Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making

Jianlong Zhou, Sunny Verma‡, Mudit Mittal‡, and Fang Chen
Data Science Institute, University of Technology Sydney, Sydney, Australia
Jianlong.Zhou@uts.edu.au

***Abstract*—The adoption of Artificial Intelligence (AI) is currently under scrutiny due to various concerns such as fairness, and how does the fairness of an AI algorithm affects user's trust is much legitimate to pursue. In this regard, we aim to understand the relationship between induced algorithmic fairness and its perception in humans. In particular, we are interested in whether these two are positively correlated and reflect substantive fairness. Furthermore, we also study how does induced algorithmic fairness affects user trust in algorithmic decision making. To understand this, we perform a user study to simulate candidate shortlisting by introduced (manipulating mathematical) fairness in a human resource recruitment setting. Our experimental results demonstrate that different levels of introduced fairness are positively related to human perception of fairness, and simultaneously it is also positively related to user trust in algorithmic decision making. Interestingly, we also found that users are more sensitive to the higher levels of introduced fairness than the lower levels of introduced fairness. Besides, we summarize the theoretical and practical implications of this research with a discussion on perception of fairness.**

***Index Terms*—Introduced fairness, perception of fairness, trust**

I. Introduction

Artificial Intelligence (AI) has powerful capabilities in prediction, automation, planning, targeting, and personalisation [1]. It has been increasingly used to make important decisions that affect human lives in different areas ranging from social and public management to promote productivity for economic well-being. For example, AI can be used to decide the loan approval in banks and manage engagement and outcomes of job for workers within an organization. These algorithms are also utilized by various hiring platforms to recommend and recruit candidates in human resource settings [2], [3] (such AI-informed decision making is also called algorithmic decision making). Besides all these functionalities of AI, a paramount concern with AI's decision making is equal treatment or equitability of decision based on people's performance or needs [4] is required [5], [6]. This setting of equitable treatment is also known as fairness in AI. On the other hand, unintentional (or intentional) discrimination can cause unfairness in AI and lead to poor decision making. Thus fairness becomes critical as a fair decision making system amplifies the satisfaction levels with algorithmic decision making [5], [6]. Often, the fairness is a consequence of either the training data or the design of machine learning models, which is the fairness human actually perceives in algorithmic decision making, ultimately affect their adoptions in real-world applications [7]. Meanwhile, AI models are usually "black-boxes" for users and even for AI experts [8], [9], which causes trust issues in AI-informed decision making. Considerable research on fairness has evidenced that fairness perceptions are linked to trust such as in management and organizations [10], [11].

Different from above, in algorithmic decision making, mathematical fairness introduced by AI models and/or data (also refers to *introduced fairness* in this paper) is perceived by humans (also refers to *perception of fairness* in this paper) implicitly or explicitly. The perceived fairness is a central component of maintaining satisfactory relationships with humans in decision making [12]. Given various mathematical formulations of fairness, three major findings are: 1) demographic parity most closely matches human perception of fairness [13]; 2) effects of transparency and outcome control on perceived fairness [14]; and 3) factors affecting perceptions of fairness in algorithmic decision making [15]. While the fairness (or discrimination) is either introduced by AI models and/or the data, it is critical to understand whether an introduced level of fairness is affecting its perception by humans in algorithmic decision making. Therefore, in this work we aim to investigate the relations between the introduced fairness and human perception of fairness.

This paper aims to understand what is the perception of fairness by humans in particular, is it positive or negative to introduced fairness? Importantly, we further dwell to understand whether the introduced fairness affects users trust in algorithmic decision making. In this regard, we utilise the statistical parity difference [16] as the actual fairness level of an AI system as it has been widely accepted as a metric to measure fairness. We then design a user study to investigate the perception of fairness by simulating a human resource recruitment for candidate shortlisting by manipulating introduced fairness. Our experimental results demonstrate two important findings: 1) introduced fairness is positively related to human perception; and 2) simultaneously, high level of fairness leads to the increased trust in algorithmic decision making. These findings illustrate that trust judgments can be influenced by fairness information which are comprehensively discussed both theoretical and practically.

