A Study on Fairness and Trust Perceptions in Automated Decision Making

Jakob Schoeffer[a], Yvette Machowski[a] and Niklas Kuehl[a]

[a] Karlsruhe Institute of Technology (KIT), Germany

Abstract

Automated decision systems are increasingly used for consequential decision making—for a variety of reasons. These systems often rely on sophisticated yet opaque models, which do not (or hardly) allow for understanding *how* or *why* a given decision was arrived at. This is not only problematic from a legal perspective, but non-transparent systems are also prone to yield undesirable (e.g., unfair) outcomes because their sanity is difficult to assess and calibrate in the first place. In this work, we conduct a study to evaluate different attempts of explaining such systems with respect to their effect on people's perceptions of fairness and trustworthiness towards the underlying mechanisms. A pilot study revealed surprising qualitative insights as well as preliminary significant effects, which will have to be verified, extended and thoroughly discussed in the larger main study.

Keywords

Automated Decision Making, Fairness, Trust, Transparency, Explanation, Machine Learning

1. Introduction

Automated decision making has become ubiquitous in many domains such as hiring [1], bank lending [2], grading [3], and policing [4], among others. As automated decision systems (ADS) are used to inform increasingly high-stakes consequential decisions, understanding their inner workings is of utmost importance—and undesirable behavior becomes a problem of societal relevance. The underlying motives of adopting ADS are manifold: They range from cost-cutting to improving performance and enabling more robust and objective decisions [1, 5]. One widespread assumption is that ADS can also avoid human biases in the decision making process [1]. However, ADS are typically based on artificial intelligence (AI) techniques, which, in turn, generally rely on historical data. If, for instance, this underlying data is biased (e.g., because certain socio-demographic groups were favored in a disproportional way in the past), an ADS may pick up and perpetuate existing patterns of unfairness [6]. Two prominent examples of such behavior from the recent past are the discrimination of black people in the realm of facial recognition [7] and recidivism prediction [8]. These and other cases have put ADS under enhanced scrutiny, jeopardizing trust in these systems.

In recent years, a significant body of research has been devoted to detecting and mitigating unfairness in automated decision making [6]. Yet, most of this work has focused on formalizing the concept of fairness and enforcing certain statistical equity constraints, often without explicitly taking into

Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA

jakob.schoeffer@kit.edu (J. Schoeffer); yvette.machowski@alumni.kit.edu (Y. Machowski); niklas.kuehl@kit.edu (N. Kuehl)
0000-0003-3705-7126 (J. Schoeffer); 0000-0002-9271-6342 (Y. Machowski); 0000-0001-6750-0876 (N. Kuehl)

account the perspective of individuals affected by such automated decisions. In addition to how researchers may define and enforce fairness in technical terms, we argue that it is vital to understand people's *perceptions* of fairness—vital not only from an ethical standpoint but also with respect to facilitating trust in and adoption of (appropriately deployed) socio-technical systems like ADS. Srivastava et al. [9], too, emphasize the need for research to gain a deeper understanding of people's attitudes towards fairness in ADS.

A separate, yet very related, issue revolves around how to *explain* automated decisions and the underlying processes to affected individuals so as to enable them to appropriately assess the quality and origins of such decisions. Srivastava et al. [9] also point out that subjects should be presented with more information about the workings of an algorithm and that research should evaluate how this additional information influences people's fairness perceptions. In fact, the EU General Data Protection Regulation (GDPR)[1], for instance, requires to disclose "the existence of automated decision-making, including [. . .] meaningful information about the logic involved [. . .]" to the "data subject". Beyond that, however, such regulations remain often vague and little actionable. To that end, we conduct a study to examine in more depth the effect of different explanations on people's perceptions of fairness and trustworthiness towards the underlying ADS in the context of lending, with a focus on

- the amount of information provided,
- the background and experience of people,
- the nature of the decision maker (human vs. automated).

