A Framework for Predicting Fairness Perception – Towards Semi-Personalized Explanations

Avital Shulner-Tal [1], Doron Kliger [2] and Tsvi Kuflik [1]

[1] *Information system department, University of Haifa, Haifa, Israel*
[2] *Economics department, University of Haifa, Haifa, Israel*

Abstract

Systems based on Machine Learning (ML) and Artificial Intelligence (AI) are becoming part of our everyday lives. Potential risks and discrimination in algorithmic decision-making systems (ADMSs) has recently led to a growing interest among developers, practitioners, regulators, as well as in the research community, in ensuring the fairness of this kind of systems. The fairness of ADMS is usually examined using a variety of objective metrics. However, while such examination may ensure their computational fairness, in order to convince users to trust these systems there is a need to address the users' fairness perception as well, as it plays an important role in accepting novel technologies and especially ADMSs that are considered to be "Black boxes". Users' perception of fairness towards ADMSs may be affected by various personal characteristics, as well as by the explanations (whether for a model or for an outcome) provided by the system. Hence, following an experiment that revealed inter-personal differences, we propose a framework for prediction of individuals' fairness perception that may be used by developers for generating personalized explanations that will match their users' preferences. The proposed framework is based on three main aspects, including: system/scenario characteristics, user's demographic characteristics and user's personality characteristics. Our experimentation demonstrated the potential benefit of explanation's personalization in enhancing users' fairness perceptions. As far as we know, this is the most comprehensive framework for predicting fairness perceptions and selecting the most beneficial explanation style which is based on individual's personal characteristics.

Keywords

Algorithmic Fairness, Perceived Fairness, Algorithmic Transparency, Explainability, Algorithmic Systems, Decision Support Systems, Users Perception

1. Introduction

Recent research in the area of algorithmic fairness focusses mainly on ensuring the fairness of algorithmic decision-making systems (ADMSs) using a variety of objective computational fairness metrics, while there is a lack of studies about users' fairness perception and its impact on users' decisions to use a system and to trust its results [3, 18]. Understanding users' perceptions regarding the fairness of ADMSs is essential in order to deal with trust-related problems [1, 6, 7, 11, 12, 16, 17, 19, 20, 21]. As users differ in their characteristics and preferences, it is well known that the "one size fits all" paradigm is inappropriate when a system aims at providing a service to its users and personalization is a must [13]. Explanations about the decision-making process and the outcome of the system may enhance users' fairness perception [1, 3, 15, 17, 21]. The various factors that affect users' fairness perceptions and how they can be modified should be examined and personalization of explanations may be applied given the result of such examination. This aspect was dealt by relatively few recent studies, which indicated the

Proceedings Name, Month XX–XX, YYYY, City, Country
EMAIL: email1@mail.com (A. 1); email2@mail.com (A. 2); email3@mail.com (A. 3)
ORCID: XXXX-XXXX-XXXX-XXXX (A. 1); XXXX-XXXX-XXXX-XXXX (A. 2); XXXX-XXXX-XXXX-XXXX (A. 3)


importance of examining the factors that affect users' fairness perceptions [5, 15, 16] and the use of personalized explanations in AI/ML – based systems [5, 8, 14]. Currently, there is a lack of studies that examines the personalization of explanation based on explicit data that is collected from the users [14]. Our intention in this study is to fill this gap by suggesting an initial framework that enables predicting users' fairness perception based on her characteristics with respect to the specific scenario at hand and to the explanations that may be provided by the system.

2. Framework for predicting fairness perception

Shulner-Tal et al. [15] devised a between-subject online experiment that examined various factors and their effect on users' fairness perception regarding algorithmic decision-making (see [15] for detailed procedure and results). Their experiment was performed using a description of a recruitment decision support system as a case study.