II. Related Work

The current research on fairness in machine learning focuses on the formalisation of the definition of fairness and quantify-

‡ Work done while working at UTS.

ing the unfairness (bias) of an algorithm with different metrics [17]–[19]. These work typically begins by outlining fairness in the context of different protected attributes (sex, race, origin, culture, etc.) receiving equal treatments by algorithms [16], [20]. Despite the proliferation of fairness definitions and unfairness quantification approaches [21], little work is found to investigate human's perceived fairness (perception of fairness) when the fairness defined by a specific definition is introduced. This paper uses statistical parity as the definition of fairness to investigate human perception of fairness in algorithmic decision making.

Various researches have been investigated to learn user trust variations in algorithmic decision making. Zhou et al. [22], [23] argued that communicating user trust benefits the evaluation of effectiveness of machine learning approaches. Kizilcec [24] proposed that the transparency of algorithm interfaces can promote awareness and foster user trust. Other studies that empirically tested the importance of explanation to users, in various fields, consistently showed that explanations significantly increase users' confidence and trust [25], [26]. Zhou et al. [27] investigated the effects of presentation of influence of training data points on predictions to boost user trust. In addition, researchers found that user trust had significant correlations with users' experience of system performance [8]. These previous work primarily focuses on the investigation of effects of explanation and model performance on user trust in algorithmic decision making. However, less attention has been paid to the perception of fairness and its effects on trust, which is investigated by this paper.

It was found that perceptions of fair treatment on customers are important in driving trustworthiness and engendering trust in the banking context [11]. Previous work pays more attention to relations between the perception of fairness especially procedural fairness [28] and user trust in social interaction context such as marketing and services [7], [29], however, little work is found on the effects of fairness on user trust in algorithmic decision making. This study investigates whether the introduced fairness is positively received by humans and how such fairness affects user trust by simulating a candidate shortlisting in a human resource recruitment setting in algorithmic decision making.

III. Hypotheses

This paper poses the following hypotheses:

- H1 The human perceived fairness will be positively related to the introduced fairness. That is, the high level of introduced fairness will result in the high level of perceived fairness by humans, and vice versa;
- H2 The user trust will be positively related to the introduced fairness. That is, the high level of introduced fairness will result in the high level of trust in algorithmic decision making, and vice versa;
- H3 Humans will be more sensitive to the changes of high level of introduced fairness than the low level of introduced fairness.

IV. Method

A. Case Study

In this study, a company is supposed to recruit staff for a position. The company posted the job description and different applicants submitted their applications. Participants are supposed to be a recruiter to shortlist applicants for the position. A machine learning system named Automatic Recruiting Assistant (ARA) is used to help participants make decisions on the shortlisting of applicants. ARA is a simulated candidate assessment software that uses historical recruiting data to train a machine learning model and predict whether a candidate will be shortlisted. The ARA's accuracy is assumed to keep constant for all shortlisting.

B. Fairness and Trust Scales

This study investigates the perception of fairness, which is measured with a single scale that focuses on a global perception of appropriateness [30]. In this study, the questionnaire on the fairness is "overall, female and male applicants are treated fairly by ARA". The scale is on a 5-point Likert-type response scale ranging from 1 (strongly disagree) to 5 (strongly agree).

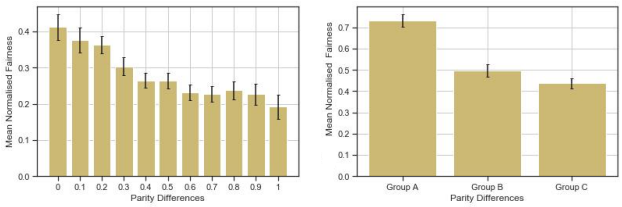
Trust is assessed with six items using self-report scales based on [31]. The scale is on a 5-point Likert-type response scale ranging from 1 (strongly disagree) to 5 (strongly agree).

