




Does a Compromise on Fairness Exist in Using AI Models?

Jianlong Zhou¹✉, Zhidong Li¹, Chun Xiao², and Fang Chen¹

¹ Data Science Institute, University of Technology Sydney, Sydney, Australia
{jianlong.zhou,zhidong.li,fang.chen}@uts.edu.au

² Research Office, University of Technology Sydney, Sydney, Australia
chun.xiao@uts.edu.au

Abstract. Artificial Intelligence (AI) has been increasingly used to assist decision making in different domains. Multiple parties are usually affected by decisions in decision making, e.g. decision-maker and people affected by decisions. While various parties of users may have different responses to decisions regarding ethical concerns such as fairness, it is important to understand whether a compromise on fairness exists in using AI models. This paper takes AI-assisted talent shortlisting as a case study and investigates perception of fairness, trust, and satisfaction with decisions of both recruiters and applicants in AI-informed decision making. The compromises on fairness between decision-maker and people affected by decisions are identified which are then explained by social and psychological theories. The findings can be used to help find compromising points between decision-maker and people affected by decisions so that both parties can reach for a balanced state in decision making.

Keywords: AI ethics · Fairness · Trust · Satisfaction · Compromise

1 Introduction

Artificial Intelligence (AI) and Machine Learning (ML) algorithms have been increasingly used in shaping our everyday lives and activities in different domains especially human related decision making such as allocation of social benefits, hiring, and criminal justice [4, 8, 11]. As a result, the ethical issues of AI are becoming key concerns in algorithmic decision making. AI algorithms, trained on a large amount of historical data, may not only replicate, but also amplify existing biases or discrimination in historical data [32]. Therefore, fairness has especially been becoming one of actively discussed ethical concerns in AI-informed decision making tasks where multiple parties are usually involved and affected by decisions. Fairness is defined as a global perception of appropriateness – a perception that tends to lie theoretically downstream of justice [9]. In the algorithmic context, fairness means that algorithmic decisions should not create discriminatory or unjust consequences [28]. Examples of bias discrimination are 1) banks evaluating credit risks based on race or gender and not on financial score,

and 2) courts judging the recidivism rate of prisoners based on races. Algorithmic fairness is a complicated topic and extensive research has been investigated focusing on fairness definitions (ranging from statistical bias, group fairness, to individual fairness) and unfairness quantification [10, 12, 22].

Taken the recruiting scenario in the human resources context as an example, AI algorithms are often used to shortlist applicants. The laws such as Australia's Anti-Discrimination Law require that different groups (e.g. male and female) should have equal employment opportunity, which implies that the shortlisting should keep a similar proportion of both male and female candidates for the fairness (equal opportunity for male and female candidates). The AI algorithm is designed and trained to meet such fairness requirement. When the AI algorithm is used to shortlist candidates, female candidates are hurt by the AI algorithm if they are shortlisted with a less proportional number than male candidates. This means that the level of fairness of the AI algorithm is not high enough. In addition, AI model accuracy is another factor that affects user's responses to AI solutions such as user trust [29]. For example, if the AI model accuracy is low, it may affect recruiters' trust because they may not get the most appropriate candidates for a position. However, if the AI model accuracy is very high, the applicants may have questions on the fairness of decisions since fairness usually comes with a trade-off over AI model accuracy [19, 23, 27]. As it can be seen from this recruiting scenario example, at least two parties are involved in AI-informed decision making: decision-makers (recruiters in this example) and people affected by decisions (applicants in this example). The influence of AI-informed decision making on them and their expectations are different: recruiters prefer high model accuracy to get the most appropriate candidates, while applicants prefer high fairness in recruiting to get equal opportunities. However, decisions usually cannot meet preferences from both parties at the same time so that both parties agree to and are satisfied with decisions.

As a result, important questions are posed on the use of AI:

- Whether people in different roles in AI-informed decision making have different perception of fairness, trust, and satisfaction with decision making?
- Whether there is a compromise on fairness between people in different roles in AI-informed decision making?