2. Background and Related Work

It is widely understood that AI-based technology can have undesirable effects on humans. As a result, topics of fairness, accountability and transparency have become important areas of research in the fields of AI and human-computer interaction (HCI), among others. In this section, we provide an overview of relevant literature and highlight our contributions.

Explainable AI Despite being a popular topic of current research, explainable AI (XAI) is a natural consequence of designing ADS and, as such, has been around at least since the 1980s [15]. Its importance, however, keeps rising as increasingly sophisticated (and opaque) AI techniques are used to inform evermore consequential decisions. XAI is not only required by law (e.g., GDPR, ECOA[2]); Eslami et al. [16], for instance, have shown that users' attitudes towards algorithms change when transparency is increased. When sufficient information is not presented, users sometimes rely too heavily on system suggestions [17]. Yet, both quantity and quality of explanations matter: Kulesza et al. [18] explore the effects of soundness and completeness of explanations on end users' mental models and suggest, among others, that oversimplification is problematic. We refer to [15, 19, 20] for more in-depth literature on the topic of XAI.

Perceptions of fairness and trustworthiness A relatively new line of research in AI and HCI has started focusing on *perceptions* of fairness and trustworthiness in automated decision making. For instance, Binns

[1] https://eur-lex.europa.eu/eli/reg/2016/679/oj (last accessed Jan 3, 2021)

[2] Equal Credit Opportunity Act: https://www.consumer.ftc.gov/articles/0347-your-equal-credit-opportunity-rights (last accessed Jan 3, 2021)

Table 1
Overview of related work.

Reference	Explanation styles provided	Amount of provided information evaluated	Understandability tested	Computer / AI experience evaluated	Human involvement in context considered
Binns et al. [10]	distinct	no	single question	no	no
Dodge et al. [11]	distinct	no	not mentioned	no	no
Lee [12]	distinct	no	no	knowledge of algorithms	individual in management context
Lee and Baykal [13]	n/a due to study setup	no	no	programming / algorithm knowledge	group decision in fair division context
Wang et al. [14]	distinct	partly	no	computer literacy	algorithmic decision, reviewed by group in crowdsourcing context
Our work	distinct and combined	yes	construct with multiple items	AI literacy	individual in provider-customer context

et al. [10] and Dodge et al. [11] compare fairness perceptions in ADS for four distinct explanation styles. Lee [12] compares perceptions of fairness and trustworthiness depending on whether the decision maker is a person or an algorithm in the context of managerial decisions. Lee and Baykal [13] explore how algorithmic decisions are perceived in comparison to group-made decisions. Wang et al. [14] combine a number of manipulations, such as favorable and unfavorable outcomes, to gain an overview of fairness perceptions. An interesting finding by Lee et al. [21] suggests that fairness perceptions decline for some people when gaining an understanding of an algorithm if their personal fairness concepts differ from those of the algorithm. Regarding trustworthiness, Kizilcec [22], for instance, concludes that it is important to provide the right amount of transparency for optimal trust effects, as both too much and too little transparency can have undesirable effects.

Our contribution We aim to complement existing work to better understand *how much* of *which* information of an ADS should be provided to *whom* so that people are optimally enabled to understand the inner workings and appropriately assess the quality (e.g., fairness) and origins of such decisions. Specifically, our goal is to add novel insights in the following ways: First, our approach combines multiple explanation styles in one condition, thereby disclosing varying amounts of information. This differentiates our method from the concept of distinct individual explanations adopted by, for instance, Binns et al. [10]. We also evaluate the understandability of explanations through multiple items; and we add a novel analysis of the effect of people's AI literacy [23] on their perceptions of fairness and trustworthiness. Finally, we investigate whether perceptions of fairness and trustworthiness differ between having a human or an automated decision maker, controlling for the provided explanations. For brevity, we have summarized relevant aspects where our work can complement existing literature in Table 1.

3. Study Design and Methodology

With our study, we aim to contribute novel insights towards answering the following main questions:

Q1 Do people perceive a decision process to be fairer and/or more trustworthy if more information about it is disclosed?