The experiment consisted of six consecutive steps, in which the participants: (1) were provided with information about the system; (2) received input information (candidate description); (3) viewed the simulated output, alongside its explanation; (4) were requested to express their views about the fairness of the system; (5) filled a demographic questionnaire (gender, age, residence, education and income); and (6) completed a TIPI questionnaire for examining the Big Five personality domains (openness, conscientiousness, extraversion, agreeableness and emotional stability).

Steps (1), (2) and (3) created 72 different permutations to which the participants were randomly-assigned. The study followed [2, 4, 5, 8, 9, 10] in the selection of the demographic and personality characteristics. 3,196 randomly-assigned participants were recruited for the study using Amazon Mechanical Turk (see Appendix 1 for the various permutations and their descriptions).

We used the dataset[2] , collected in Shulner-Tal et al. [15], in order to classify the 3,196 participants according to the various factors. Table 1 presents the classes of the factors, as well as the target variable – participants' fairness perception regarding the permutation they received in the experiment.

Table 1
Characteristics' Classification Values

Factor	Characteristic	Values
System	input	High-quality/ Low-quality
	output	Positive/ Borderline/ Negative
	Input- Output Correlation	Compatible/ Contrasting
	Explanation	no explanation/ Case-based/ Certification-based/ Demographic-based/ Input influence-based/ Sensitivity-based
	Certification	Uncertificated / Certificated
Demographic	Gender	Male/ Female
	Age	18-34 / 35-50/ 50+
	Residence	USA/ India/ Other
	Education	High school degree or less / Bachelor's degree / Master's or doctoral degree
	Income	Above average / Average / Below average
Personality	Openness	High / Low
	Conscientiousness	High / Low
	Extraversion	High / Low
	Agreeableness	High / Low
	Emotional Stability	High / Low
Target variable	Fairness Perception	Extremely fair/ Moderately fair/ Slightly fair/ Slightly unfair/ Moderately unfair/ Extremely unfair

[2] https://doi.org/10.5281/zenodo.5075110.

See appendix 1 for the description of the characteristics. We referred to each characteristics' possible value, presented in Table 1, as binary variable (1- if the value is true, 0- else), when the sum of the possible variables for each characteristic amounts to 1. Then, we classified each participant according to her reported demographic characteristic, personality characteristics and according to the permutation (system characteristics) she received in the experiment.

The framework, presented in Figure 1, enables to predict users' views regarding the fairness of the system on a 6-point Likert scale, from "Extremely fair" to "Extremely unfair". The various characteristics that were examined in the experiment (system characteristics in blue, user's personality characteristics in purple and user's demographic characteristics in orange) and their effect on the fairness perception (ordinal regression coefficient (STD) and significance) are presented in the framework.

The implementation of the framework will include the following steps: (1) the user fills a demographic and personality questionnaire in order to determine the values of the demographic and personality characteristics, this step may be implemented once for each user or in any periodically time; (2) the system provides its classification (values) regarding the input, output, input-output correlation and certification of the system; (3) the system aggregate the true values (values that are determined as 1) with the ordinal regression coefficients and predicts the fairness perception level for each explanation style; (4) the most beneficial explanation style, in which the result of the prediction is the highest level of fairness perception, is presented.

We referred to the reported fairness perception of the participants, participated in the experiment, as the "gold standard" levels of fairness perception. To evaluate the framework, we preformed 10-fold cross validation using various models. The models predict the fairness perception level for each explanation style, according to the characteristics' values and compare it with the "gold standard" level for each of possible values. Table 2 present the evaluation results for each model (AUC, Accuracy, F1, precision and recall) and Figure 2 presents the ROC Analysis Curve of the various models. See Appendix 1 for models' confusion matrixes.

The results of the evaluation suggest that using AdaBoost or Neural Network models can predict the fairness perception of the users with high accuracy (0.938 and 0.877 respectively).

3. Discussion, conclusions and future work

While most of the characteristics (the input, output, input-output correlation, certification of the system, demographic and personality characteristics of the users) may be difficult or even impossible to change, the explanation provided by the system can be relatively easily modified in order to increase the fairness perception of the users. Hence, using this framework will enable to present semi-personalized explanations which are based on the input, output, input-output correlation and certification of the system, as well as the demographic and personality characteristics of the users.