In order to answer these questions, this paper takes AI-assisted talent shortlisting as a case study and investigates perception of fairness, trust, and satisfaction with decision making of both recruiters and applicants in AI-informed decision making. Different introduced fairness (refers to the inherent algorithmic fairness) and model performance are introduced and manipulated in AI-informed decision making tasks. The responses in perception of fairness, trust, and satisfaction from recruiters and applicants at each introduced fairness level and model performance are compared to find any differences in responses from recruiters and applicants. Compromises on fairness between decision-maker and people affected by decisions are identified if both parties have the same responses in perception of fairness, trust, or satisfaction under a given introduced fairness level and model performance. A user study has been conducted to answer research questions.

2 Related Work

2.1 Fairness-Accuracy Trade-Off

A large amount of work has shown that fairness usually comes with a trade-off over accuracy. Zliobaite [38] presented a theoretical and empirical analysis of trade-offs between accuracy and fairness. They argued that comparison of non-discriminatory classifiers needs to account for different rates of positive predictions, otherwise conclusions about performance may be misleading in binary classification. Martinez et al. [20] used Pareto frontiers to dynamically re-balance subgroups' risks to minimize performance discrepancies across sensitive groups without causing unnecessary harm. They argue that even in domains where fairness at cost is required, finding a non-unnecessary-harm fairness model is the optimal initial step. Pleiss et al. [23] investigated the tension between minimizing unfairness across different population groups while maintaining calibrated predictions. It shows that maintaining cost parity and calibration is desirable yet often difficult in practice. They argue that as long as calibration is required, no lower-error solution can be achieved.

Wang et al. [27] showed that traditional approaches that mainly focus on optimizing the Pareto frontier of multi-task accuracy might not perform well on the trade-off between group fairness and accuracy. They proposed a new set of metrics to better capture the multi-dimensional Pareto frontier of fairness-accuracy trade-offs uniquely presented in a multi-task learning setting. Zhao and Gordon [31] theoretically and empirically investigated the problem of quantifying the trade-off between utility and fairness in learning group-invariant representations. They proved a lower bound to characterize the trade-off between fairness and the utility across different population groups.

2.2 Human Responses to AI

Since AI is often used by humans and/or for human-related decision making [26], humans' responses to AI play an important role in AI-informed decision making. This section reviews some of the most investigated human responses to AI including human's perceived fairness (perception of fairness), trust, and satisfaction.

The perception of fairness is a central component of maintaining satisfactory relationships with humans in decision making [1]. The perception of fair treatment on customers is found to be important in driving trustworthiness and engendering trust in the banking context [24].

In AI-informed decision making, algorithmic factors have been studied on how the technical design of an AI system affects people's fairness perceptions. For example, Lee et al. [16] found that people had different variations in the preferences for the three fairness metrics (equality, equity, efficiency) impacted by the decision. Human-related information has also been investigated on their effects on the perception of fairness. For example, education and age have been found affecting both perceptions of algorithmic fairness and people's reasons for

the perception of AI fairness [14]. Zhou et al. [37] found that introduced fairness is positively related to perception of fairness.

User trust in AI-informed decision making has been extensively investigated from different perspectives. Zhou et al. [33, 36] argued that communicating user trust benefits the evaluation of effectiveness of machine learning approaches. Confidence score, model accuracy and users' experience of system performance have been studied on their effects on user trust [30, 34]. Zhou et al. [35] found that the presentation of influences of training data points significantly increased the user trust in predictions, but only for training data points with higher influence values under the high model performance condition.

Theoretical arguments and empirical evidence suggests that satisfaction be among the most important of reactions to the appraisal process [15]. User's satisfaction is another factor that affects the effectiveness of AI-informed decision making. For example, Allam and Mueller [2] found that visual and example-based explanations integrated with rationales had a significantly impact on patient satisfaction in AI diagnostic systems.

These previous work primarily focuses on responses from one party such as decision-maker's response or response of people affected by decisions in AI-informed decision making. However, less attention has been paid to responses from both sides of decision-makers and people affected by decisions in AI-informed decision. This study investigates the responses from both sides in AI-informed decision to find their differences and whether there is a compromise over decisions.

3 Preliminary Knowledge

Fairness is a complex and multi-faceted concept that depends on context and culture [3]. Various mathematical definitions of fairness have been summarised because of various reasons such as different contexts/applications, different stakeholders, impossibility theorems, as well as allocative versus representational harms. It shows that it is impossible to satisfy all definitions of fairness at the same time [3].