Q2 Does people's experience / knowledge in the field of AI have an impact on their perceptions of fairness and trustworthiness towards automated decision making?

Q3 How do people perceive human versus automated (consequential) decision making with respect to fairness and trustworthiness?

We choose to explore the aforementioned relationships in the context of lending—an example of a provider-customer encounter. Specifically, we confront study participants with situations where a person was denied a loan. We choose a between-subjects design with the following conditions: First, we reveal that the loan decision was made by a human or an ADS (i.e., automated). Then we provide one of four explanation styles to each study participant. Figure 1 contains an illustration of our study setup, the elements of which will be explained in more detail shortly. Eventually, we measure four different constructs: *understandability* (of the given explanations), *procedural fairness* [24], *informational fairness* [24], and *trustworthiness* (of the decision maker); and we compare the results across conditions. Additionally, we measure *AI literacy* of the study participants. Please refer to Appendix A for a list of all constructs and associated measurement items for the case of automated decisions. Note that for each construct we measure *multiple* items.

Our analyses are based on a publicly available dataset on home loan application decisions[3], which has been used in multiple `Kaggle` competitions. Note that comparable data—reflecting a given finance company's individual circumstances and approval criteria—might in practice be used to train ADS. The dataset at hand consists of 614 labeled (loan Y/N) observations and includes the following features: *applicant income, co-applicant income, credit history, dependents, education, gender, loan amount, loan amount term, marital status, property area, self-employment.* After removing data points with missing values, we are left with 480 observations, 332 of which (69.2%) involve the positive label (Y) and 148 (30.8%) the negative label (N). We use 70% of the dataset for training purposes and the remaining 30% as a holdout set.

As groundwork, after encoding and scaling the features, we trained a random forest classifier with bootstrapping to predict the held-out labels, which yields an out-of-bag accuracy estimate of 80.1%. Our first explanation style, *(F)*, consists of disclosing the features including corresponding values for an observation (i.e., an applicant) from the holdout set whom our model denied the loan. We refer to such an observation as a *setting*. In our study, we employ different settings in order to ensure generalizability. Please refer to Appendix B for an excerpt of questionnaires for one exemplary setting (male applicant). Note that all explanations are derived from the data—they are *not* concocted. Next, we computed permutation feature importances [25] from our model and obtained

[3] https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset (last accessed Jan 3, 2021)

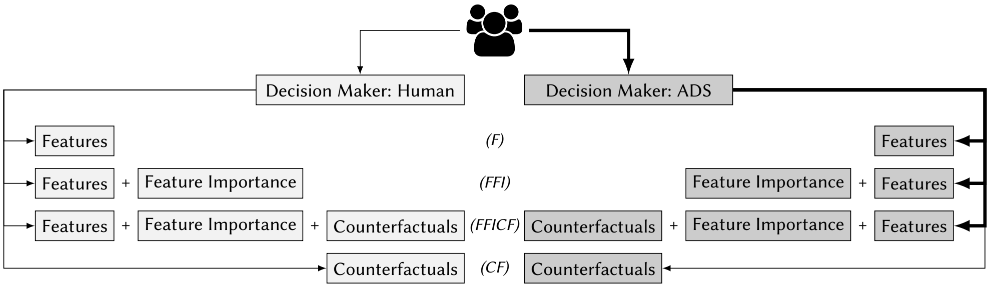

Figure 1: Graphical representation of our study setup. Thick lines indicate the subset of conditions from our pilot study.

