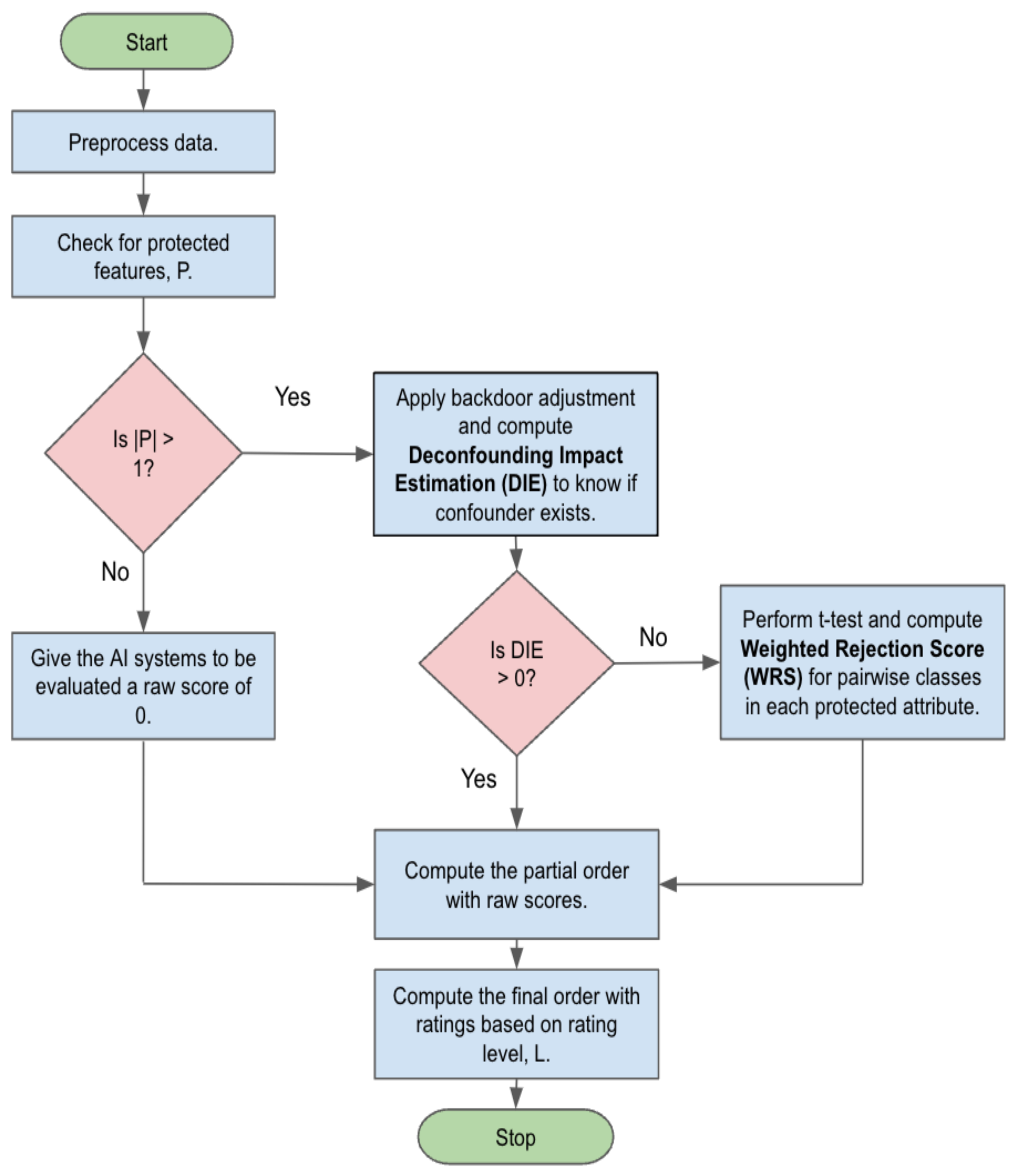
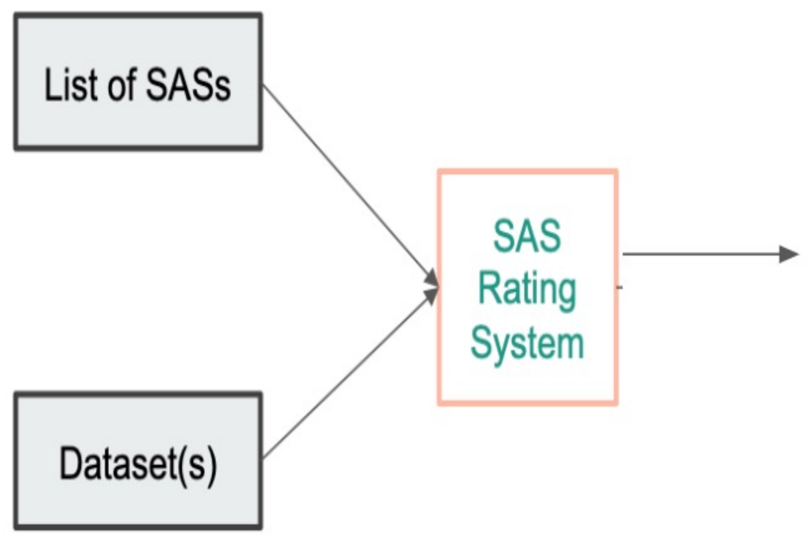