In this study, the statistical parity, one of group fairness definitions, is used to represent fairness. The statistical parity suggests that a predictor is fair if the prediction \hat{Y} is independent of the protected attribute Z so that

$$P(\hat{Y}|Z) = P(\hat{Y}). \quad (1)$$

It also means that subjects in both protected and unprotected groups have equal probability (P) of being assigned to the positive predicted class. Taken the recruitment as an example, this would imply equal probability for male and female applicants to have positive predicted recruitment:

$$P(\hat{Y} = 1|Z = 0) = P(\hat{Y} = 1|Z = 1) \quad (2)$$

where $Z = 0$ represents male applicants and $Z = 1$ represents female applicants. Based on these preliminaries, statistical parity difference (PD) is defined as:

$$PD = \left| P\left(\hat{Y} = 1|Z = 0\right) - P\left(\hat{Y} = 1|Z = 1\right) \right| \quad (3)$$

where PD is in the range of $[0, 1]$. $PD = 0$ represents the complete fairness, and $PD = 1$ represents the complete unfairness. This paper manipulates various fairness levels of PD between $[0, 1]$ (called introduced fairness in this paper) to learn how introduced fairness is perceived and affects user responses in algorithmic decision making.

4 Method

4.1 Case Study

A company needs to recruit staff for a position. They posted the position description and a large number of applicants submitted their applications for the position. A machine learning system named Automatic Recruiting Assistant (ARA) is simulated to help to process applications and shortlist applicants for interviewing. ARA is a laboratory simulated candidate assessment tool that is supposed to use historical recruiting data to train a machine learning model and predict whether a candidate will be shortlisted.

In this study, a participant is told to act as either a Recruiter (R) or an Applicant (A) but not both. The participant is then required to conduct tasks and answer questions by giving information on the ARA performance information and shortlisting information of male and female applicants as a role of recruiter or applicant.

4.2 Fairness-Performance Space

In this study, introduced fairness (defined in Eq. 3) and model performance of ML models are manipulated and presented to participants to investigate responses of participants on the perception of fairness, trust, and satisfaction. Therefore, introduced fairness and performance form a 2D space. In this 2D space, each point represents a task condition of introduced fairness and model performance pair (f, p) . The values in the dimension of model performance investigated include 70%, 80%, and 90% which correspond to low, middle, and high model performance respectively.

In the fairness dimension of the 2D space, the gender of applicants is used as the protected attribute in the recruitment scenario. The PD is used to measure the fairness and defined as the difference of shortlisted rate by the gender. In this study, fairness is introduced by manipulating PD with its discrete values of 0, 0.1, 0.2, 0.3, ..., 0.8, 0.9, and 1.0, where each PD 's discrete value was used as a measure of fairness to define the number of male and female applicants as well as number of male and female applicants shortlisted in each task respectively.

4.3 Task Design

In this study, tasks with different model performance and introduced fairness pair conditions were designed to investigate their effects on user’s perception of fairness, trust, and satisfaction in AI-informed decision making. Table 1 shows 11 fairness presentation examples corresponding to different PD values. In this table, “Rate (M)” and “Rate (F)” represent the predicted success rate of male and female applicants respectively, “Male #” and “Female #” represent the number of male and female applicants respectively, and “Listed Male #” and “Listed Female #” represent the number of shortlisted male and female applicants respectively. All together 33 (11×3) tasks were designed and conducted by each participant based on eleven (11) fairness presentation examples and three (3) model performance levels (70%, 80%, 90%). Two additional training tasks were also conducted by each participant before the formal tasks. The order of formal tasks was randomized during the experiment to avoid any bias.

Table 1. Examples of fairness presentation in tasks.

