Why We Need Personalized AI-Advisors for Biased Human Decision-Makers (and how to produce them)

Nicholas Wolczynski[1], Maytal Saar-Tsechansky[1], Tong Wang[2]
[1]McCombs School of Business, The University of Texas,
[2]School of Management, Yale University

GOOD SYSTEMS
Ethical AI at UT Austin

The University of Texas at Austin
Research Development
Vice President for Research, Scholarship and Creative Endeavors

The University of Texas at Austin
McCombs School of Business

Yale SCHOOL OF MANAGEMENT

Introduction

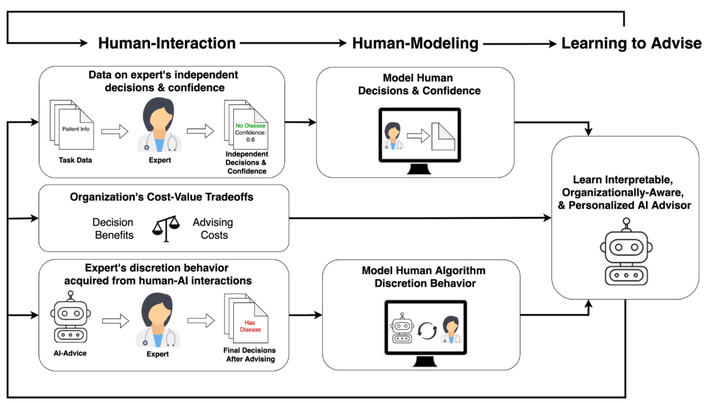
While there have been growing concerns regarding bias in AI-driven decisions, a vast array of high-stakes decisions, such as medical and judicial decisions, will continue to be undertaken by human decision-makers (DMs) in the foreseeable future. Importantly, human experts, such as physicians, have been shown to produce racially-biased[1] and gender-biased[2] decisions. In high-stakes AI-assisted decision-making (AIaDM) settings, experts must also reconcile advice from AI systems before making final decisions[3]. Even when AI models possess similar or superior accuracy to human DMs, their advice rarely improves team decision-making outcomes in practice[4]. We first identify distinct properties of AIaDM settings which are key to developing AIaDM models that can effectively benefit team performance and reduce decision-making bias: (1) DMs often exhibit imperfect and biased decision-making and algorithm discretion behaviors, (2) DMs face reconciliation costs from exerting decision-making resources (e.g., time and effort) when AI advice contradicts their own judgment, (3) DMs require interpretability to understand the reasoning behind the AI's advice. Building on the properties above, we propose the AIaDM-learning framework and use it to develop the TeamRules (TR) algorithm for producing organizationally-aware and personalized rule-based advisors. Using data on different decision-making domains, we simulate varied and biased human decision-making and algorithm discretion behaviors and demonstrate how TR generated advisors consistently produce either superior or otherwise comparable team decisions relative to alternatives.

Research Goal

Our goal is to develop personalized and organizationally-aware advisors that reduce decision-making bias and improve team decision-making outcomes when paired with biased, high-stakes decision-makers.

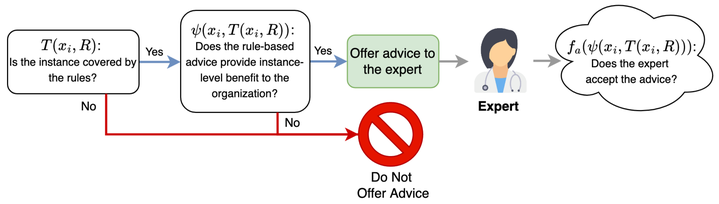
Framework

We propose the AI-assisted Decision Making (AIaDM)-Learning Framework. This framework includes three iterative phases to develop an organizationally-aware and personalized AI advisor for a human DM in AI-advised Decision-Making contexts.



Methods

We utilize the AIaDM-learning framework to develop TeamRules (TR), an algorithm for developing an organizationally-aware and personalized rule-based advisor. Below is a visualization of TR in deployment:



TR models are produced using a simulated annealing algorithm which optimizes an objective which can be broken down into:

1. Discretion-Adjusted Decision-Loss: Decision-loss from accepted and rejected advice (given human's expected decisions and expected algorithm discretion behavior).
2. Cost of Advising: Costs incurred from reconciling contradictory advice.