V. Experiment

A. Task Design

In this study, statistical parity difference (PD) is used to measure the fairness [21]. PD is defined as the probability difference that protected ($Z = 1$) and unprotected ($Z = 0$) groups being assigned to the positive predicted class Y: $PD = \left| P\left(\hat{Y} = 1 | Z = 0\right) - P\left(\hat{Y} = 1 | Z = 1\right) \right|$. 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 $[0, 1]$ to learn how introduced fairness is perceived and affects trust in algorithmic decision making.

Tasks were designed to investigate effects of different fairness levels on user trust in algorithmic decision making. The protected attribute in this study is the gender of applicants. In this case, the PD is the difference of shortlisted rate by the gender. In this study, fairness was introduced by manipulating PD with its discrete values of 0, 0.1, 0.2, 0.3, 0.4, ..., 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. Table I shows 11 task examples corresponding to different PD values. In this table, "Rate (Male)" represents the predicted success rate for male applicants, "Rate (Female)" represents the predicted success rate for female applicants, "Male #" represents the number of male applicants, "Female #" represents the number of female applicants, "Listed Male #" represents the number of shortlisted male applicants, and "Listed Female #" represents the number of shortlisted female applicants. With the same



(a) Non-grouped (b) Grouped

Fig. 1: Mean normalised perceived fairness over introduced fairness.

settings of PD as in the table, different number of male and female applicants were used to generate another 11 tasks. All together 22 tasks were conducted by each participant. Two additional training tasks were also conducted by each participant before the formal tasks. The order of tasks was randomized during the experiment to avoid any bias.

B. Experiment Setup

Due to social distancing restrictions and lockdown policies during the COVID-19 pandemic, our experiment was implemented using the flask framework in Python and was deployed on a cloud server online. The deployed application link was then shared with participants to invite them to conduct tasks.

C. Participants and Data Collection

20 participants were invited via various means of communications such as emails, text messages and social media posts who are mainly university students around the age group of 20-30 years with the average of around 25 years old. After each task was displayed on the screen, the participants were asked to answer seven questions based on the task. The first question was on fairness of applicant shortlisting shown in the task while the other six questions were on the trust of the participant in the decision making from the ARA.

VI. Results

This study aims to understand: 1) how the introduced fairness is perceived by humans, and 2) how the introduced fairness affects user trust. In order to perform the analyses, we first normalised the collected trust and fairness data. We then performed one-way ANOVA tests on the normalised data followed by post-hoc comparison using Tukey HSD tests. The fairness and trust values were normalised with respect to each subject to minimise individual differences in rating behavior using the equation given as: $T_i^N = \frac{T_i - T_i^{min}}{T_i^{max} - T_i^{min}}$, where T_i and T_i^N are the original fairness or trust ratings and the normalised fairness or trust rating respectively from the user i, T_i^{min} and T_i^{max} are the minimum and maximum of the ratings respectively from the user i in all of his/her tasks.