Example#	PD	Rate (M)	Rate (F)	Male#	Female#	Listed Male#	Listed Female#
1	0	0.8	0.8	10	10	8	8
2	0.1	0.7	0.8	10	5	7	4
3	0.2	0.6	0.8	5	5	3	4
4	0.3	0.8	0.5	5	10	4	5
5	0.4	0.8	0.4	5	5	4	2
6	0.5	0.7	0.2	10	5	7	1
7	0.6	0.8	0.2	5	5	4	1
8	0.7	0.1	0.8	10	5	1	4
9	0.8	0.9	0.1	10	10	9	1
10	0.9	0.1	1	10	10	1	10
11	1	1	0	5	10	5	0

During the task time, each pair of fairness and model performance is firstly presented to participants with visualisations. Figure 1 shows the screenshot of visualisations in a task conducted in the experiment. The left barchart shows the number of applicants and number of applicants shortlisted by ARA for both males and females, which implies the fairness status in shortlisting for males and females. The right circular chart represents the model accuracy in shortlisting. After reading these information, participants are then asked to agree or reject decisions made by the ARA followed by different survey questions on perception of fairness, trust, and satisfaction in AI-informed decision making.

4.4 Scales of User Responses

Different questionnaires with Likert-type response scales are used in this study to collect responses of perception of fairness, trust, and satisfaction of users. The

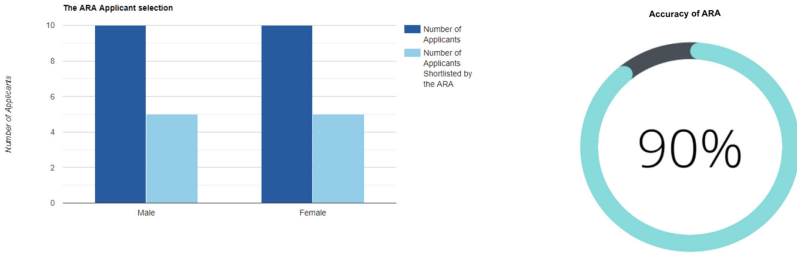


Fig. 1. Visualisation of fairness and model performance (accuracy of ARA).

scale is on a 5-point Likert-type response scale ranging from 1 (strongly disagree) to 5 (strongly agree) for each questionnaire on perception of fairness, trust, and satisfaction respectively.

Trust Scales. Trust is assessed with four items using self-report scales as the following [21].

- I am happy with help provided by the ARA.
- I have confidence in the advice given by the ARA.
- I can depend on the ARA.
- I can trust the ARA to make the correct selection.

Scales of Perception of Fairness. The perception of fairness of participants is assessed with the following two items.

- Overall, female and male applicants are treated fairly by ARA.
- I believe the ARA is a competent performer for both men and women.

Scales of Satisfaction. The satisfaction of participants is assessed with the following item [15,25]: overall, I am satisfied with the recruiting by considering both the performance of ARA and the fairness.

4.5 Experiment Setup

Due to social distancing restrictions and lockdown policies during the COVID-19 pandemic, this experiment was implemented using Python web framework and was deployed on the cloud server online. The deployed application link was then shared with participants to invite them to conduct tasks. In this study, participant responses to tasks were stored in a MySQL database.

4.6 Participants and Data Collection

In this study, 60 participants were recruited to conduct experimental tasks via various means of communications such as emails, text messages and social media posts who were mainly university students and 19 participants were females. Of all participants, 30 participants randomly acted as job applicants and other 30 participants acted as HR recruiters in the experiment.

After each task was displayed on the screen, the participants were asked to answer questions based on the task on perception of fairness, trust, and satisfaction in the AI-informed decision making respectively.

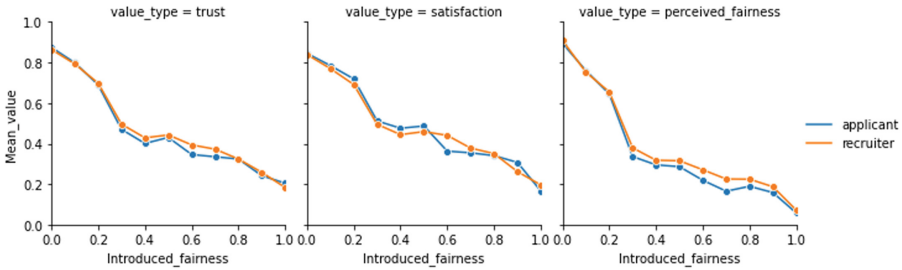


Fig. 2. Overall average responses in trust, satisfaction, and perception of fairness regardless of model performance.