The implementation of this framework requires performing demographic and personality questionnaires for the users and implantation of some kind of explanation generator that can create case-based, demographic-based, input influence-based and sensitivity-based explanation styles, as well as, performing an auditing process in order to be able to create some certification/certification-based explanation to the system.

The framework can be easily extended with more characteristics as well as more explanation styles and can be generalized and evaluated for various domains.

4. Acknowledgements

Partial financial support was received from the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105 (CyCAT – Call: H2020-WIDESPREAD-05-2017-Twinning), by a scholarship program for doctoral students in High-Tech professions at the University of Haifa, Israel and by Data Science Research Center (DSRC) at the University of Haifa, Israel.

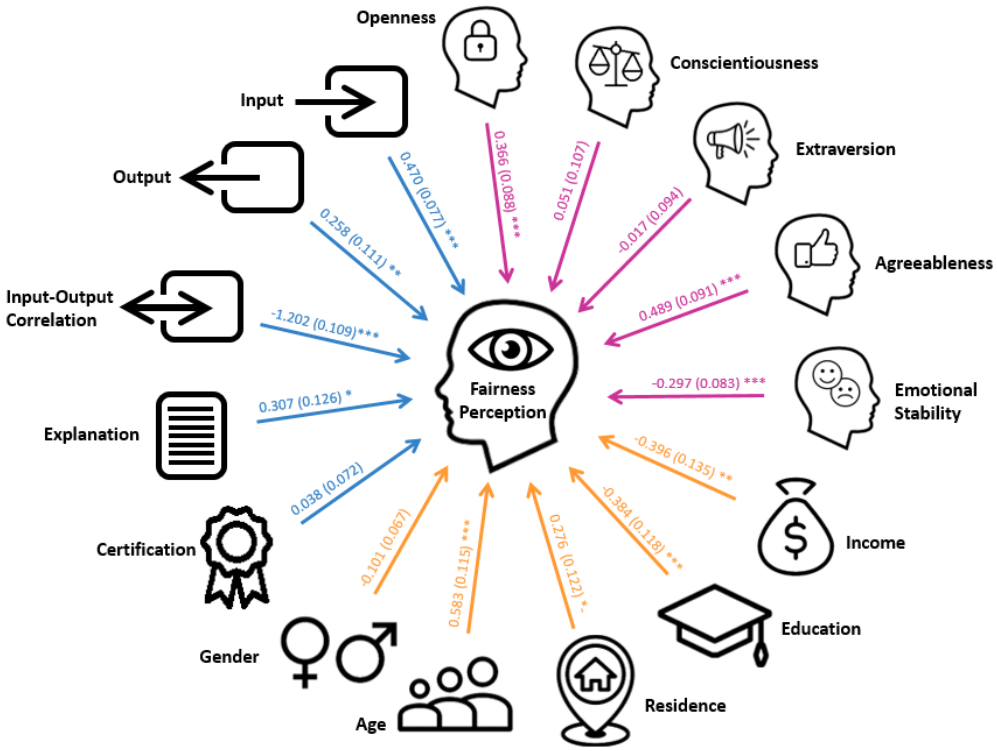

Figure 1: Fairness Perception Prediction Framework

Table 2

Framework Evaluation Results

Model	AUC	Accuracy	F1	Precision	Recall
Naïve Bayes	0.636	0.363	0.318	0.332	0.363
SVM	0.719	0.367	0.360	0.386	0.367
Logistic regression	0.648	0.396	0.295	0.400	0.396
KNN	0.709	0.415	0.339	0.417	0.415
Decision Tree	0.943	0.680	0.673	0.691	0.680
Random Forest	0.979	0.842	0.841	0.849	0.842
Neural Network	0.984	0.877	0.876	0.880	0.877
AdaBoost	0.998	0.938	0.938	0.938	0.938

