ARC*: A Tool to Rate AI Models for Robustness Through a Causal Lens

Kausik Lakkaraju[1] , **Siva Likitha Valluru**[1] , **Biplav Srivastava**[1,2] and **Marco Valtorta**[2]
[1]AI Institute, University of South Carolina, Columbia, USA
[2]Department of Computer Science and Engineering, University of South Carolina, Columbia, USA
kausik@email.sc.edu, svalluru@email.sc.edu, biplav.s@sc.edu, mgv@cse.sc.edu

Abstract

As Artificial Intelligence (AI) systems become more powerful, concerns about trust issues such as bias hinder their large-scale adoption. Bias may arise with respect to protected attributes, including well-studied factors like gender, race, and age. In this paper, we introduce ARC, a tool to rate AI models for robustness, encompassing both bias and robustness against perturbations, along with accuracy through a causal lens. The main objective of the tool is to assist developers in building better models and aid end-users in making informed decisions based on the available data. The tool is extensible and currently supports four different AI tasks: binary classification, sentiment analysis, group recommendation, and time-series forecasting. It allows users to select data for a task and rate AI models for robustness, assessing their stability against perturbations while also identifying biases related to protected attributes. The rating method is model-independent, and the ratings produced are causally interpretable. These ratings help users make informed decisions based on the data at hand. The demonstration video is available at: https://tinyurl.com/muwujxfv.

1 Introduction

Most AI models deployed in critical domains such as healthcare and education are blackbox [Adadi and Berrada, 2018; Durán and Jongsma, 2021; Pedreschi *et al.*, 2019; Srinivasu *et al.*, 2022]. They learn correlations between data attributes [Fischer *et al.*, 2020] rather than causal relationships (cause and effect), making trust and interpretability key concerns [Shin, 2021; Schmidt and Biessmann, 2019]. [Srivastava and Rossi, 2019] proposed a two-step method to rate AI models for bias, and developed a tool to explore gender bias versus accuracy trade-offs in translators [Bernagozzi *et al.*, 2021; Dutta *et al.*, 2020; Wang *et al.*, 2021]. This approach was also used to rate chatbots [Srivastava *et al.*, 2020] and search engines [Tian *et al.*, 2023]. Various statistical fairness definitions [Hedden, 2021; Li *et al.*, 2022; Li *et al.*, 2023; Pitoura *et al.*, 2022; Zhang and Wang, 2021], such as statistical parity [Yao and Huang, 2017] and equalized odds [Garg *et al.*, 2020], have been proposed, but they are often mutually exclusive and insufficient [Verma and Rubin, 2018]. Causality-based fairness [Ehyaei *et al.*, 2023; Su *et al.*, 2022] aligns better with human values and fosters collaboration with social sciences [Carey and Wu, 2022].

In [Lakkaraju, 2022; Srivastava *et al.*, 2023], we introduced the idea of rating AI models through a causal lens. This method was applied to sentiment analysis systems (SASs) [Lakkaraju *et al.*, 2024b] and to composite systems combining SASs with translators [Lakkaraju *et al.*, 2023]. In [Lakkaraju *et al.*, 2024a], the method was extended to assess time-series forecasting models (TSFMs), evaluating their robustness by accounting for bias, as well as the impact of perturbations on their outcomes. Despite recent advances, there is no existing tool that uses causal analysis to rate AI models across multiple robustness dimensions and support model selection. To address this gap, we introduce ARC (AI Rating through Causality) tool. Beyond accuracy, robustness is a key component of trustworthiness [Chander *et al.*, 2024; Wei and Liu, 2024]. ARC estimates both and offers a comprehensive rating system for comparing AI models.

The key benefits of ARC are that it (a) helps users select the most suitable AI model for a specific task, and (b) is designed to support a range of AI tasks and models, with potential for extension to new tasks. We, hence, make the following contributions: **(1)** Present a general, extensible rating tool that supports various AI models based on causal analysis with a choice of metrics. **(2)** Demonstrate its applicability in multiple domains, including classification, sentiment analysis, group recommendation, and time-series forecasting. **(3)** Discuss how ratings generated by the tool can support informed AI model selection.

2 Problem

In this section, we introduce the generalized causal model used by ARC and the key research questions it could answer.