5 Analyses and Results

This section analyses the collected data to answer our questions. We aim to understand whether two parties affected by decisions from AI have the same responses to AI-informed decision making from the perspectives of perception of fairness, satisfaction and trust.

When two parties have similar responses to decisions from AI under a given introduced fairness condition, it shows that they both agree with the effects of the specific introduced fairness on the decision. We can say that there is a compromise between two parties regarding the introduced fairness despite the decision maybe affecting them differently. When two parties show different responses to decisions under a given introduced fairness condition, it implies that there is a disagreement between two parties regarding the introduced fairness. The outcomes of the study can be used to customise user interface or take different measures when there is no compromise.

In order to perform the analyses, we first normalised the collected data of trust, satisfaction, and perception of fairness with respect to each subject to minimise individual differences in rating behavior using the following equation:

$$V_i^N = \frac{V_i - V_i^{min}}{V_i^{max} - V_i^{min}} \tag{4}$$

where V_i and V_i^N are the original rating values and the normalised rating values respectively from the user i , V_i^{min} and V_i^{max} are the minimum and maximum of the ratings of trust, satisfaction, or perception of fairness respectively from the user i in all of his/her tasks.

Figure 2 shows the overall average responses of participants in trust, satisfaction, and perception of fairness (or perceived fairness) regardless of model performance. t-tests were used to compare differences in trust, satisfaction, and perception of fairness between applicants and recruiters at each introduced fairness level. There are no statistically significant differences found in both trust and perception of fairness between applicants and recruiters at each introduced fairness level. However, it was found that recruiters showed statistically significantly higher trust than applicants at the introduced fairness level of 0.6 ($t = 1.9905$, $p < .048$), and no significant differences were found in trust between recruiters and applicants at other introduced fairness levels. The results also show the decreasing trends of trust, satisfaction, and perception of fairness with the increase of PD values on the horizontal axis (the decrease of introduced fairness levels), which is consistent with the previous research [37].

Figure 3 shows the average responses of participants in trust, satisfaction, and perception of fairness per different model performances. t-tests were applied to compare differences in trust, satisfaction, and perception of fairness between applicants and recruiters at each introduced fairness level under different model performances. From Fig. 3, it was found that:

- As it is expected, the recruiters have overall lower satisfaction when performance is low (at the region 1 in Fig. 3), and the applicants have overall lower satisfaction when fairness is low while performance is high. However, we observed the higher satisfaction at the region 1 for applicants even if the fairness is low. If we compare it to the region 2, then we can see the actual value at the region 1 is lower than the region 2. Here we argue that the applicants' satisfaction is higher than recruiters due to the low model performance.
- We observed that there was a significantly higher level of satisfaction from recruiters than applicants at the region 2 ($t = 2.4918$, $p < .0156$). This can be explained that even recruiters thought the fairness was poor, they were still satisfied with ARA.
- We also observed that recruiters showed lower trust under low model performance (at the region 3), this is further affected by fairness. If we compare the region 3 to the region 4, we can see that recruiters trust less at the region 3. We assume that the recruiters may consider fairness-accuracy trade-off here, since we can observe that their trust at the region 4 is higher than applicants, where the fairness is lower.
- We observed that the compromised setting can be achieved. It is obvious that high model performance (90%) and high fairness (close to 0 of introduced fairness) were highly rated and satisfied by both parties (the region 5). And low performance (70%) and low fairness were rated low and less satisfied by both parties (the region 1 and the region 4). The more compromised setting is at the region 7 that both parties had the same satisfaction.