A. Perception of Fairness

Figure 1a shows the mean normalised perceived fairness (perception of fairness) over introduced fairness (error bars represent the 95% confidence interval of a mean and it is the same in other figures). A one-way ANOVA test found that there were statistically significant differences in perceived fairness among 11 introduced fairness levels ($F(10, 429) = 29.872, p < .000$). The further post-hoc comparison with Tukey HSD tests were conducted to test pair-wised differences in perceived fairness between two introduced fairness levels. It was found that the perceived fairness at $PD = 0, 0.1$, and 0.2 had significant differences with all other PD levels from 0.4 to 1.0 respectively (for all, $p < .001$). The perceived fairness at $PD = 0$ ($p < .001$) and 0.1 ($p < .005$) also had significant differences with $PD = 0.3$ respectively. However, there were no significant differences found in perceived fairness among any pair of PD at $0, 0.1$, and 0.2. It was also found that the perceived fairness at $PD = 0.3$ had significant differences with $PD = 0.6, 0.7, ..., 1.0$ respectively (for all, $p < .017$). Furthermore, the perceived fairness at $PD = 0.4$ ($p < .006$) and 0.5 ($p < .005$) had significant differences with $PD = 1.0$ respectively. Despite no other significant difference found in perceived fairness among introduced fairness levels, Figure 1a shows that the perceived fairness has a clear decreasing trend with the decrease of introduced fairness (increase of PD levels). The results suggest that participants' perception of fairness was positively related to the introduced fairness (H1), but was not sensitive to the small changes of introduced fairness. Moreover, participants were more sensitive to the perceived fairness with high levels than low levels as we expected (H3). These findings also imply that the introduced fairness can be safely used to validate the perception of fairness of humans.

Following the findings of the trend of perceived fairness, we divided introduced fairness into three groups:

- Group A (high level of introduced fairness group): $PD = 0, 0.1, 0.2, 0.3$;
- Group B (middle level of introduced fairness group): PD=0.4, 0.5, 0.6, 0.7;
- Group C (low level of introduced fairness group): $PD = 0.8, 0.9, 1.0$.

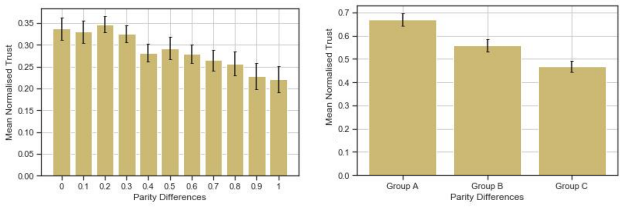
A one-way ANOVA test found that there were statistically significant differences in perceived fairness among three introduced fairness levels ($F(2, 117) = 104.725, p < .000$). The post-hoc comparison with Tukey HSD tests found that the perceived fairness at the introduced fairness level of Group A was significantly higher than that at levels of Group B ($p < .001$) and Group C ($p < .001$) respectively. The perceived fairness at the introduced fairness level of Group B was also significantly higher than that at the level of Group C ($p < .001$). The results show that human perception of fairness was positively related to the introduced fairness. The findings imply that the introduced fairness based on PD can safely reflect perception of fairness in algorithmic decision making.

B. Fairness and Trust

Figure. 2a shows mean normalised trust ratings over introduced fairness (PD) levels. A one-way ANOVA test found that there were statistically significant differences in trust

TABLE I: Experiment tasks.

Task#	PD	Rate (Male)	Rate (Female)	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



(a) Non-grouped (b) Grouped

Fig. 2: Mean normalised trust over fairness.

ratings among 11 fairness levels ($F(10, 429) = 11.550, p <$ .000). Then the post-hoc comparison using Tukey HSD tests found significant differences in trust responses between introduced fairness level pairs. It shows that participants had significantly higher trust in AI-informed decisions under high introduced fairness levels (low PD values) than that under low introduced fairness levels (high PD values). For example, participants had significantly higher trust under $PD = 0$ than that under $PD = 0.7$, $p < .003$. However, user trust did not show significant differences under high introduced fairness levels (e.g. PD = 0, 0.1, 0.2, 0.3).

We further analyse trust differences under three introduced fairness group levels of A, B, C. as described above, a one-way ANOVA test found that there were statistically significant differences in user trust among three introduced fairness group levels ($F(2, 117) = 48.272, p <$ .000). The further post-hoc comparison with Tukey HSD tests found that user trust was significantly higher at the introduced fairness level of Group A than that at the levels of Group B ($p <$ .001) and Group C ($p <$ .001) respectively. User trust was also significantly higher at the introduced fairness level of Group B than that at the level of Group C ($p <$ .001).

