Supplementary Material for: Mitigating Bias in Algorithmic Systems—A Fish-eye View

KALIA ORPHANOU, Open University of Cyprus
JAHNA OTTERBACHER and STYLIANI KLEANTHOUS, Open University of Cyprus & CYENS Centre of Excellence
KHUYAGBAATAR BATSUREN, National University of Mongolia
FAUSTO GIUNCHIGLIA, The University of Trento
VERONIKA BOGINA, AVITAL SHULNER TAL, ALAN HARTMAN, and TSVI KUFLIK, The University of Haifa

1 SUPPLEMENTARY MATERIALS

Table 7 summarizes the methods used for auditing and discrimination discovery within each of the research domains analyzed in this survey. In ML systems, bias detection is mostly done using discrimination or fairness metrics. Auditing in ML systems can be achieved by auditing software tools or when *developers/regulators* act as auditors of the algorithmic system. However, in IR, HCI, and RecSys systems, *users* often act as auditors by submitting different queries in search engines and social networks or by taking the role of crowdworker in the crowdsourcing conducted studies. Discrimination discovery approaches used in IR, HCI, and RecSys systems are similar to auditing but with a more concrete methodology on detecting bias.

Table 8 summarizes the methods used for fairness management within each of the research domains analyzed in this survey. In ML algorithmic systems, the most popular techniques are data re-sampling, removal of sensitive attributes and data transformation to mitigate bias in the data, optimization and regularization approaches to mitigate bias during the model training and re-labeling of the outcome decision to mitigate bias on the output of the system. In ranking systems such as RecSys and IR, the most popular approaches are re-sampling for mitigating data bias, learning to rank methods to mitigate bias in the ranking algorithms and re-ranking methods as for modifying the ranking outcomes. Two approaches that are common in RecSys and ML communities are the data transformation (fairness pre-processing) and optimization approaches (fairness in-processing). In the HCI community, since the *user* is the main stakeholder, most of the papers examine the user perception on fairness. Approaches to mitigate bias referred to the use of a human-in-the-loop on the decision-making [22]. Fairness certification techniques use fairness constraints or defining new fairness notions, i.e., counterfactual fairness and metrics for certifying the fairness of systems in all the four research domains. In IR, some studies also use user evaluation to certify the fairness of the system.

Table 9 provides a comparison of the solutions focusing on Explainability Management. Explainability approaches have primarily been developed in the context of ML algorithms and systems. The best known methods for explaining the model decision-making process use interpretable models to mimic the behavior of black-box models, i.e., decision trees, decision rules, and ontologies.

0360-0300/2022/12-ART87 $15.00
https://doi.org/10.1145/3527152

Table 7. Comparison of Discrimination Detection Approaches Across the Four Domains

Domain	Problem	Solution Space	Reference(s)
Bias Detection			
ML	Data/Model	Auditing	Automatic auditing tool [14, 174]
			Developers as auditors [24, 130, 226]
	Data		Discrimination/Fairness metrics [98, 222, 233]
	Data		Metrics [56, 125, 154]
	Data/Model/Output	Discrimination Discovery	ML methods [43, 49, 123, 157, 226, 234]
IR	User/Data/Output	Auditing	Submit queries to search engines/Twitter [92, 105, 119, 122, 132, 147, 204]
	Model/User		Sock-puppet auditing [6]
	User/Data/Output	Discrimination Discovery	Analysis of Web logs [13, 35, 94, 150, 208, 211–213, 221]
	User/Data/Output		Word embedding [66, 95, 164]
	User/Third Party/Data		Crowdsourcing studies [59, 127]
	User/Third Party		Direct discrimination of perceived bias [10, 96, 209, 210]
HCI	Output/Model/User	Auditing	Analysing system behavior [101, 135]
	Data/User/Third Party	Discrimination Discovery	Crowdsourcing studies [11, 48, 78, 139, 161]
	Model/User		Use of ML methods [89, 178]
	Data/User		Data-driven personas [175]
RecSys	Data/User	Auditing	Developers as auditors [57, 62]
	Model/User		Sock-puppet auditing [6]
	User/Model/Output	Discrimination Discovery	Discrimination detection in advertising recommendation systems [2, 188, 192]
	Output/Model		Discrimination detection in evaluation metrics [15, 60]
	Output/User		Discrimination in social networks [34]

