PHIL 408Q/PHPE 308D
Fairness

Eric Pacuit, University of Maryland

May 9, 2024

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kühl (2024). *Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making*. In Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24).

Navita Goyal, Connor Baumler, Tin Nguyen, and Hal Daumé III (2024). *The Impact of Explanations on Fairness in Human-AI Decision-Making: Protected vs Proxy Features*. In 29th International Conference on Intelligent User Interfaces (IUI '24).

"[C]ompanies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums."

Jennifer Kite-Powell (2022). *Explainable AI is trending and here's why*. Forbes.

"[C]ompanies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums."

Jennifer Kite-Powell (2022). *Explainable AI is trending and here's why*. Forbes.

Explanations "...provide a more effective interface for the human-in-the-loop, enabling people to identify and address fairness and other issues"

Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan (2019). *Explaining models: An empirical study of how explanations impact fairness judgment*. In Proceedings of the 24th International Conference on Intelligent User Interfaces. 275–285.

In order for a human-in-the-loop to addresses fairness issues, they should have the capacity to identify mistaken recommendations, reducing the false negative errors affecting that group.

In this case, the goal of explanations should be to help humans identify such errors, yielding AI-assisted decisions that have better distributive fairness properties than the AI alone.

Note that this is different to the perceptions that humans may have of an AI system, and it also differs from the overall accuracy or reliance behavior.

Fairness through Unawareness

"Fairness through unawareness": an AI system is fair if it does not make use of information that is evidently indicative of a person's demographics.

- Neither a sufficient nor a necessary condition for fairness

Sam Corbett-Davies and Sharad Goel (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv preprint arXiv:1808.00023.

Explanation

Focus on *feature-based explanations*: LIME is used in the experiments, due to its popularity in the literature as well as in practice and, importantly, the fact that LIME has been claimed to enable fairness assessments.

Vaishnavi Bhargava, Miguel Couceiro, and Amedeo Napoli (2020). *LimeOut: An ensemble approach to improve process fairness*. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases. Springer, pp. 475 - 491.

Joymallya Chakraborty, Kewen Peng, and Tim Menzies (2020). *Making fair ML software using trustworthy explanation*. In Proceedings of the 35th IEEE/ACM International Conference on Automated Software Engineering, pp. 1229-1233.

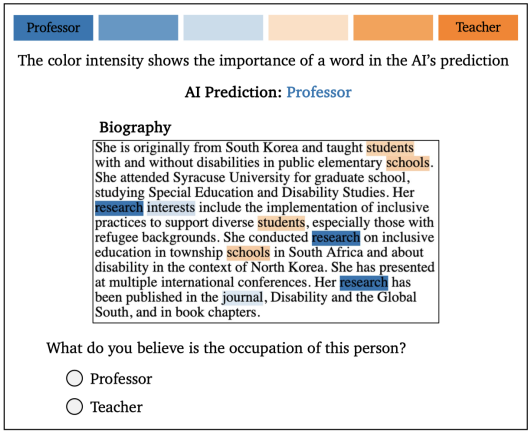
Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online. An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI systems, it is susceptible to gender bias and discrimination.

Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online. An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI systems, it is susceptible to gender bias and discrimination.

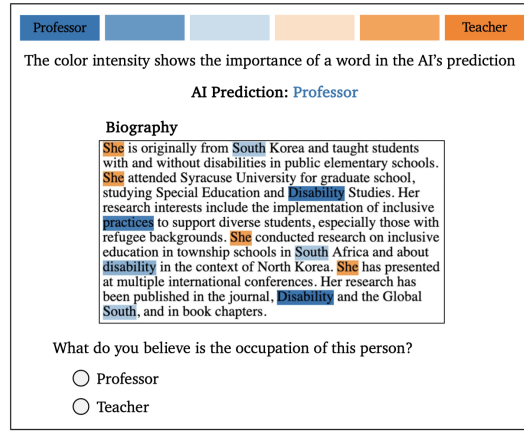
This study: instantiate an AI-assisted decision-making setup where participants see short textual bios and are asked—with the help of an AI recommendation to predict whether a given bio belongs to a professor or a teacher. Professors are historically a men-dominated occupation, whereas teachers have been mostly associated with women.