The findings suggest that user trust had a positive relationship with the introduced fairness as we expected (H2). The higher the introduced fairness level was, the higher trust in decisions users had.

VII. Discussion

Our study found that the introduced fairness was positively related to the perceived fairness by humans. Besides, it also showed that high levels of introduced fairness resulted in high levels of human perception of fairness. These findings confirm that the introduced fairness level can be safely used to evaluate the human perception of fairness. Furthermore, the introduced fairness was also positively related to user's trust in algorithmic decision making. Once again we see that the high level of introduced fairness benefited user trust. It was also found that participants were more sensitive to the introduced fairness with high levels than low levels.

These findings have significant implications in algorithmic decision making applications. For example, when the trust is difficult to examine in algorithmic decision making, human perception of fairness can be used to estimate user trust in algorithmic decision making. While human perception of fairness is positively related to introduced fairness. Our findings also imply that the introduced fairness can be safely used to validate the human perception of fairness. Furthermore, since human is more sensitive to the high level of fairness, the high level of fairness instead of the low level of fairness can be explicitly presented in the user interface of AI applications to boost user trust in algorithmic decision making.

Overall, the findings from this study at least have the following implications: 1) the estimation of user trust in algorithmic decision making by human perception of fairness; 2) the user interface design of AI applications to boost user trust by explicitly presenting high level of fairness to users; 3) manipulation of human perception of fairness by manipulating level of introduced fairness.

VIII. Conclusion and Future Work

This paper understood the relations between the introduced fairness and human perception of fairness and investigated how the introduced fairness affected user trust in algorithmic decision making. Experimental results showed that the introduced fairness was positively related to human perception of fairness, and concurrently it was also positively related to user's trust. Interestingly, the users were more sensitive to fairness with high levels than those with low levels. The findings can be used to help to estimate trust in algorithmic decision making and user interface design for AI solutions.

A future work of this study will focus on the introduction of AI explanations into the pipeline to understand their effects on user trust in algorithmic decision making.

REFERENCES

[1] F. Chen and J. Zhou, "AI in the public interest," in *Closer to the Machine: Technical, Social, and Legal Aspects of AI*, C. Bertram, A. Gibson, and A. Nugent, Eds. Office of the Victorian Information Commissioner, 2019.

[2] C. Hughes, L. Robert, K. Frady, and A. Arroyos, "Artificial intelligence, employee engagement, fairness, and job outcomes," in *Managing Technology and Middle- and Low-skilled Employees*, 7 2019, pp. 61–68.

[3] A. Gugnani and H. Misra, "Implicit skills extraction using document embedding and its use in job recommendation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, 4 2020, pp. 13 286–13 293.

[4] G. S. Leventhal, "What should be done with equity theory?" in *Social Exchange: Advances in Theory and Research*, K. J. Gergen, M. S. Greenberg, and R. H. Willis, Eds. Springer US, 1980, pp. 27–55.

[5] J. J. Lavelle, G. C. McMahan, and C. M. Harris, "Fairness in human resource management, social exchange relationships, and citizenship behavior: testing linkages of the target similarity model among nurses in the united states," *The International Journal of Human Resource Management*, vol. 20, no. 12, pp. 2419–2434, 12 2009.

[6] L. P. Robert, C. Pierce, L. Marquis, S. Kim, and R. Alahmad, "Designing fair ai for managing employees in organizations: a review, critique, and design agenda," *Human–Computer Interaction*, pp. 1–31, 2020.

[7] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, p. 205395171875668, June 2018.

[8] J. Zhou and F. Chen, Eds., *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*. Cham: Springer, 2018.

[9] J. Zhou, A. H. Gandomi, F. Chen, and A. Holzinger, "Evaluating the quality of machine learning explanations: A survey on methods and metrics," *Electronics*, vol. 10, no. 5, 2021.