2.1 Preliminaries

Let an AI model take input attributes X (which can be unimodal or multi-modal) and predict an output $\hat{Y}$. The model functions as a black box:$\hat{Y} = f(T_n(X); \theta)$, where $f(.)$ represents a pre-trained AI model with parameters θ, and

*Stands for AI Rating through Causality.

$T_n : X \rightarrow X$ represents the family of treatments (or perturbations) that, when applied, transforms the input. T_0 represents the control group (no perturbation). Each T_n represents an input modification such as adding noise, changing words in text, or altering image properties. The input X may inherently encode sensitive information. Let Z denote the protected attributes (e.g., gender, race), which can either explicitly appear in X or be inferred through proxies. The AI model's outcome, denoted as O, is derived from $\hat{Y}$ and varies based on the application. For example, in regression tasks, $O = |\hat{Y} - Y|$ represents the residual error.

2.2 Causal Model

The causal model $\mathcal{M}$ (Fig. 1) illustrates causal relationships. Arrowheads indicate causal direction. If the *Protected Attribute* (Z) influences both *Treatment* (T) and *Outcome* (O), it introduces a *confounding effect*, creating a *backdoor path* between T and O (highlighted in red), leading to biased analysis. Backdoor adjustment techniques can mitigate this effect [Liu *et al.*, 2021; Xu and Gretton, 2022; Fang *et al.*, 2024]. The adjusted distribution, $(O|do(T))$, captures the true causal effect of T on O. Solid '?' arrows in Fig. 1 indicate causal links that can be validated using our tool, while the dotted arrow represents a potential causal link affected by how T varies with Z.

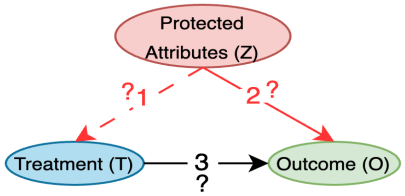


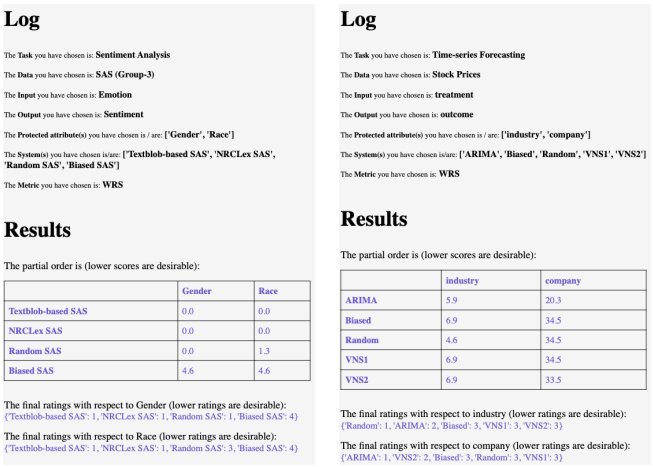
Figure 1: Our proposed generalized causal model. The validity of link '1' depends on the conditional distribution $(T|Z)$, while the validity of the links '2' and '3' can be tested using the evaluation metrics.

2.3 Key Research Questions ARC Can Help Answer

- **RQ1:** *Does Z influence O, even when Z has no effect on T?* evaluates the statistical bias exhibited by the model.
- **RQ2:** *Does Z affect the relationship between T and O when Z influences T?* evaluates the confounding bias exhibited by the model.
- **RQ3:** *Does T affect O when Z may also influence O?* examines the causal effect of treatments on the model's outcome in the presence of potential direct influence from the protected attributes, measuring robustness under varying treatments.
- **RQ4:** *Does T affect the accuracy of the model?* evaluates the impact of treatments on model performance by evaluating changes in task-specific accuracy metrics.

3 System Demonstration

The ARC tool was built using the Django framework and is available here: http://casy.cse.sc.edu/causal_rating. Table 1 summarizes the tasks, datasets, attributes, and AI models.



(a) Sentiment Analysis: TextBlob and NRCLex exhibited no measurable bias.

(b) Time-Series Forecasting: ARIMA exhibited the least bias, while VNS1 showed the most.

Figure 2: Weighted Rejection Score (WRS) quantifies statistical bias, helping to answer RQ1 across different AI models. While WRS was computed for multiple tasks, we present results for two representative ones here.