The objective accounts for the decision-maker's expected behavior both when advice is provided and accepted, and when advice is not provided or not accepted. By doing so, we can minimize expected loss from the *team's final decisions.*

Experiments

Data:
We pair our TR-generated and alternative advisors with simulated human behaviors on real-world decision-making tasks. We include results for the Heart Disease and Adult task datasets when paired with a human that exhibits decision bias for each task. The human simulated behaviors are generally described as follows:

Task	Sub-Population	Human Accuracy
Heart Disease	Male	High
	Female	Low
Income Prediction	White	High
	Non-White	Low

Alternative Frameworks:
We compare TR to rule-based models produced by alternative frameworks:

1. **Task-Only:** AI optimized for decision accuracy on the task only (current practice).
2. **TR-No(ADB, OrgVal):** AI optimized for the task while accounting for DM's decision expertise.
3. **TR-No(ADB):** AI optimized for the task, accounts for the DM's decision expertise, and accounts for costs incurred from contradictory advice (does not consider DM's algorithm discretion behavior).

***Full comparisons available in paper. Poster includes comparison to the Task-Only framework which is predominantly used currently in practice.

Evaluation Metrics:
Team Decision Loss (TDL): Loss from final team decisions (after advising)
Costs Incurred: Loss from reconciling contradictory advice (cost from single contradiction * number of contradictions).
Total Team Loss (TTL): TDL + Costs Incurred

Results

Below we demonstrate how TR produces personalized advice to improve human decisions in the expert's area of weakness while maintaining performance in their area of strength. TDL and TTL outcomes below are results after advice has been provided to expert DM, and final team decision has been made.

Income Prediction Task Reconciliation Cost: 0.0	Non-White		White		Entire Population	
	TeamRules	Task-Only	TeamRules	Task-Only	TeamRules	Task-Only
Advising Rate	0.38	1.00	0.20	1.00	0.23	1.00
Contradiction Rate	0.37	0.43	0.12	0.22	0.16	0.25
Contradiction Accuracy	0.90	0.83	0.28	0.19	0.50	0.35
Improvement in TTL w.r.t. Human	**+0.019**	+0.017	**-0.003**	-0.011	**+0.019**	+0.006

Heart Disease Classification Task Reconciliation Cost: 0.2	Female		Male		Entire Population	
	TeamRules	Task-Only	TeamRules	Task-Only	TeamRules	Task-Only
Advising Rate	0.34	1.00	0.05	1.00	0.11	1.00
Contradiction Rate	0.33	0.44	0.01	0.24	0.08	0.29
Contradiction Accuracy	0.86	0.73	0.28	0.16	0.82	0.35
Improvement in TDL w.r.t. Human	**+0.045**	+0.039	**0.000**	-0.010	**+0.045**	+0.029
Reconciliation Costs Incurred w.r.t. Human	**-0.014**	-0.019	**-0.001**	-0.039	**-0.015**	-0.058
Improvement in TTL w.r.t. Human	**+0.031**	+0.020	**-0.001**	-0.049	**+0.030**	-0.029

TR provides advice selectively, advising less than traditional approaches optimized just for the task. Further, TR focuses on producing advice in the DM's area of weakness and avoids advising in their area of strength. By giving selective, higher accuracy advice for instances on which the DM has lower accuracy, TR improves team decisions and leads to fairer outcomes. This occurs because TR improves the DM's accuracy for sub-populations on which the DM tends to perform worse, while reducing the harm done to team decisions for sub-populations on which the DM performs better.

Conclusion

In this work, we identify key properties of AI that can benefit AI systems' ability to learn to advise experts in high-stakes AIaDM settings. Methods that advise humans can improve team outcomes if they include mechanisms that leverage the individual human partner's unique expertise and optimize for the final organizational value provided by the advice. By personalizing to the DM's expertise, AI advising methods can selectively give advice for instances that are part of a sub-population that is originally disadvantaged by the DM's decisions, improving outcomes for those sub-populations while minimizing harm done to instances that are part of the advantaged sub-population.

Acknowledgments

We would like to thank Good Systems for providing the funding that makes this research possible and for providing the opportunity to share this work at the Good Systems Symposium.

References

[1]Breathett, Khadijah et al. "African Americans Are Less Likely to Receive Care by a Cardiologist During an Intensive Care Unit Admission for Heart Failure." *JACC. Heart failure* vol. 6,5 (2018): 413-420.
[2]Lee, K, Ferry, A, Anand, A. et al. Sex-Specific Thresholds of High-Sensitivity Troponin in Patients With Suspected Acute Coronary Syndrome. J Am Coll Cardiol. 2019 Oct, 74 (16) 2032–2043.
[3]Tamer Boyacı, Caner Canyakmaz, and Francis de Véricourt. Human and machine: The impact of machine input on decision making under cognitive limitations. Management Science, March 2023.
[4]Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. Organization Science, page orsc.2021.1549, January 2022.