the following hierarchy, using "≻" as a shorthand for "is more important than": *credit history* ≻ *loan amount* ≻ *applicant income* ≻ *co-applicant income* ≻ *property area* ≻ *marital status* ≻ *dependents* ≻ *education* ≻ *loan amount term* ≻ *self-employment* ≻ *gender*. Revealing this ordered list of feature importances in conjunction with *(F)* makes up our second explanation style *(FFI)*. To construct our third and fourth explanation styles, we conducted an online survey with 20 quantitative and qualitative researchers to ascertain which of the aforementioned features are actionable—in a sense that people can (hypothetically) act on them in order to increase their chances of being granted a loan. According to this survey, the top-5 actionable features are: *loan amount, loan amount term, property area, applicant income, co-applicant income.* Our third explanation style *(FFICF)* is then—in conjunction with *(F)* and *(FFI)*—the provision of three counterfactual scenarios where one actionable feature each is (minimally) altered such that our model predicts a loan approval instead of a rejection. The last explanation style is *(CF)*, without additionally providing features or feature importances. This condition aims at testing the effectiveness of counterfactual explanations in isolation, as opposed to providing them in conjunction with other explanation styles. We employ only model-agnostic explanations [20] in a way that they could plausibly be provided by both humans and ADS.

4. Preliminary Analyses and Findings

Based on Section 3, we conducted an online pilot study with 58 participants to infer preliminary insights regarding **Q1** and **Q2** and to validate our study design. Among the participants were 69% males, 29% females, and one person who did not disclose their gender; 53% were students, 28% employed full-time, 10% employed part-time, 3% self-employed, and 5% unemployed. The average age was 25.1 years, and 31% of participants have applied for a loan before. For this pilot study, we only included the ADS settings (right branch in Figure 1) and limited the conditions to *(F)*, *(FFI)*, and *(FFICF)*. The study participants were randomly assigned to one of the three conditions, and each participant was provided with two consecutive questionnaires associated with two different settings—one male and one female applicant. Participants for this online study were recruited from all over the world via `Prolific`[4] [26] and asked to rate their agreement with multiple state-

[4] https://www.prolific.co/

Table 2
Pearson correlations between constructs for pilot study.

Construct 1	**Construct 2**	**Pearson's r**
Procedural Fairness	Informational Fairness	0.47
Procedural Fairness	Trustworthiness	0.78
Procedural Fairness	Understandability	0.23
Informational Fairness	Trustworthiness	0.72
Informational Fairness	Understandability	0.69
Trustworthiness	Understandability	0.41

ments on 5-point Likert scales, where a score of 1 corresponds to "strongly disagree", and a score of 5 denotes "strongly agree". Additionally, we included multiple open-ended questions in the questionnaires to be able to carry out a qualitative analysis as well.

4.1. Quantitative Analysis

Constructs As mentioned earlier, we measured four different constructs: understandability (of the given explanations), procedural fairness [24], informational fairness [24], and trustworthiness (of the decision maker); see Appendix A for the associated measurement items. Note that study participants responded to the same (multiple) measurement items per construct, and these measurements were ultimately averaged to obtain one score per construct. We evaluated the reliability of the constructs through Cronbach's alpha—all values were larger than 0.8 thus showing good reliability for all constructs [27]. We proceeded to measure correlations between the four constructs with Pearson's r to obtain an overview of the relationships between our constructs. Table 2 provides an overview of these relationships: Procedural fairness and informational fairness are each strongly correlated with trustworthiness, and informational fairness is strongly correlated with understandability. Overall, we found significant correlations ($p < 0.05$) between all constructs besides procedural fairness and understandability.

Insights regarding Q1 We conducted multiple ANOVAs followed by Tukey's tests for post-hoc analysis to examine the effects of our three conditions. The individual scores for each construct and condition are provided in Table 3. We found a significant effect between different conditions on fairness perceptions for procedural fairness ($F(2, 55) = 3.56, p = 0.035$) as well as for informational fairness ($F(2, 55) = 10.90, p < 0.001$). Tukey's test for post-hoc analysis showed that the effect for procedural fairness was only significant between the conditions *(F)* and *(FFICF)* ($p = 0.040$). When controlling for different variables, such as study participants' gender, the effect for procedural fairness is reduced to marginal significance ($p > 0.05$). For informational fairness the effect in the post-hoc analysis without control variables is significant between *(F)* and *(FFICF)* ($p < 0.001$) as well as between *(FFI)* and *(FFICF)* ($p = 0.042$), and it is marginally significant between *(F)* and *(FFI)* ($p = 0.072$). Controlling for study participants' gender reduces the significance between *(FFI)* and *(FFICF)* to marginal significance ($p = 0.059$); controlling for study par-

Table 3
Construct scores by condition for pilot study. The scores, ranging from 1 (low) to 5 (high), were obtained by averaging across all measurement items for each construct.