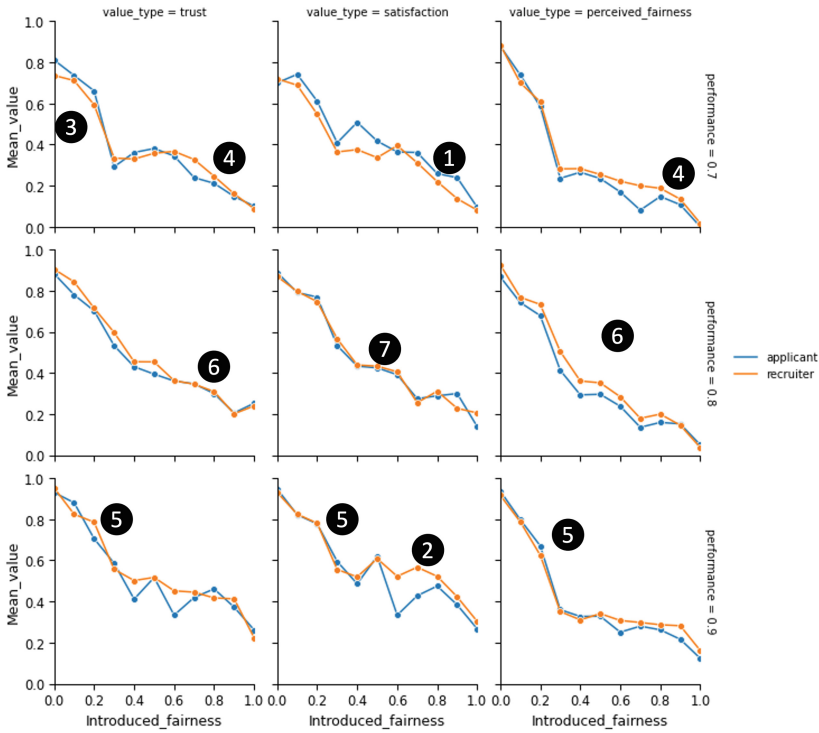


Fig. 3. Average responses of participants in trust, satisfaction, and perception of fairness per different model performances.

- Recruiters showed statistically significantly higher level of perception of fairness than applicants at the introduced fairness level of 0.7 when the model performance is 70% ($t = 2.8366, p < .0062$). Furthermore, we can see that almost all recruiters rated the perception of fairness higher than applicants when the model performance is 70%. This is maybe because that the recruiters may expect lower fairness to improve the performance given the trade-off between accuracy and fairness. However, recruiters showed statistically significantly lower level of satisfaction than applicants at the introduced level of 0.4 when the model performance is 70% ($t = 3.1949, p < .0023$). Under each studied model performances, we have not found other significant differences between recruiters and applicants in trust, satisfaction, and perception of fairness at different introduced fairness levels.

6 Discussions

Multiple parties are usually involved in an AI-informed decision making, e.g. decision-maker and people affected by decisions. Different parties may have different responses to a decision from AI-informed decision making. This study took the AI-assisted talent shortlisting as a case study and investigated satisfaction, trust, and perception of fairness of parties (recruiters and applicants) related to decisions respectively. The results showed that compromises on fairness did exist in AI-informed decision making under given model performances and introduced fairness levels.

Fairness heuristic theory [6, 18] suggests that when individuals face uncertain circumstances they rely on impressions of fairness to determine whether to cooperate and enter into exchange relationships with the other party, which suggests that individuals use fairness judgements to form their perceptions of trust. The social exchange theory [5] also argues that fair actions and the treatment by one party generate reciprocation in the form of trust by the other party in the exchange. In the context of talent shortlisting in human resource settings used in this paper, recruiters were unsure about the outcomes from the Automatic Recruiting Assistant when the model performance was low, resulting in the low perception of fairness as shown in the region 4 in the right diagram of the first row in Fig. 3, and therefore also resulting in low trust as shown in the region 4 in the left diagram of the first row in Fig. 3. The similar conclusion was observed for applicants as stated in the previous section.

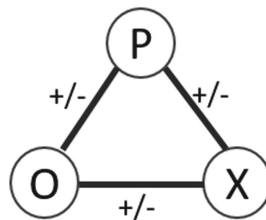


Fig. 4. Heider's POX model.

In the psychology of motivation, *balance theory* proposed by Fritz Heider [13] conceptualizes the cognitive consistency motive as a drive toward psychological balance. It assumes that individuals retain their psychological balance and develop their relationships with others or things within their circumstances. They prefer to maintain a balanced state through a series of cognitive operations that balance out their likes (represented by “+”) and dislikes (represented by “-”) to create equilibrium. Balance theory is often termed POX theory, representing the balanced/imbalanced state of individuals from the relationships among one person (P), the other person (O), and an attitudinal thing or object (X), as shown in Fig. 4. In this triadic relationship, a balance is achieved when