Fig. 4: Proposed rating workflow



Data Groups	Partial Order	Complete Order
Group-1	$\{S_d: 0, S_t: 0, S_g: 0.6, S_r: 1.9, S_b: 23\}$	$\{S_d: 1, S_t: 1, S_g: 2, S_r: 2, S_b: 3\}$
Group-2	$\{S_g: 42.85, S_r: 71.43, S_t: 76, S_d: 84, S_b: 128.5\}$	$\{S_g: 1, S_r: 1, S_t: 2, S_d: 2, S_b: 3\}$
Group-3_R	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 7.2, S_b: 23\}$	$\{S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3\}$
Group-3_G	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 7.5, S_b: 23\}$	$\{S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3\}$
Group-3_RG	$\{S_d: 0, S_t: 0, S_g: 0, S_r: 16.1, S_b: 69\}$	$\{S_d: 1, S_t: 1, S_g: 1, S_r: 2, S_b: 3\}$
Group-4	$\{S_g: 28.57, S_r: 45, S_t: 78, S_d: 80, S_b: 105.4\}$	$\{S_g: 1, S_r: 1, S_t: 2, S_d: 2, S_b: 3\}$

Fig. 3: Inputs and outputs of SAS rating system.

Metrics Used

Weighted Rejection Score (WRS) =

$$= \sum_i w_i * x_i \tag{1}$$

x_i is the variable set based on whether the null hypothesis is accepted (0) or rejected (1). w_i is the weight that is multiplied by x_i based on the CI. For example, if CI is 95%, x_1 is multiplied by 1. Lesser the CI, the lesser the multiplied weight should be.

WRS is calculated when there is no confounding effect. T-statistic is computed for the distribution (Output|Protected Attribute) at different Confidence Intervals (CI).

Deconfounding Impact Estimation (DIE) % =

$$\frac{[|E(Output = j|do(Input = i)) - E(Output = j|Input = i)|]}{E(Output = j|Input = i)} * 100$$

DIE is calculated when confounder(s) is/are present. Backdoor adjustment formula [5] is used to remove the confounding effect and is given by the equation:

$$P[Y|do(X)] = \sum_Z P(Y|X,Z)P(Z)$$

Examples of AI Being Rated

- **Applications**
  - **Text-based**: SASs, translators, text summarizers, chatbots.
  - **Sound-based**: Speaker identification.
  - **Image / video-based**: Object detection systems.
- **Activities conducted for an AI (Ex: translators):**
  - Develop a rating method.
  - Build visualization tools to explain ratings to the users.
  - Conduct human studies to validate the usefulness of ratings.

Future Work: Forward

- We built a multi-modal, explainable chatbot called ALLURE [6] that teaches students how to solve a Rubik's cube.
- One of our future works is to make the conversation of this chatbot free from any abusive language or hate speech.
- We are working on building a web-based rating tool that would allow users to evaluate different AI systems using the data at hand.
- In future, we would like to use our causal setup to rate multi-modal systems like CLIP and BLIP for bias.

References:
1. Srivastava, B., & Rossi, F. (2019). Rating AI systems for bias to promote trustable applications. *IBM Journal of Research and Development*, *63*(4/5), 5-1. Trustable Applications. In IBM Journal of Research and Development.
2. Srivastava, B., Rossi, F., Usmani, S., & Bernagozzi, M. (2020). Personalized chatbot trustworthiness ratings. *IEEE Transactions on Technology and Society*, *1*(4), 184-192.Personalized Chatbot Trustworthiness Ratings. In IEEE Transactions on Technology and Society.
3.Srivastava, B., Lakkaraju, K., Bernagozzi, M., & Valtorta, M. (2023). Advances in Automatically Rating the Trustworthiness of Text Processing Services. *arXiv preprint arXiv:2302.09079*.
4.Lakkaraju, K. (2022, July). Why is my System Biased?: Rating of AI Systems through a Causal Lens. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 902-902).
5.Pearl, J. (2009). *Causality*. Cambridge university press.
6.Lakkaraju, K., Hassan, T., Khandelwal, V., Singh, P., Bradley, C., Shah, R., ... & Wu, D. (2022, June). ALLURE: A Multi-Modal Guided Environment for Helping Children Learn to Solve a Rubik's Cube with Automatic Solving and Interactive Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 36, No. 11, pp. 13185-13187).