Methods for explaining the decision outcome include feature-relevance, local and global explainability, and visualization methods. There is also a growing literature on explainability within the HCI community. These works suggest that explainability, and judgement of the outcome or decision of the system should be provided to enhance the trust of the end user in the system. Also in HCI, we found a few works that connect explainability to fairness perception. Finally, explainability approaches have also been widely discussed in RecSys and IR systems. The difference between these approaches and the ones used in ML are that they take into consideration the user's perception and have the specific goal of increasing the trust of the end user in the system. The most popular explainability techniques in the RecSys and IR literature are the visualization methods (outcome explainability) that have been applied to justify the ranking results.

Table 8. Comparison of Fairness Management Methods in the Different Domains

Fairness Management			
Domain	**Problem**	**Solution Space**	**Reference(s)**
ML	Data	Fairness Pre-processing	Removal of protected attributes & Data Transformation [26, 100, 158, 224]
			Causal BN [121, 226]
			Data Re-labeling [65, 102]
			Re-sampling methods [101, 182]
	Model	Fairness In-processing	Regularization approach [103, 219]
			Optimization approach [144, 173]
			Constraints [165]
	Model/Output		Counterfactual fairness [120]
	Third Party/Output	Fairness Post-processing	Altering of labels [84, 102, 157]
	User/Third Party	Fairness Perception	[134, 189]
	Data/Model/Output	Fairness Certification	Fairgroups [64]
			Counterfactual Fairness [109, 182]
			Techno-moral graphs [97]
			Fairness Constraints/Metrics [31, 46, 52, 79, 108, 111, 216, 225]
IR	Data	Fairness Pre-processing	Data sampling [51, 53, 76, 184]
	Model	Fairness In-processing	Learn-to-rank methods [47, 117, 149, 220]
	Output	Fairness Post-processing	Re-ranking [104, 110, 118, 126]
	Model/Output/User	Fairness Certification	[61, 90, 141]
	User/Output	Fairness Perception	[136, 152]
HCI	Data	Fairness Pre-processing	Data sampling [101]
			Data transformation [32]
	Output	Fairness Perception	Human-in-the-loop [22]
	User/Output		Metrics [206]
	Output/User	Fairness Certification	[124, 214]
RecSys	Data	Fairness Pre-processing	Data sampling [25, 104, 130]
			Data transformation [215]
	Model/Output		Optimization approaches [138, 217]
	Model	Fairness In-processing	Learn-to-rank [117, 220]
	Output	Fairness Post-processing	Re-ranking [104, 155, 186, 223]
	Model/Output	Fairness Certification	Metrics [106]

Table 9. Comparison of Explainability Management Approaches for the Different Research Domains

Explainability Management			
Domain	**Problem**	**Solution Space**	**Reference(s)**
ML	Model	Model Expainability	Use of decision tree [38, 54, 74, 115, 177, 193, 230]
	Model		Use of decision rules [44, 99, 128]
	Model		Ontologies [19, 42, 166]
	Output	Outcome Explainability	Local explanations [167, 168, 200]
	Output/User		Visualization methods [17, 67, 180, 185, 218, 230, 232]
	Output/User		Counterfactual explanations [182, 206]
			Feature-relevance explanations [1, 87, 187, 203]
IR	Output/User	Outcome Explainability	Global explanaions [9]
HCI	User/Data	Model Explainability	Data-centric explanations [5]
	Output/Data	Outcome Explainability	Feature-relevance explanation [91]
	User/Output		Taxonomy of explanations & Styles [16, 58, 69]
	User/Output		Raise user awareness [162]
RecSys	Model/User	Model Explainability	Taxonomy of concepts [145]
	Model/User		Based on user opinions [37, 205]
	Output/User		Personalized explanations [151]
	Output/User		Knowledge graph [29, 86]
	Output/User	Output Explainability	Visualization methods [20, 114, 198, 201]