[10] M. Komodromos, "Employees' perceptions of trust, fairness, and the management of change in three private universities in cyprus," *Journal of Human Resources Management and Labor Studies*, vol. 2, no. 2, pp. 35–54, July 2014.

[11] S. K. Roy, J. F. Devlin, and H. Sekhon, "The impact of fairness on trustworthiness and trust in banking," *Journal of Marketing Management*, vol. 31, no. 9-10, pp. 996–1017, 2015.

[12] P. Aggarwal and R. P. Larrick, "When consumers care about being treated fairly: The interaction of relationship norms and fairness norms," *Journal of Consumer Psychology*, vol. 22, no. 1, SI, pp. 114–127, 2012.

[13] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *Proceedings of the 25th ACM KDD*, 2019, p. 2459–2468.

[14] M. K. Lee, A. Jain, H. J. Cha, S. Ojha, and D. Kusbit, "Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, pp. 1–26, November 2019.

[15] R. Wang, F. M. Harper, and H. Zhu, "Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences," in *Proceedings of CHI 2020*, 2020, p. 1–14.

[16] R. K. E. Bellamy and et al., "AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias," *arXiv:1810.01943 [cs]*, 2018.

[17] S. Corbett-Davies and S. Goel, "The measure and mismeasure of fairness: A critical review of fair machine learning," *arXiv preprint arXiv:1808.00023*, 2018.

[18] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of the... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, vol. 2018. NIH Public Access, 2018, p. 1931.

[19] B. Glymour and J. Herington, "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*, 2019, pp. 269–278.

[20] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems*, 2017, pp. 656–666.

[21] A. Narayanan, "Translation tutorial: 21 fairness definitions and their politics," in *ACM Conference on Fairness, Accountability, and Transparency*, 2 2018.

[22] J. Zhou, C. Bridon, F. Chen, A. Khawaji, and Y. Wang, "Be Informed and Be Involved: Effects of Uncertainty and Correlation on User Confidence in Decision Making," in *Proceedings of CHI2015 Works-in-Progress*, Korea, 2015.

[23] J. Zhou, J. Sun, F. Chen, Y. Wang, R. Taib, A. Khawaji, and Z. Li, "Measurable Decision Making with GSR and Pupillary Analysis for Intelligent User Interface," *ACM Transactions on Computer-Human Interaction*, vol. 21, no. 6, p. 33, 2015.

[24] R. F. Kizilcec, "How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface," in *Proceedings of CHI2016*, 2016, pp. 2390–2395.

[25] M. Bilgic and R. Mooney, "Explaining recommendations: Satisfaction vs. promotion," in *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at 2005 IUI*, 2005.

[26] P. Symeonidis, A. Nanopoulos, and Y. Manolopoulos, "Moviexplain: A recommender system with explanations," in *Proceedings of the Third ACM Conference on Recommender Systems*, 2009, pp. 317–320.

[27] J. Zhou, H. Hu, Z. Li, K. Yu, and F. Chen, "Physiological indicators for user trust in machine learning with influence enhanced fact-checking," in *Machine Learning and Knowledge Extraction*, 2019, pp. 94–113.

[28] T. C. Earle and M. Siegrist, "On the relation between trust and fairness in environmental risk management," *Risk Analysis*, vol. 28, no. 5, pp. 1395–1414, October 2008.

[29] D. Nikbin, I. Ismail, M. Marimuthu, and I. Abu-Jarad, "The effects of perceived service fairness on satisfaction, trust, and behavioural intentions," *Singapore Management Review*, vol. 33, no. 2, pp. 58–73, 2011.

[30] J. A. Colquitt and J. B. Rodell, "Measuring justice and fairness," in *The Oxford Handbook of Justice in the Workplace*, R. S. Cropanzano and M. L. Ambrose, Eds. Oxford University Press, 2015.

[31] S. M. Merritt, H. Heimbaugh, J. LaChapell, and D. Lee, "I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system," *Human Factors*, vol. 55, no. 3, pp. 520–534, 2013.