Fair Machine Guidance to Enhance Fair Decision Making

Mingzhe Yang

The University of Tokyo
mingzhe-yang@g.ecc.u-tokyo.ac.jp

Abstract

Human judgment is often subject to bias, leading to unfair decisions. This is particularly problematic when assessments have significant consequences, underscoring the importance of guiding humans towards fairness. Although recent advancements in AI have facilitated decision support, it is not always feasible to employ AI assistance in real-world scenarios. Therefore, this study focuses on developing and evaluating a method to guide humans in making fair judgments. Our experimental results confirmed that our approach effectively promotes fairness in human decision-making.

Introduction

In various situations, humans often evaluate others, such as in loan assessments and job interview processes. While these evaluations are expected to be fair, they often result in unfair judgments, posing a significant issue. This tendency towards unfairness stems from biases formed through experiences and beliefs acquired, often unconsciously, over one's life.

AI-assisted decision-making is extensively practiced, where human operators utilize AI outputs to enhance decision-making accuracy. Current AI-assisted decision-making can be likened to using AI as a "training wheel." However, it is not always feasible to implement AI-assisted decisions in real-world scenarios. Therefore, we propose a method to guide humans towards making fair judgments independently of AI assistance.

Related Work

In efforts to mitigate biases in human judgment, various methods have been explored thus far. Among these, the Implicit Association Test has been employed to make individuals aware of their biases (Greenwald, McGhee, and Schwartz 1998). Although this approach confirmed an inclination to eliminate biases consciously, it didn't change into fair decision. This gap was primarily because, while individuals could recognize the presence of biases, they lacked clear guidance on how to alter their judgments accordingly.

Methods

To guide humans towards fair judgment, fair machine guidance integrates two frameworks: Fair-awareness machine learning (Agarwal et al. 2018) and machine teaching (Liu et al. 2017). Fair-awareness machine learning emphasizes fairness in model outputs within the machine learning domain. Machine teaching focuses on creating educational materials to enhance human learning efficiency when acquiring concepts. In fair machine guidance, we first learn a human judgment model (the unfairness model) based on past human decisions. Then, using fair-awareness machine learning, we derive a modified model (the fairness model) that corrects for unfairness in the human judgment model. Subsequently, machine teaching is employed to simulate and generate materials that transition the unfairness model towards the fairness model. Figure 1 shows samples presented to humans under fair machine guidance (Ma, Correll, and Wittenbrink 2015). By visualizing and comparing the criteria of the fair model and the unfairness model, we aim to make individuals aware of their judgment biases and foster an understanding of fair judgment criteria.

Experiments

We conducted an experiment with 99 participants using our fair machine guidance system. The experimental tasks included income prediction and loan approval scenarios. Participants were presented with profiles for person assessment and asked to make fair evaluations. In this study, a fairness metric *unfairness* is defined as follow: **unfainress** $= P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)$ where $\hat{Y}$ represents the labels predicted by humans, while S denotes protected attributes such as race or gender. This metric requires that the same response trends are fair for certain attributes, e.g., race and gender. A lower value of unfairness, closer to 0, signifies greater fairness.

The experimental process is divided into three parts: a pre-test, an intervention, and a post-test. Initially, the pre-test measures participants' unfairness and judgment criteria before intervention. Next, during the intervention, educational materials based on fair machine guidance are presented to support the learning of fair judgment. Finally, the post-test evaluates improvements in fairness due to the intervention. It is important to note that in the post-test, AI-

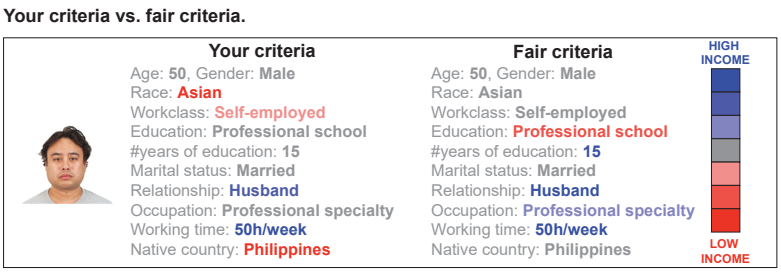

Figure 1: Example of teaching materials. Teaching materials provided by fair machine guidance to teach how to make fair decisions. Materials were selected by iterative machine teaching. The interpretations of the student and teacher models were presented visually.

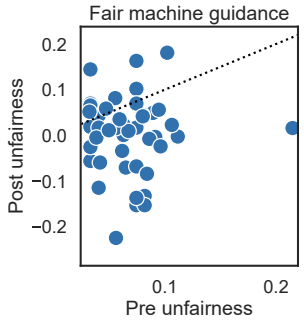

Figure 2: Scatter plots of unfairness of each participant in the pre- and post-tests. The dotted line represents equal levels of unfairness in both the pre- and post-tests; points below this line indicate an improvement in fairness in the post-test.

assisted decision-making is not employed; judgments are made solely by humans. An open-ended questionnaire was conducted with the participants at the end of both the pre- and post-test.

Results

Figure 2 shows the change in unfairness between the pre- and post-test. The results confirmed that the proposed method improved fairness for many participants. Figure 3 shows the rate of change in judgment criteria between the pre- and post-test. It was observed that many participants, through fair machine guidance, altered the attributes they focused on for judgment between the pre- and post-test.

Discussion

This experiment confirmed that our proposed method helped participants learn fair judgment and encouraged changes in their judgement criteria. The following discussion is based on participants' open-ended responses, focusing on the internal changes they experienced. Analysis of their open-ended responses revealed that many participants became aware of their own biases and tendencies to adhere to specific perspectives in their judgments. Realizing their unfair judgments led them to learn the importance of considering a broader range of perspectives and alternate viewpoints,

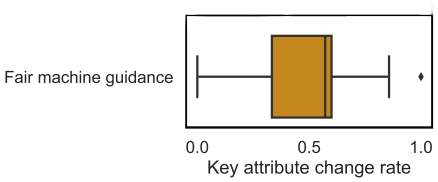

Figure 3: Degree of change in attributes of focus between pre- and post-tests indicates the extent of alteration in the attributes under consideration. Values closer to zero suggest a greater change in the attributes focused on between pre- and post-tests.

thereby shifting towards fairer judgment. This suggested that presenting and comparing fair and unfair judgment criteria using our proposed method was beneficial in prompting participants to reconsider their judgements.

Conclusion

In this study, we promoted changes in judgment criteria towards fairness by visualizing and presenting both fair judgment standards and human judgment criteria to participants. This approach proved to be effective in enabling humans to make fair decisions, even in scenarios without AI assistance.

References

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. In *Proceedings of the 2018 International conference on machine learning*, 60–69.

Greenwald, A. G.; McGhee, D. E.; and Schwartz, J. L. K. 1998. Measuring individual differences in implicit cognition: The implicit association test. *Journal of personality and social psychology*, 74(6): 1464.

Liu, W.; Dai, B.; Humayun, A.; Tay, C.; Yu, C.; Smith, L. B.; Rehg, J. M.; and Song, L. 2017. Iterative machine teaching. In *Proceedings of the 2017 International Conference on Machine Learning*, 2149–2158.

Ma, D. S.; Correll, J.; and Wittenbrink, B. 2015. The chicago face database: A free stimulus set of faces and norming data. *Behavior Research Methods*, 47(4): 1122–1135.