Toward Human-in-the-Loop AI fairness with Crowdsourcing: Effects of Crowdworkers' Characteristics and Fairness Metrics on AI Fairness Perception

Yuri Nakao
Fujitsu Limited
Kawasaki-city, Kanagawa-prefecture, Japan
nakao.yuri@fujitsu.com

Abstract

Fairness-aware machine learning technology has been developed to support people's fair decisions. To aggregate the opinion about the fairness in machine learning from diverse stakeholders, we need to use crowdsourcing methods. However, crowdworkers perception is easy to be affected by the setting and the situations of problems. In this research, We examine the difference in participants' perceptions using four scenarios that differ in the acceptance rates of the screening process for job candidates and the ways of calculating disparity.

Introduction

As machine learning technology is used in diverse decision-making, it has been pointed out that potentially reproduces historical discriminatory bias and various technologies have been developed to remove such bias from data and models (Mehrabi et al. 2021). These technologies have been used to attempt to mitigate bias based on protected attributes, such as gender, ethnicity, and age.

However, there has not been any numerical consensus on the fairness criteria (Srivastava, Heidari, and Krause 2019; Nakao et al. 2019). Hence, there is a need for crowdsourcing methods to aggregate diverse people's opinions to make the machine learning models fair in a human-in-the-loop way. On the other hand, human perceptions are easily changed even by small stimuli. In behavioral economics, for example, it has been pointed out that people's estimations of numbers such as a person's age at death or the heights of a tree are affected by the number shown right before the estimation (Kahneman 2011). Therefore, we need to examine how such subtle differences in shown information affect perceived fairness.

To investigate the effect of the minor differences in fairness information, we conduct a crowdsourcing study for the participants who reside in the US via Amazon Mechanical Turk (MTurk). We take a virtual case in which a company uses artificial intelligence (AI). From the responses, we explore what gender balance from the AI model the participants perceive as being fair. We address the following research questions:

RQ1. How are people's perceptions of the fair gender balance in AI affected by the difference in shown information?

RQ2. What human characteristics tend to be affected by the difference in shown information?

RQ3. In what types of shown information do differences in perceived fairness based on differences in human characteristics tend to be observed?

Methods

First, we give the background information scenarios explained to the participants. The summary is as follows:

> A company called X Ltd. decided to start to recruit new professional workers requiring certain qualifications (e.g., accountants, or in-house lawyers). X Ltd. is trying to select candidate applicants to be interviewed by using artificial intelligence (AI). However, because only a few cases where female applicants were hired were included in the training data, the AI tends to mark female candidates with lower scores than male candidates. You, an HR representative, have to decide on a fair gender balance to make the AI model provide fair results.

Next, The participants are randomly assigned to one of the following scenarios to investigate the differences in participants' perceived fair female-to-male ratio (fair gender ratio) among the scenarios:

Scenario H-R (*H*igh Acceptance rate & *R*atio Metric): Out of 100 candidates including 50 males and 50 females, 50 pass the screening process (i.e., the acceptance rate is 50%). The ratio of acceptance rates (i.e., disparate impact) is used to evaluate fairness.

Scenario H-D (*H*igh Acceptance rate & *D*ifference Metric): Out of 100 candidates including 50 males and 50 females, 50 pass the screening process (i.e., the acceptance rate is 50%). The difference in acceptance rates (i.e., statistical parity difference) is used to evaluate fairness.

Scenario L-R (*L*ow Acceptance rate & *R*atio Metric): Out of 2500 candidates including 1250 males and 1250 females, 50 pass the screening process (i.e., the acceptance rate is 2%). The ratio of acceptance rates is used to evaluate fairness.

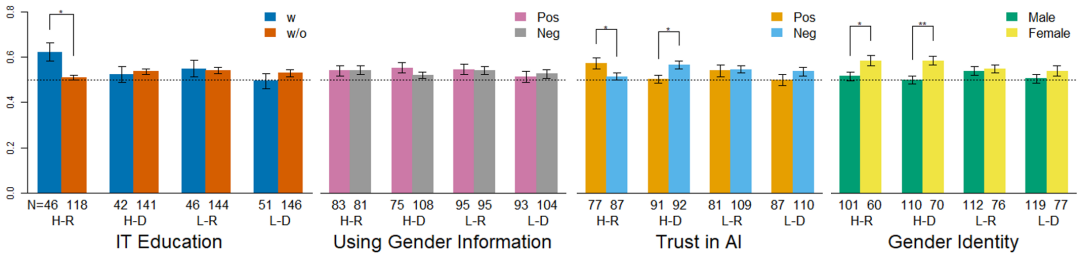

Figure 1: The perceived fair gender ratio for each characteristics in each scenario. Numbers below characteristics are that of participants belonging to each group. Error bars indicate standard errors, * indicates $p < 0.05$, and ** indicates $p < 0.005$.