Experiment

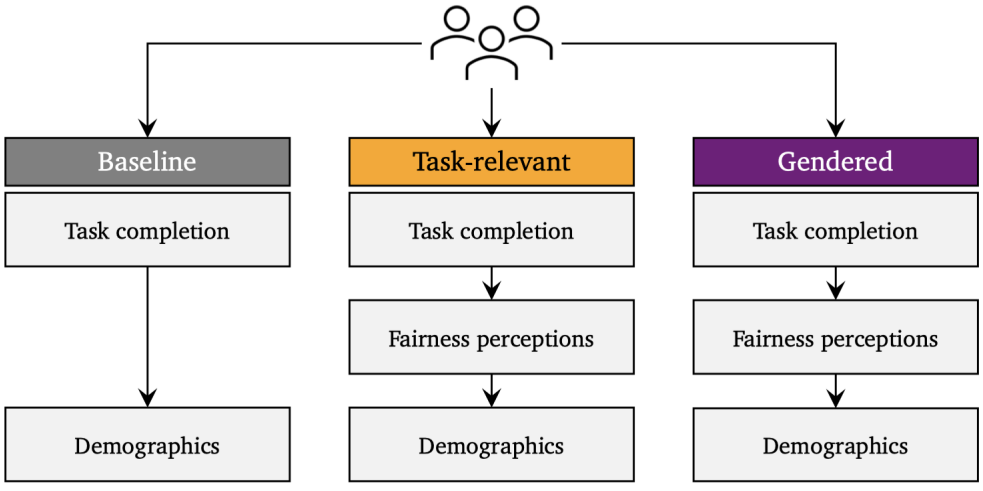
- Each participant sees 14 bios one by one, each including the AI recommendation as well as an explanation highlighting the most predictive words. We also include a baseline condition without explanations.
- Participants are assigned to conditions where they see recommendations and explanations either from (i) an AI model that uses task-relevant features, or (ii) an AI model that uses gendered (i.e., sensitive) features.
- Participants in each condition first complete the task of predicting occupations for 14 bios, and—if assigned to a condition with explanations—answer several questions regarding their fairness perceptions.



***Task-relevant* condition**



***Gendered* condition**


Baseline
Task completion
Demographics
Task-relevant
Task completion
Fairness perceptions
Demographics
Gendered
Task completion
Fairness perceptions
Demographics

Gender of bio	True occupation	AI recommendation	AI correct?	Acronym	#Bios
Woman	Teacher	Teacher	✓	WTT	3
Woman	Professor	Teacher	✗	WPT	3
Woman	Professor	Professor	✓	WPP	1
Man	Teacher	Teacher	✓	MTT	1
Man	Teacher	Professor	✗	MTP	3
Man	Professor	Professor	✓	MPP	3

Table 2: We distinguish four types of reliance in AI-assisted decision-making: humans can adhere to or override correct AI recommendations, or they can adhere to or override incorrect AI recommendations.

	Human adherence to AI	Human overriding of AI
AI correct	Correct adherence	Detrimental overriding
AI incorrect	Detrimental adherence	Corrective overriding

Results 1

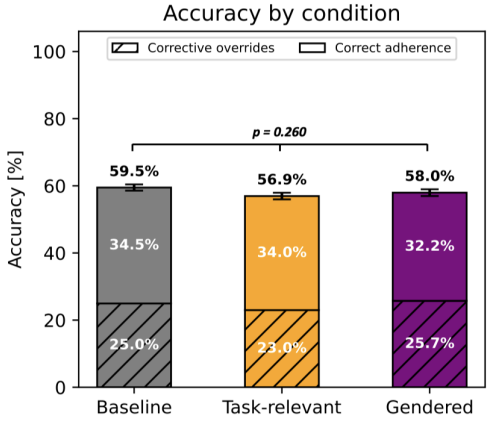


Figure 3: Accuracy is not higher when explanations are provided, compared to the baseline.

Results 2

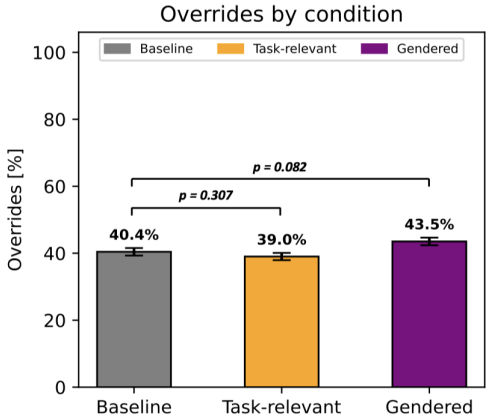


Figure 4: Overrides are highest in the *gendered* condition.