This section details the rating workflow, represented in the flowchart (Fig. 3) and demonstrated in the tool (Fig. 4). We use Time-Series Forecasting as a running example, with further details available in [Lakkaraju *et al.*, 2024a].

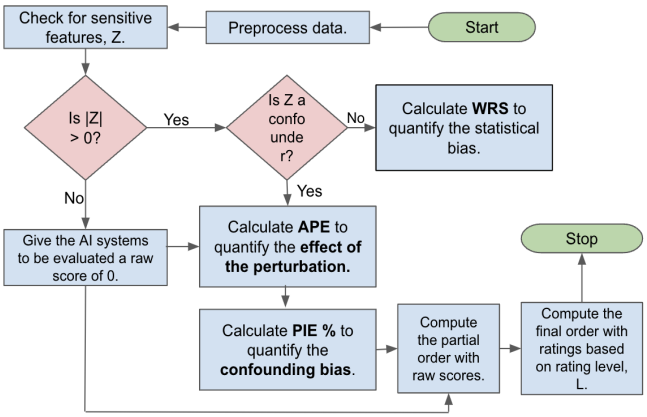


Figure 3: Workflow for performing statistical and causal analysis to compute raw scores and assign final ratings to the test models.

1. Select a Task (Fig. 4a): The user starts by choosing a task, such as *Binary Classification, Sentiment Analysis, Group Recommendation*, or *Time-Series Forecasting*.
2. Choose a Dataset (Fig. 4b): The user selects a dataset relevant to the chosen task.
3. Choose Attributes (Fig. 4c): The user specifies *input, output,* and *protected attributes* that will be used in the analysis.
4. Select AI Models (Fig. 4d): The user picks one or more AI models from the available options for comparison.
5. Choose Evaluation Metrics (Fig. 4e): The user selects evaluation metrics that can be used to answer research ques-

Tasks	Data	Attributes	Metrics	Models
Binary Classification	German Credit Dataset [Dua and Graff, 2017].	**Treatment**: Credit Amount (low, medium, high); **Protected**: Age, Gender; **Outcome**: Risk (good/bad).	WRS, DIE %, and PIE %	Logistic Regression, Random
Sentiment Analysis (SAS)	EEC Dataset [Kiritchenko and Mohammad, 2018] with emotion word variations and protected attributes (Gender, Race).	**Treatment**: Emotion Word (positive, negative); **Protected**: Gender, Race; **Outcome**: Sentiment.	WRS, DIE %, and PIE %	TextBlob, NRCLex, Biased, Random
Group Recommendation	Public data from funding agencies (RFPs) and researcher profiles [Valluru *et al.*, 2024a; Valluru *et al.*, 2024b].	**Treatment**: Request For Proposals (RFPs) and researcher profiles; **Protected**: Gender; **Outcome**: Goodness Scores (for recommended teams).	WRS	Random Matching (*M0*), String Matching (*M1*), Semantic Matching (*M2*), Boosted Bandit Learning (*M3*)
Time-series Forecasting (TSFM)	Stock prices (Mar 2023 - Apr 2024) from *Yahoo! Finance*.	**Treatment**: Semantic, Input-specific, and Composite perturbations; **Protected**: Company, Industry; **Outcome**: Residual.	WRS, DIE %, PIE %, APE, and Accuracy Metrics	ARIMA, Random, Biased, ViT-num-spec-large (*VNS1*), ViT-num-spec-small (*VNS2*)

Table 1: Summary of tasks that include Binary Classification, Sentiment Analysis [Lakkaraju *et al.*, 2023; Lakkaraju *et al.*, 2024b], Group Recommendation [Srivastava *et al.*, 2022; Valluru *et al.*, 2024c; Nagpal *et al.*, 2024b; Nagpal *et al.*, 2024a], and Time-series Forecasting [Lakkaraju *et al.*, 2024a], data attributes, evaluation metrics, AI models, and references with implementation details used in the ARC tool.

tions stated in Section 2, including robustness metrics such as the Weighted Rejection Score (WRS) (to answer RQ1), Deconfounding Impact Estimation (DIE %), (to answer RQ2) Propensity Score Matching - DIE (PIE %) (to answer RQ2), Average Perturbation Effect (to answer RQ3), and Forecasting Accuracy metrics (to answer RQ4). The tool provides the description of these metrics in a popup block as shown in Fig. 4e. The complete formulation of these metrics can be found in [Lakkaraju *et al.*, 2024a].