Scenario L-D (*L*ow Acceptance rate & *D*ifference Metric): Out of 2500 candidates including 1250 males and 1250 females, 50 pass the screening process (i.e., the acceptance rate is 2%). The difference in acceptance rates is used to evaluate fairness.

Binary Search For the participants to express their perceived fair gender ratios, we use the binary search. The participants were first shown the original result from the AI model, where 33 male and 17 female candidates passed the screening process. The participants were then given the first question asking if they thought it is better to have more male or female interviewees.

Reflecting the answer, the number of female interviewees was increased to 37 (more female interviewees) or decreased to 12 (more male interviewees). After that, the participants were asked three times whether it was preferable to increase the number of female or male interviewees. Here, the unit of change decreased to 6 or 7 (in the second question), 3 (in the third question), and 1 or 2 (in the fourth question). Through this process, the participants expressed their perceived fair gender ratios at intervals of 6.25 % (3 or 4 people) of all interviewees.

Measures We compare the ratios of female interviewees expressed by participants among different scenarios, and characteristics to assess the difference in the perceived fair gender balance. For the characteristics of the participants to evaluate, we ask them if the participants feel it is good or bad to consider gender information in the process of recruitment (i.e., Gender Info Use - Pos., and Neg.), if the participants think it is preferable for HR representatives to trust AI when deciding who to hire (i.e., Trust in AI - Pos, and Neg.), if they received IT education and those who did not (i.e., IT Education - w, and w/o), and their gender identity.

Participants We recruited participants on MTurk. The conditions for the participants are residing in the US, being at least 18 years old, having completed at least 500 Human Intelligence Tasks (HITs, MTurk's task unit) which was approved, and having at least a 95% HIT approval rate. The median time to complete the task was 18 minutes and 15 seconds. Participants were compensated $2.5 for their time. We omitted the participants who completed the survey multiple times, did not pass the attention check and declared that they did not understand the meaning of fairness metrics. After this screening process, we obtained 734 responses.

Results

From the Kruskal-Wallis test conducted to compare the perceived fair gender ratios among all scenarios to answer RQ1, there was no significant difference ($H = 0.390$, $p = 0.94$).

Regarding RQ2 we initially conducted the Mann-Whitney U test among the participants with different characteristics within each attribute regardless of scenarios. There was a significant difference only in gender identity (male: $N = 442$ $M = 0.514$ $SD = 0.20$, female: $N = 283$ $M = 0.562$ $SD = 0.17$, $U = 53314.5$, $p < 0.001$).

With the Mann-Whitney U test to compare the perceived fair gender ratio between the different characteristic in each attribute within each scenario to address RQ3, we observed five significant differences (Fig. 1). In scenario H-R, the participants who had IT education tended to perceive a higher gender ratio as fair than those who did not have IT education ($U = 1984.0$, $p = 0.006$). In scenario H-R, participants with positive attitudes toward trust in AI tended to perceive a higher gender ratio as fair than those with negative attitudes ($U = 2761.0$, $p = 0.045$). Conversely, in scenario H-D, the participants with negative attitudes toward trusting in AI perceive a higher gender ratio as fair than those with positive attitudes ($U = 4892.5$, $p = 0.040$). Finally, female participants perceived a higher gender ratio as fair than male participants in both scenarios H-R ($U = 2429.5$, $p = 0.030$) and H-D ($U = 2862.0$, $p = 0.002$).

Discussions and Future works

From these results, for RQ1, the difference in the problem settings alone did not generally affect the participants. And, for RQ3, in problem settings with a high acceptance rate, the difference in perceived fair gender ratios between different characteristics is more significant. Although in this experiment, there were a few significant differences, there are some points to improve to explore the effects of shown information to crowdworkers. For example, there might be other more flexible ways than the binary search method for the participants to express their perceived fair gender balance considering fairness metrics enough.

Finally, so far, despite the need for the sophistication of the crowdsourcing method in the context of AI fairness, few studies focus on the effect of fairness metrics on crowdworkers' perceptions. Toward the future where AI is embedded in diverse decision makings, we need to proceed with this line of research.

References

Kahneman, D. 2011. *Thinking, fast and slow*. Macmillan.

Mehrabi, N.; Morstatter, F.; Saxena, N.; Lerman, K.; and Galstyan, A. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6).

Nakao, Y.; Shigezumi, J.; Yokono, H.; and Takagi, T. 2019. Requirements for Explainable Smart Systems in the Enterprises from Users and Society Based on FAT. In *IUI Workshops*.

Srivastava, M.; Heidari, H.; and Krause, A. 2019. Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, 2459–2468. New York, NY, USA: Association for Computing Machinery. ISBN 9781450362016.