there are three positive (+) links or two negatives (-) with one positive. Balance theory has been used in social psychology to understand various interpersonal relationships such as service quality, customer behaviour understanding [7, 17]. Such balance theory can be used to explain the satisfaction of applicants and recruiters across different model performances in the talent shortlisting example conducted in this paper. As shown in Fig. 3, applicants showed an overall higher satisfaction level with AI than recruiters when the model performance is 70%, and vice versa when the model performance is 90%. All these result in “tensions” between applicants and recruiters. To reduce “tensions”, this study modulated the model performance to 80%, and recruiters and applicants reached a balanced state (the region 7), where recruiters and applicants compromised and had the similar level of satisfaction.

The findings from this study can be used to help find compromising points between decision-maker and people affected by decisions so that both parties can reach for a balanced state in AI-informed decision making. Such findings also suggest AI developers as well as AI users that different stakeholders can be considered together in AI-informed decision making so that all stakeholders can satisfy with decisions.

7 Conclusion and Future Work

Since multiple parties are usually affected by decisions in AI-informed decision making and they have different responses regarding the fairness, this paper investigated whether there is a compromise on fairness in using AI models by examining user’s satisfaction, trust, and perception of fairness in AI-informed decision making. The paper took the AI-assisted talent shortlisting as a case study to compare responses to decisions from recruiters and applicants. The results showed that compromises on fairness did exist in AI-informed decision making under given model performances and introduced fairness levels, which can be used to help find compromising points between decision-maker and people affected by decisions so that both parties can reach a balanced state. The future work of this study will focus on the setup of a compromise profile for an AI-informed decision making through investigation of wider model performances such as from 50% to 100% and such profile can be used to guide the use of AI solutions for more effective decision making.

References

1. Aggarwal, P., Larrick, R.P.: When consumers care about being treated fairly: the interaction of relationship norms and fairness norms. *J. Consumer Psychol.* **22**(1, SI), 114–127 (2012)
2. Alam, L., Mueller, S.: Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Med. Inform. Decision Making* **21**(1), 178 (2021). <https://doi.org/10.1186/s12911-021-01542-6>
3. Bellamy, R.K.E., et al.: AI fairness 360: an extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv:1810.01943 [cs]* (2018)

4. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: the state of the art. *Sociological Methods & Research* p. 0049124118782533 (2018)
5. Blau, P.M.: *Exchange and Power in Social Life*. Wiley, New York, NY (1964)
6. van den Bos, K.: Uncertainty management: the influence of uncertainty salience on reactions to perceived procedural fairness. *J. Person. Soc. Psychol.* **80**(6), 931–941 (2001)
7. Carson, P.P., Carson, K.D., Knouse, S.B., Roe, C.W.: Balance theory applied to service quality: a focus on the organization, provider, and consumer triad. *J. Bus. Psychol.* **12**(2), 99–120 (1997). <https://doi.org/10.1023/A:1025061816323>,
8. Chen, F., Zhou, J.: *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*. Springer, Cham (2022). <https://doi.org/10.1007/978-3-030-72188-6>
9. Colquitt, J.A., Rodell, J.B.: Measuring justice and fairness. In: Cropanzano, R.S., Ambrose, M.L. (eds.) *The Oxford Handbook of Justice in the Workplace*, pp. 187–202. Oxford University Press (2015)
10. Corbett-Davies, S., Goel, S.: The measure and mismeasure of fairness: a critical review of fair machine learning. arXiv preprint [arXiv:1808.00023](https://arxiv.org/abs/1808.00023) (2018)
11. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: *Proceedings of KDD2015*, pp. 259–268 (2015)
12. Glymour, B., Herington, J.: Measuring the biases that matter: the ethical and casual foundations for measures of fairness in algorithms. In: *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 269–278 (2019)
13. Heider, F.: *The Psychology of Interpersonal Relations*. Wiley (1958)
14. Helberger, N., Araujo, T., de Vreese, C.H.: Who is the fairest of them all? public attitudes and expectations regarding automated decision-making. *Comput. Law Secur. Rev.* **39**, 105456 (2020)
15. Jawahar, I.M.: The influence of perceptions of fairness on performance appraisal reactions. *J. Labor Res.* **28**(4), 735–754 (2007)
16. Lee, M.K., Jain, A., Cha, H.J., Ojha, S., Kusbit, D.: Procedural justice in algorithmic fairness: leveraging transparency and outcome control for fair algorithmic mediation. *Proc. ACM Hum. Comput. Interact.* **3**, 1–26 (2019)
17. Lin, C.F., Fu, C.S., Chen, Y.T.: Exploring customer perceptions toward different service volumes: an integration of means-end chain and balance theories. *Food Qual. Preferen.* **73**, 86–96 (2019)
18. Lind, E.: Fairness heuristic theory: justice judgments as pivotal cognitions in organizational relations. In: *Advances in Organizational Justice*, pp. 56–88. Stanford University Press (2001)
19. Liu, S., Vicente, L.N.: Accuracy and fairness trade-offs in machine learning: a stochastic multi-objective approach. arXiv preprint [arXiv:2008.01132](https://arxiv.org/abs/2008.01132) (2020)
20. Martinez, N., Bertran, M., Sapiro, G.: Fairness with minimal harm: a pareto-optimal approach for healthcare. arXiv preprint [arXiv:1911.06935](https://arxiv.org/abs/1911.06935) (2019)
21. Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I trust it, but i don't know why: effects of implicit attitudes toward automation on trust in an automated system. *Hum. Fact.* **55**(3), 520–534 (2013)
22. Nabi, R., Shpitser, I.: Fair inference on outcomes. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 2018, p. 1931. NIH Public Access (2018)
23. Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., Weinberger, K.Q.: On fairness and calibration. *Adv. Neural Inform. Process. Syst.* **30** (2017)