6. View Results (Fig. 4f): The tool presents a log of user selections, causal analysis results, and a causal diagram. **ARC provides both nuanced raw scores and final ratings for easy interpretation and comparison across AI models.**

4 Discussion

In this paper, we applied our robustness assessment tool to four diverse tasks and showed that the rating methodology generalizes well.

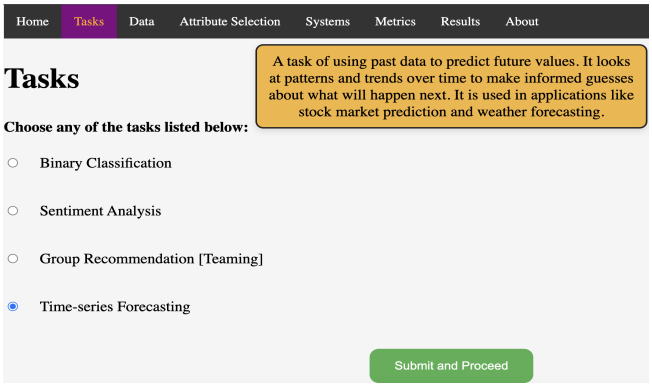
ARC uncovered the following task-specific observations: **1. The German Credit dataset is widely recognized as biased with respect to gender and age [Liao and Naghizadeh, 2023; Bhargava *et al.*, 2020]. ARC was able to identify statistical and confounding biases; 2. For SASs, ARC identified statistical and confounding biases related to gender and race, with TextBlob and NRCLex emerging as the least-biased; 3. In group recommender systems for team settings, we identified statistical bias concerning the protected attribute *gender*, where *M2* emerged as most biased; 4. *VNS1* and *VNS2* exhibited least confounding bias, whereas ARIMA showed least statistical bias.** The proposed rating methodology not only estimates attributes driving bias but also quantifies these biases. By providing end-users with valuable insights, the tool helps them make informed choices among the available models. As future work, we plan to conduct user studies to evaluate the tool's usability.

Limitations. A central limitation of ARC is its reliance on a predefined causal model, which requires domain knowledge and may not reflect the true underlying data-generating process. It is rarely possible to specify a perfect causal model. ARC begins with a hypothesized structure in which protected attributes act as confounders, and it explores different scenarios within this structure. In practice, such models are often informed by expert knowledge, controlled studies, or causal discovery algorithms that rely on statistical dependencies. Even when imperfect, a causal model provides a principled framework for reasoning about interventions and counterfactuals, capabilities that purely statistical methods lack. Causal models are designed to answer "what if I intervene?" questions, which are especially critical in the presence of confounding (e.g., Simpson's paradox [Pearl, 2022]).

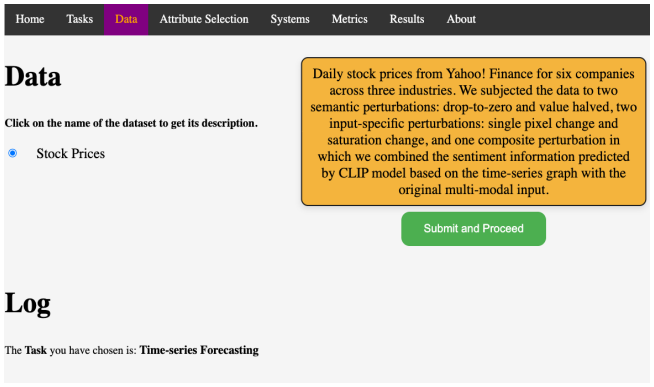
A second limitation is that ARC uses a fixed causal graph structure across all tasks. This means we do not account for domain-specific differences in causal structure, which could matter in some settings. That said, keeping the structure fixed has a practical upside: it lets us compare models under the same set of assumptions, without the results being driven by changes in the causal setup itself. In future work, it would be useful to explore how ARC behaves when the graph is adapted to specific domains or learned from data.

