

Trust in Fair Algorithms: Pilot Experiment

Mattia Cerrato^{1,†}, Marius Köppel^{2,†}, Kiara Stempel^{1,†}, Alesia Vallenias Coronel^{1,†} and Niklas Witzig^{1,*,†}

¹Johannes Gutenberg-University Mainz, Saarstraße 21, 55122 Mainz DE

²ETH Zürich (IPA), Otto-Stern-Weg 5, 8093 Zürich CH

Abstract

We study human reliance on (un)fair algorithmic recommendations. We train two separate models to predict the employment status of (anonymized) persons in the US-Census 2018. The two models differ in their *fairness*, with their overall accuracy being the same for (differing by) binary¹ gender ("overall accuracy equality") [2]. We use the predictions by the two models in a pilot experiment, in which participants performed the same prediction task while receiving assistance from either the fair or unfair model. We find that people rely more on the predictions by the fair algorithm with unknown gender, yet stronger on unfair models once the gender is revealed. The present data remains limited in size, but we aim to identify these effects and potential sources of individual treatment heterogeneity using a much larger experimental sample in the future.

Keywords

Algorithmic Fairness, Trust, User Study

1. Introduction

The fairness of algorithms is a topic that garnered a lot of attention in recent years, with biased systems producing unwanted consequences for marginalized groups and often perpetuating societal inequalities. At the same time, human-centric AI – the active involvement of users in the use and deployment of AI systems – is increasing. However, only little is known about how human decision-makers, imagine e.g., in managerial positions, interact and use predictions by (un)fair models. In this project, we study the impact of the *fairness* of a prediction model on human reliance in a setting where there are known and existing biases (both from machine and humans): employment, a high-risk category according to the EU AI-Act (Annex III). Specifically, we test if humans rely more strongly on predictions by *fair* versus *unfair* models when tasked to predict the employment status of (anonymized) individuals in a preliminary online experiment. We find tentative evidence that reliance on fair models is slightly, yet significantly, higher with unknown sensitive attributes yet *lower* once the sensitive attribute of the anonymized person is revealed. We investigate treatment heterogeneity and find provisional (yet statistically

¹We fully acknowledge that this dichotomy is far from ideal (see also the discussions in Pinney et al. [1]), but the data availability does not allow the use of non-binary gender attributes so far.

EWAF'24: European Workshop on Algorithmic Fairness, July 01–03, 2024, Mainz, Germany


*Corresponding author.

†These authors contributed equally.

✉ mcerrato@uni-mainz.de (M. Cerrato); mkoeppl@phys.ethz.ch (M. Köppel); stempel@uni-mainz.de (K. Stempel); a.vallenias@uni-mainz.de (A. Vallenias Coronel); niklas.witzig@uni-mainz.de (N. Witzig)



© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

insignificant) evidence that prior beliefs, i.e., the general and potentially biased expectation about the employment status of men and women are a potential source of treatment heterogeneity.

2. (Un-)fairness in Predicting Employment Status

The main goal is to deploy two models with differing fairness levels, measured by accuracy, to offer predictions to participants. Both models are trained on the same 2018 U.S. Census data, predicting employment status using variables like age, education, disability, marital status, and other sociodemographic factors. We exclude the sensitive attribute, gender, from the model’s input. We use BinaryMI [?] for both models, which aims to make a fair prediction by learning a representation of the data that is invariant to the sensitive attribute, by using a stochastically quantized binary neural layer. All parameters for training are set to the same values for both models, except for γ , which is the tradeoff parameter in the loss function weighting the importance of representation invariance and accuracy on the target variable, i.e., the employment status. For the fair model, we set γ to 0.05, whereas for the unfair one, we set it to 0. Both models are trained for 50 epochs. Before model training, the dataset is subsampled to retain only 10% of the instances of employed women, reducing their representation in the learning process. While the fair model reaches an accuracy of 79.37% for male and 79.18% for female, the unfair version reaches an accuracy of 81.98% for male, but only 79.24% for female. In the experiment, we select 10 random profiles from the validation dataset, ensuring a balance of gender and employment status, and representing two profiles each from the prediction quintiles of the fair, unfair models, and their difference.