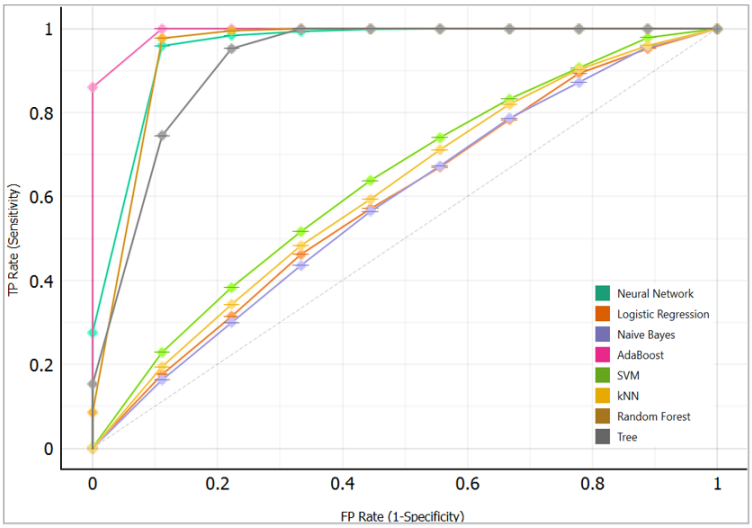

Figure 2: ROC Analysis Curve

5. References

[1] AA Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. In Human and Machine Learning (pp. 21-35). Springer, Cham.

[2] OO Ahnert, G., Smirnov, I., Lemmerich, F., Wagner, C., & Strohmaier, M. (2021, June). The FairCeptron: A Framework for Measuring Human Perceptions of Algorithmic Fairness. In Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (pp. 401-403).

[3] BB Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

[4] PP Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... & Zhang, Y. (2019). One explanation does not fit all: A toolkit and taxonomy of ai explainability techniques. arXiv preprint arXiv:1909.03012.

[5] QQ Conati, C., Barral, O., Putnam, V., & Rieger, L. (2021). Toward personalized XAI: A case study in intelligent tutoring systems. Artificial Intelligence, 298, 103503.

[6] CC Došilović, F. K., Brčić, M., & Hlupić, N. (2018, May). Explainable artificial intelligence: A survey. In 2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO) (pp. 0210-0215). IEEE.

[7] DD Ferrario, A., Loi, M., & Viganò, E. (2020). In AI we trust Incrementally: a Multilayer model of trust to analyze Human-Artificial intelligence interactions. Philosophy & Technology, 33(3), 523-539.

[8] RR Jung, A., & Nardelli, P. H. (2020). An information-theoretic approach to personalized explainable machine learning. IEEE Signal Processing Letters, 27, 825-829.

[9] SS Kouki, P., Schaffer, J., Pujara, J., O'Donovan, J., & Getoor, L. (2020). Generating and Understanding Personalized Explanations in Hybrid Recommender Systems. ACM Transactions on Interactive Intelligent Systems (TiiS), 10(4), 1-40.

[10] TT Millecamp, M., Htun, N. N., Conati, C., & Verbert, K. (2020, July). What's in a User? Towards Personalising Transparency for Music Recommender Interfaces. In Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization (pp. 173-182).

[11] EE Rai, A. (2020). Explainable AI: From black box to glass box. Journal of the Academy of Marketing Science, 48(1), 137-141.

[12] FF Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). Why should i trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (pp. 1135-1144). ACM.

[13] GG Rich, E. (1983). Users are individuals: individualizing user models. International journal of man-machine studies, 18(3), 199-214.

[14] UU Schneider, J., & Handali, J. (2019). Personalized explanation in machine learning: A conceptualization. arXiv preprint arXiv:1901.00770.

[15] HH Shulner-Tal, A., Kuflik, T., Kliger, D. Enhancing Fairness Perception – Towards Human-centred AI and Personalized Explanations. In review. International Journal of Human-Computer Interaction (HCI).