5 Acknowledgments

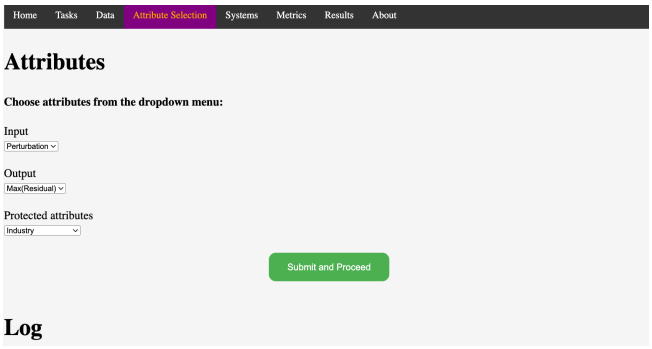
We acknowledge funding support from Cisco Research and the South Carolina Research Authority (SCRA).



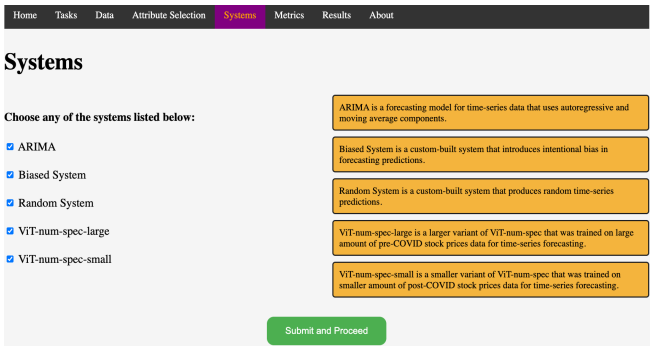
(a) Task Selection



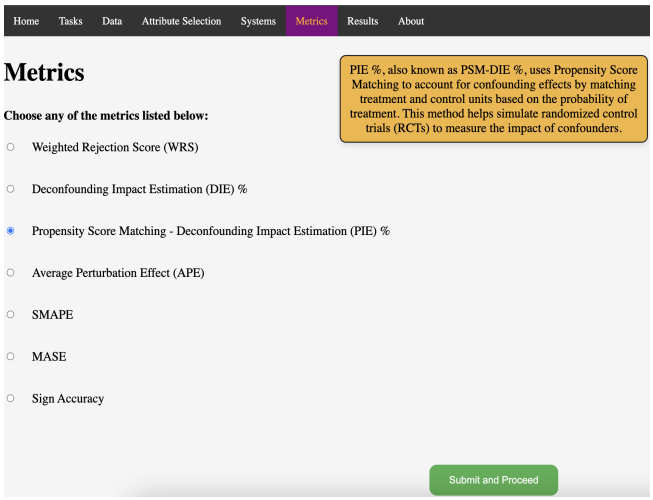
(b) Dataset Selection



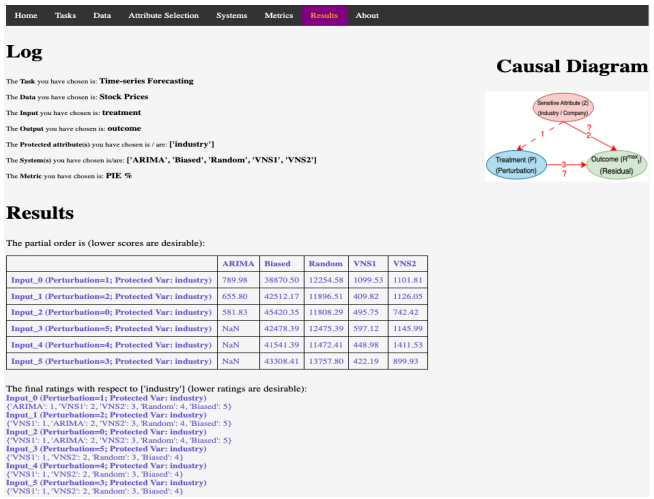
(c) Attributes Selection



(d) Models Selection



(e) Metric Selection



(f) Displayed Results

Figure 4: Figure showing step-by-step workflow of the ARC tool, illustrating selection of task, dataset, attributes, AI models, and metric. The result will be displayed at the end.

References

[Adadi and Berrada, 2018] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: a survey on explainable artificial intelligence (xai). *IEEE access*, 6:52138–52160, 2018.

[Bernagozzi *et al.*, 2021] Mariana Bernagozzi, Biplav Srivastava, Francesca Rossi, and Sheema Usmani. Vega: a virtual environment for exploring gender bias vs. accuracy trade-offs in ai translation services. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(18):15994–15996, May 2021.

[Bhargava *et al.*, 2020] Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli. Limeout: An ensemble approach to improve process fairness, 2020.