3. Online Experiment

The experiment, centering on a prediction task and supplemented by auxiliary measures, was conducted on Prolific (www.prolific.com) in December 2023 with 103 participants. It lasted an average of 27.23 minutes, with 77% of participants receiving a 2-pound bonus. **Main Task:** For participants in the online experiment, the main task is to give an accurate probabilistic statement if a given (anonymized) person is employed or not based on shown personal characteristics. The statement ranges from 0 to 100 and the closer the statement is to the true employment status (either 0 or 1 (=100)), the higher the chances that a participant obtains a bonus prize. We apply the randomized quadratic scoring rule [3], which is proper even for risk-averse subjects and incentivizes participants to report their beliefs truthfully. The main task consists of **three phases**, which entail 10 statements each (s.t. participants give 30 statements in total). We randomly draw one statement to determine participants’ payoffs. Panels (a) - (c) of Figure 1 show screenshots of the three phases. In **phase 1** of the main task, participants obtain information about six personal characteristics of a given anonymized person, i.e., about their age, education level, marital status, whether they have a disability, if they moved in the previous year, and if they have single or multiple ancestries. Information about the gender of the person remains undisclosed. Based on this information, participants need to form and enter a belief about the employment status of the person. In **phase 2**, participants receive the prediction by either the fair or unfair model, which is randomly determined at the beginning of the

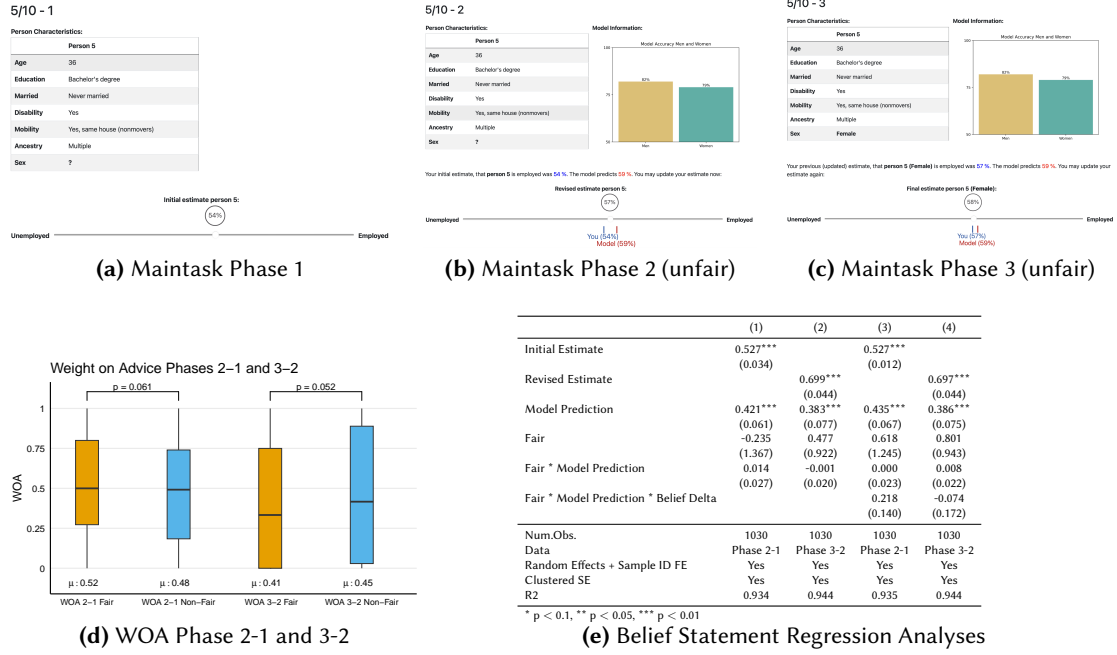


Figure 1: (a) Participants enter their probabilistic belief based only on the characteristics of the anonymized person. (b) Participants obtain the prediction by the model and may update their initial estimate and get a reminder of group (un)fairness. (c) The sex is revealed and participants may update a final time. In the fair group, reminders are replaced accordingly. (d) Average WOA for Phases 2-1 and 3-2, separately for Fair and unfair. (e) Intercept and remaining interactions remain omitted.

experiment (and remains constant throughout, "between-subject" treatment). In addition, they are reminded of the (un)fairness by the global accuracy plot next to the personal characteristics. Participants now may update their belief estimates. Phases 1 and 2 are carried out sequentially for each anonymized person, s.t. participants first enter their initial estimate and then may revise their estimate. After phases 1 and 2 (for each of the 10 profiles), they are introduced to **phase 3**, which discloses the gender of the profile and participants may revise their (possibly updated) estimate a final time. This task allows us to construct the central measure of **trust** on algorithmic predictions, namely the weight of advice (WOA). This measure stems from the advice-taking literature and we use the censored measure by Greiner et al. [4]: $WOA = \min\left(\max\left(0, \frac{|\text{updated estimate} - \text{initial estimate}|}{|\text{model prediction} - \text{initial estimate}|}\right), 1\right)$. This measure is 0 if participants — in their updated estimate — ignore the model prediction, and 1 if they react as much (or more strongly) as suggested by the model prediction. Crucially, we estimate the WOA separately for phases 2 and 1, i.e., based on the first update with a model estimate, and for phases 3 and 2, the final estimate once the profiles' gender was revealed. **Additional tasks and measures:** We carry out several additional tasks and surveys. We gather data on prior beliefs about the employment status of men and women, a survey with attitudes towards the statistical model by Zhou et al. [5], and a general trust in various institutions survey [6]. We use this data to investigate heterogeneous treatment effects, i.e., if our chosen treatment variation invoked systematically

different behavior across participants depending on these characteristics. For brevity, we focus on the role of prior beliefs in the later analyses.