[16] II Shulner-Tal, A., Kuflik, T., Kliger, D. Fairness, Explainability and In-Between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. Accepted. Journal of Ethics and Information Technology (EIT). ETIN-D-21-00053R1

[17] JJ Singh, C., Murdoch, W. J., & Yu, B. (2018). Hierarchical interpretations for neural network predictions. arXiv preprint arXiv:1806.05337.

[18] KK Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2021). Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. arXiv preprint arXiv:2103.12016

[19] LL Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time

inspection of autonomous robots. Connection Science, 29(3), 230-241.
[20] MM Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. Journal of Management Information Systems, 23(4), 217-246.
[21] NN Wortham, R. H., Theodorou, A., & Bryson, J. J. (2016, June). What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In Ijcai-2016 ethics for artificial intelligence workshop

6. Appendix

Table 3
Manipulations used in the experiment [15]

Manipulation	Values	Description shown to the participants
Certification of the system	Uncertificated system	-
	Certificated system	
Input	High-quality candidate data	The candidate is an average graduate student (ranked 48th out of 103 students in the class). The candidate worked and did voluntary service while studying. The candidate was appreciated by co-workers in both places. Interviewer's summary: The candidate has relevant professional experience for the position. According to recommendation letters from former employers, the candidate fulfills his/her job responsibilities as required. According to the internal interview, the candidate has good communication skills. We may consider proceeding with this candidate.
	Low-quality candidate data	The candidate is significantly below the average graduate student (ranked 88th out of 103 students in the class). The candidate worked and did voluntary service while studying. The candidate was not appreciated by co-workers in both places. Interviewer's summary: The candidate has some relevant professional experience for the position. According to recommendation letters from former employers, the candidate sometimes may not fulfill job responsibilities as required. According to the internal interview, the candidate has reasonable communication skills. I am doubtful whether we should proceed with this candidate.
Output	Positive outcome	Recommended by the system (R)
	Borderline outcome	Borderline (B)
	Negative outcome	Not recommended by the system (N)
Explanation	Control- no explanation	-
	Case-based	A similar case (which received the same outcome) is the following candidate: "The candidate was an average performing student with some relevant experience for the job, S/he was positively recommended by her/his co-workers and fulfills her/his job responsibilities as required. The above candidate has a similar CV to this candidate and the demographic and personality characteristics were also similar."
	Certification-based	The system was tested and verified by authorized experts and regulators for fairness towards different population segments guarding against biases and discrimination. It was found to satisfy the required fairness constraints.

Demographic-based	The outputs are distributed in a normal distribution. Furthermore, it is known that: > 17% of candidates who are ranked in the top 10% in their graduating class are positively recommended by the system. > 36% of candidates with 10 years of relevant experience are negatively recommended by the system. > 28% of candidates with good communication skills in the internal interview are negatively recommended by the system. > 41% of candidates who were appreciated by former employers are negatively recommended by the system.
Input influence-based	Our predictive model assessed the candidate's information in order to predict his/her chances of progressing in the recruitment process. The more + signs or - signs, the more positively or negatively that factor impacted the probability of being recommended. Unimportant factors are not indicated. The following features and their impact on the outcome for this particular candidate are: > Rating of the university (+ +) > Candidate's ranking in the university (+) > Candidate's CV (+) > Candidate's personality test results (-) > Candidate's experience (+ + +) > Candidate's recommendation letters (- -) > Internal interviewer's recommendation (+ +)
Sensitivity-based	Our predictive model The following changes in the input features will change the outcome of the system: > If this candidate were to be ranked in the top 10 percent of her/his graduating class – the likelihood of positive recommendation by the system would be increased by 23%. > If this candidate had another year of relevant experience to this job – the likelihood of positive recommendation by the system would be increased by 34%. > If this candidate had better communication skills in the internal interview – the likelihood of positive recommendation by the system would be increased by 15%. > 12% of candidates who were recommended by the internal interviewer are positively recommended by the system.