[Carey and Wu, 2022] Alycia N Carey and Xintao Wu. The causal fairness field guide: Perspectives from social and formal sciences. *Frontiers in Big Data*, 5, 2022.

[Chander *et al.*, 2024] Bhanu Chander, Chinju John, Lekha Warrier, and Kumaravelan Gopalakrishnan. Toward trustworthy artificial intelligence (tai) in the context of explainability and robustness. *ACM Computing Surveys*, 2024.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Durán and Jongsma, 2021] Juan Manuel Durán and Karin Rolanda Jongsma. Who is afraid of black box algorithms? on the epistemological and ethical basis of trust in medical ai. *Journal of Medical Ethics*, 47(5):329–335, 2021.

[Dutta *et al.*, 2020] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International conference on machine learning*, pages 2803–2813. PMLR, 2020.

[Ehyaei *et al.*, 2023] Ahmad-Reza Ehyaei, Golnoosh Farnadi, and Samira Samadi. Causal fair metric: Bridging causality, individual fairness, and adversarial robustness. *arXiv preprint arXiv:2310.19391*, 2023.

[Fang *et al.*, 2024] Junpeng Fang, Gongduo Zhang, Qing Cui, Caizhi Tang, Lihong Gu, Longfei Li, Jinjie Gu, and Jun Zhou. Backdoor adjustment via group adaptation for debiased coupon recommendations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 11944–11952, 2024.

[Fischer *et al.*, 2020] Lukas Fischer, Lisa Ehrlinger, Verena Geist, Rudolf Ramler, Florian Sobiezky, Werner Zellinger, David Brunner, Mohit Kumar, and Bernhard Moser. Ai system engineering—key challenges and lessons learned. *Machine Learning and Knowledge Extraction*, 3(1):56–83, 2020.

[Garg *et al.*, 2020] Pratyush Garg, John Villasenor, and Virginia Foggo. Fairness metrics: A comparative analysis. In *2020 IEEE international conference on big data (Big Data)*, pages 3662–3666. IEEE, 2020.

[Hedden, 2021] Brian Hedden. On statistical criteria of algorithmic fairness. *Phil. & Pub. Aff.*, 49:209, 2021.

[Kiritchenko and Mohammad, 2018] Svetlana Kiritchenko and Saif Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[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, 2022] Kausik Lakkaraju. 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*, AIES '22, page 902, New York, NY, USA, 2022. Association for Computing Machinery.

[Li *et al.*, 2022] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: A survey. *arXiv preprint arXiv:2205.13619*, 2022.

[Li *et al.*, 2023] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, Juntao Tan, Shuchang Liu, and Yongfeng Zhang. Fairness in recommendation: Foundations, methods, and applications. *ACM Transactions on Intelligent Systems and Technology*, 14(5):1–48, 2023.

[Liao and Naghizadeh, 2023] Yiqiao Liao and Parinaz Naghizadeh. Social bias meets data bias: The impacts of labeling and measurement errors on fairness criteria, 2023.

[Liu *et al.*, 2021] Taoran Liu, Winghei Tsang, Yifei Xie, Kang Tian, Fengqiu Huang, Yanhui Chen, Oiying Lau, Guanrui Feng, Jianhao Du, Bojia Chu, et al. Preferences for artificial intelligence clinicians before and during the covid-19 pandemic: discrete choice experiment and propensity score matching study. *Journal of medical Internet research*, 23(3):e26997, 2021.

[Nagpal *et al.*, 2024a] Vansh Nagpal, Siva Likitha Valluru, Kausik Lakkaraju, Nitin Gupta, Zach Abdulrahman, Andrew Davison, and Biplav Srivastava. A novel approach to balance convenience and nutrition in meals with long-term group recommendations and reasoning on multi-

modal recipes and its implementation in beacon. *arXiv preprint arXiv:2412.17910*, 2024.

[Nagpal *et al.*, 2024b] Vansh Nagpal, Siva Likitha Valluru, Kausik Lakkaraju, and Biplav Srivastava. Beacon: Balancing convenience and nutrition in meals with long-term group recommendations and reasoning on multimodal recipes. *arXiv preprint arXiv:2406.13714*, 2024.

[Pearl, 2022] Judea Pearl. *Comment: Understanding Simpson's Paradox*, page 399–412. Association for Computing Machinery, New York, NY, USA, 1 edition, 2022.