4. Results

Main Treatment Differences: Panel (d) in Figure 1 plots the average WOA for the two groups. On average, the WOA is 0.48 in the unfair group and 0.52 in the fair group. This difference is statistically significant on the 10% level ($p = 0.061$) in a two-sided t-test. Participants showed a stronger reaction to predictions from the fair model than those from the unfair model. Despite the small difference, it's significant considering the low "treatment intensity" between the models, making a 4 percentage point difference noteworthy. Contrary, the mean WOA values between phases 3 and 2 (i.e., once the gender of the candidate has been revealed) show a *lower* reliance on the predictions by the fair model (panel (d)). On average, the WOA on the fair model is 0.41, yet 0.45 on the unfair model ($p = 0.052$). This suggests that participants react more strongly to fair predictions if the sensitive attribute remains undisclosed but *more strongly* to unfair predictions once the sensitive attribute is known. **Regressions and Heterogeneous Treatment Effects:** The Table in panel (e) in Figure 1 regresses the (final) revised belief statement on the initial (revised) statement, the model predictions from the main task between phase 2 and 1 (3 and 2) and a dummy for the treatment group. Columns 1 and 3 include data from phases 2-1, whereas column 2 and 4 from phases 3-2. Focussing on the first two columns, we observe a stronger reliance on the participant's own initial (revised) estimates compared to the model predictions, while this asymmetry is much more pronounced between phases 3 and 2. In line with the evidence from panel (d), the *interaction effect* of the fair treatment with the model prediction has a positive (negative) coefficient, implying that participants in the fair treatment rely more (less) strongly on the model predictions during phases 2-1 (3-2). The interaction effects were not statistically significant, likely due to the low power of the pilot experiment. The analysis includes a treatment heterogeneity segment, examining an interaction between fair treatment, model predictions, and "Belief Delta" – a measure of the difference in participants guesses regarding the employment rates of men versus women. The coefficient of this interaction is positive (negative) for phases 2-1 (3-2), implying that the more "biased" a person the stronger (weaker) their reliance on the fair predictions. Again, these interactions are, however statistically insignificant.

5. Discussion and Next Steps

The key finding of this study is that participants in our pilot experiment reacted more strongly to the predictions by a fair model compared to an unfair one in the presence of unknown sensitive attributes, yet less once the sensitive attribute was revealed. This novel insight indicates that the effectiveness of fair models depends on the decision-makers using them. We recognize the limitations of our pilot study, notably its low power, leading to provisional conclusions. Methodologically, we didn't consider the statistically optimal posterior belief (assuming Bayesian participants), potentially explaining some treatment differences. We plan to overcome these limitations with a larger online experiment and by theoretically modeling task behavior.

References

- [1] C. Pinney, A. Raj, A. Hanna, M. D. Ekstrand, Much ado about gender: Current practices and future recommendations for appropriate gender-aware information access, in: Proceedings of the 2023 Conference on Human Information Interaction and Retrieval, CHIIR '23, ACM, 2023. URL: <http://dx.doi.org/10.1145/3576840.3578316>. doi:10.1145/3576840.3578316.
- [2] S. Verma, J. Rubin, Fairness definitions explained, in: Proceedings of the International Workshop on Software Fairness, ICSE '18, ACM, 2018. URL: <http://dx.doi.org/10.1145/3194770.3194776>. doi:10.1145/3194770.3194776.
- [3] T. Hossain, R. Okui, The binarized scoring rule, *The Review of Economic Studies* 80 (2013) 984–1001. URL: <http://dx.doi.org/10.1093/restud/rdt006>. doi:10.1093/restud/rdt006.
- [4] B. Greiner, P. Grünwald, T. Lindner, G. Lintner, M. Wiernsperger, Incentives, Framing, and Reliance on Algorithmic Advice: An Experimental Study, Working Paper 01/2024, WU Vienna University of Economics and Business, 2024. doi:10.57938/137467e7-9d47-4a6c-87dc-673ff4f58009.
- [5] J. Zhou, S. Verma, M. Mittal, F. Chen, Understanding relations between perception of fairness and trust in algorithmic decision making (2021).
- [6] Oecd, OECD guidelines on measuring trust, Organization for Economic Co-operation and Development (OECD), Paris Cedex, France, 2017.