24. Roy, S.K., Devlin, J.F., Sekhon, H.: The impact of fairness on trustworthiness and trust in banking. *J. Market. Manage.* **31**(9–10), 996–1017 (2015)
25. Sholihin, M.: How does procedural fairness affect performance evaluation system satisfaction? (evidence from a UK police force). *Gadjah Mada Int. J. Bus.* **15**, 231–247 (2013). <https://doi.org/10.22146/gamaijb.5445>
26. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: a systematic review of the empirical literature (2021)
27. Wang, Y., Wang, X., Beutel, A., Prost, F., Chen, J., Chi, E.H.: Understanding and improving fairness-accuracy trade-offs in multi-task learning. In: Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 1748–1757 (2021)
28. Yang, K., Stoyanovich, J.: Measuring fairness in ranked outputs. *SSDBM 2017* (2017). <https://doi.org/10.1145/3085504.3085526>
29. Yu, K., Berkovsky, S., Taib, R., Zhou, J., Chen, F.: Do i trust my machine teammate? an investigation from perception to decision. In: Proceedings of the 24th International Conference on Intelligent User Interfaces, pp. 460–468. *IUI 2019*, ACM (2019)
30. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. *FAT* 2020* (2020)
31. Zhao, H., Gordon, G.: Inherent tradeoffs in learning fair representations. *Adv. Neural Inform. Process. Syst.* **32** (2019)
32. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: reducing gender bias amplification using corpus-level constraints. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2979–2989. Copenhagen, Denmark (2017)
33. Zhou, J., Bridon, C., Chen, F., Khawaji, A., Wang, Y.: Be informed and be involved: effects of uncertainty and correlation on user’s confidence in decision making. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems, pp. 923–928 (2015)
34. Zhou, J., Chen, F. (eds.): *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Springer, Cham (2018)
35. Zhou, J., Hu, H., Li, Z., Yu, K., Chen, F.: Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In: *Machine Learning and Knowledge Extraction*, pp. 94–113 (2019)
36. Zhou, J., et al.: Measurable decision making with GSR and pupillary analysis for intelligent user interface. *ACM Trans. Comput. Hum. Interact.* **21**(6), 1–23 (2015)
37. Zhou, J., Verma, S., Mittal, M., Chen, F.: Understanding relations between perception of fairness and trust in algorithmic decision making. In: Proceedings of the International Conference on Behavioral and Social Computing (BESC 2021), pp. 1–5 (2021)
38. Zliobaite, I.: On the relation between accuracy and fairness in binary classification. *arXiv preprint [arXiv:1505.05723](https://arxiv.org/abs/1505.05723)* (2015)