[Pedreschi *et al.*, 2019] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box ai decision systems. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 9780–9784, 2019.

[Pitoura *et al.*, 2022] Evaggelia Pitoura, Kostas Stefanidis, and Georgia Koutrika. Fairness in rankings and recommendations: an overview. *The VLDB Journal*, pages 1–28, 2022.

[Schmidt and Biessmann, 2019] Philipp Schmidt and Felix Biessmann. Quantifying interpretability and trust in machine learning systems. *arXiv preprint arXiv:1901.08558*, 2019.

[Shin, 2021] Donghee Shin. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. *International journal of human-computer studies*, 146:102551, 2021.

[Srinivasu *et al.*, 2022] Parvathaneni Naga Srinivasu, N Sandhya, Rutvij H Jhaveri, and Roshani Raut. From blackbox to explainable ai in healthcare: existing tools and case studies. *Mobile Information Systems*, 2022(1):8167821, 2022.

[Srivastava and Rossi, 2019] Biplav Srivastava and Francesca Rossi. Towards composable bias rating of ai services, 2019.

[Srivastava *et al.*, 2020] Biplav Srivastava, Francesca Rossi, Sheema Usmani, and Mariana Bernagozzi. Personalized chatbot trustworthiness ratings. *IEEE Transactions on Technology and Society*, 1(4):184–192, 2020.

[Srivastava *et al.*, 2022] Biplav Srivastava, Tarmo Koppel, Sai Teja Paladi, Siva Likitha Valluru, Rohit Sharma, and Owen Bond. Ultra: A data-driven approach for recommending team formation in response to proposal calls. In *2022 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1002–1009. IEEE, 2022.

[Srivastava *et al.*, 2023] Biplav Srivastava, Kausik Lakkaraju, Mariana Bernagozzi, and Marco Valtorta. Advances in automatically rating the trustworthiness of text processing services, 2023.

[Su *et al.*, 2022] Cong Su, Guoxian Yu, Jun Wang, Zhongmin Yan, and Lizhen Cui. A review of causality-based fairness machine learning. *Intelligence & Robotics*, 2(3):244–274, 2022.

[Tian *et al.*, 2023] Xinran Tian, Bernardo Pereira Nunes, Katrina Grant, and Marco Antonio Casanova. Mitigating bias in glam search engines: A simple rating-based approach and reflection. In *Proceedings of the 34th ACM Conference on Hypertext and Social Media*, HT '23, New York, NY, USA, 2023. Association for Computing Machinery.

[Valluru *et al.*, 2024a] Siva Likitha Valluru, Biplav Srivastava, Sai Teja Paladi, Siwen Yan, and Sriraam Natarajan. Promoting research collaboration with open data driven team recommendation in response to call for proposals. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 22833–22841, 2024.

[Valluru *et al.*, 2024b] Siva Likitha Valluru, Michael Widener, Biplav Srivastava, and Sugata Gangopadhyay. Ultra: Exploring team recommendations in two geographies using open data in response to call for proposals. In *Proceedings of the 7th Joint International Conference on Data Science & Management of Data (11th ACM IKDD CODS and 29th COMAD)*, pages 547–552, 2024.

[Valluru *et al.*, 2024c] Siva Likitha Valluru, Michael Widener, Biplav Srivastava, Sriraam Natarajan, and Sugata Gangopadhyay. Ai-assisted research collaboration with open data for fair and effective response to call for proposals. *AI Magazine*, 45(4):457–471, 2024.

[Verma and Rubin, 2018] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018.

[Wang *et al.*, 2021] Jialu Wang, Yang Liu, and Xin Eric Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*, 2021.

[Wei and Liu, 2024] Wenqi Wei and Ling Liu. Trustworthy distributed ai systems: Robustness, privacy, and governance. *ACM Computing Surveys*, 2024.

[Xu and Gretton, 2022] Liyuan Xu and Arthur Gretton. A neural mean embedding approach for back-door and front-door adjustment. *arXiv preprint arXiv:2210.06610*, 2022.

[Yao and Huang, 2017] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. *Advances in neural information processing systems*, 30, 2017.

[Zhang and Wang, 2021] Dell Zhang and Jun Wang. Recommendation fairness: From static to dynamic. *arXiv preprint arXiv:2109.03150*, 2021.