Results 3

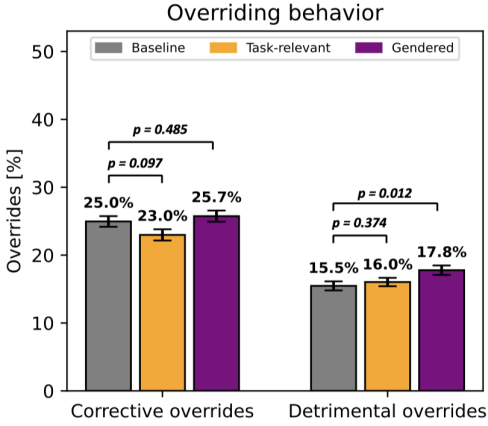


Figure 5: Explanations do not enable corrective vs. detrimental overrides.

Results 4

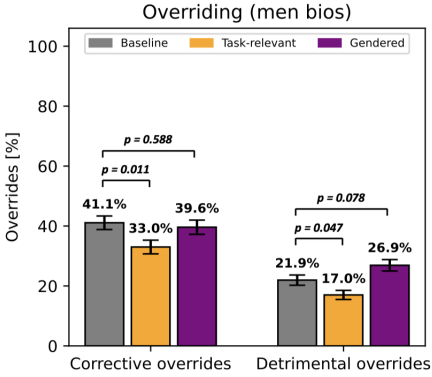


Figure 7: *Task-relevant* explanations decrease both corrective and detrimental overrides for men bios, compared to the baseline; whereas *gendered* explanations marginally increase detrimental overrides.

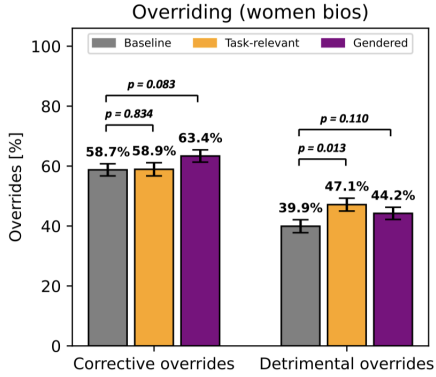


Figure 8: *Gendered* explanations marginally increase corrective overrides over the baseline for women bios; whereas *task-relevant* explanations increase detrimental overrides.

Explanations have been framed as an important mechanism for better and fairer human-AI decision-making.

Explanations have been framed as an important mechanism for better and fairer human-AI decision-making.

We find that the type of features that an explanation highlights matters: when explanations highlight only task-relevant words, people tend to reinforce stereotypical AI recommendations, ultimately increasing error rate disparities between women and men.

On the other hand, when explanations highlight gendered words, people tend to override more AI recommendations to counter stereotypical AI recommendations, which decreases error rate disparities.

On the other hand, when explanations highlight gendered words, people tend to override more AI recommendations to counter stereotypical AI recommendations, which decreases error rate disparities.

Importantly, these effects on distributive fairness do not involve an enhanced human ability to override incorrect AI recommendations (i.e., "appropriate reliance") but solely emerge from a shifting in error types.

On the other hand, when explanations highlight gendered words, people tend to override more AI recommendations to counter stereotypical AI recommendations, which decreases error rate disparities.

Importantly, these effects on distributive fairness do not involve an enhanced human ability to override incorrect AI recommendations (i.e., "appropriate reliance") but solely emerge from a shifting in error types.

For instance, if an AI system predicts that a woman is a teacher and the explanation highlights the use of gendered words, human decision-makers are more likely to override the recommendation regardless of whether the woman is indeed a teacher.

To design effective interventions for decision support, it is important to understand the psychological mechanisms at play when humans adhere to or override AI recommendations. One promising direction for follow-up work will be to study why the highlighting of gendered features results in AI aversion. On the other hand, we have also seen cases where humans perceive the use of gendered words for predicting occupations as fair, and it will be interesting to analyze when and why this is this case.

Thank you!