Construct	*(F)*	*(FFI)*	*(FFICF)*
Understandability	3.17	3.87	4.12
Procedural Fairness	3.28	3.40	3.91
Informational Fairness	2.79	3.33	3.92
Trustworthiness	2.92	3.39	3.83

ticipants' age removes the significance between these two conditions altogether.

Interestingly, significant effects on understandability between conditions ($F(2, 55) = 7.52, p = 0.001$) came from *(F)* and *(FFICF)* ($p = 0.001$) as well as *(F)* and *(FFI)* ($p = 0.020$). Significant effects of the conditions on trustworthiness ($F(2, 55) = 4.94, p = 0.011$) could only be observed between *(F)* and *(FFICF)* ($p = 0.007$). In general, we urge to exercise utmost caution when interpreting the quantitative results of our pilot study as the sample size is extremely small. We hope to generate more reliable and extensive insights with our main study and a much larger number of participants.

Insights regarding Q2 We calculated Pearson's r between each of our fairness measures including trustworthiness and the study participants' AI literacy. All three measures, procedural fairness ($r = 0.35, p = 0.006$), informational fairness ($r = 0.52, p < 0.001$) and trustworthiness ($r = 0.48, p < 0.001$) demonstrate a significant positive correlation with AI literacy. Therefore, within the scope of our pilot study, we found that participants with more knowledge and experience in the field of AI tend to perceive the decision making process and the provided explanations of the ADS at hand to be fairer and more trustworthy than participants with less knowledge and experience in this field.

4.2. Qualitative Analysis

In the following, we provide a summary of insightful responses to open-ended questions from our questionnaires.

Regarding automated decision making Perhaps surprisingly, many participants approved of the ADS as the decision maker. They perceived the decision to be less biased and argued that all applicants are treated equally, because the ADS makes its choices based on facts, not based on the likeability of a person: "*I think that an automated system treats every individual fairly because everybody is judged according to the same rules.*" Some participants directly compared the ADS to human decision makers: "*I think that [the decision making procedures] are fair because they are objective, since they are automated. Humans usually [can't] make decisions without bias.*" Other participants responded with a (somewhat expected) disapproval towards the ADS. Participants criticized, for instance, that the decisions "*are missing humanity in them*" and how an automated decision based "*only on statistics without human morality and ethics*" simply cannot be fair. One participant went so far as to formulate positive arguments for human bias in decision making procedures: "*I do not* *believe that it is fair to assess anything that greatly affects an individual's life or [livelihood] through an automated decision system. I believe some bias and personal opinion is often necessary to uphold ethical and moral standards.*" Finally, some participants had mixed feelings because they saw the trade-off between a "*cold approach*" that lacks empathy and a solution that promotes "*equality with others*" because it "*eliminates personal bias*".

Regarding explanations Study participants had strong opinions on the features considered in the loan decision. Most participants found *gender* to be the most inappropriate feature. The comments on this feature ranged from "*I think the gender of the person shouldn't matter*" to considering gender as a factor being "*ethically wrong*" or even "*borderline illegal*". *Education* and *property area* were named by many participants as being inappropriate factors as well: "*I think education, gender, property area [. . .] are inappropriate factors and should not be considered in the decision making process.*" On average, the order of feature importance was rated as equally appropriate as the features themselves. Some participants assessed the order of feature importance in general and came to the conclusion that it is appropriate: "*The most important is credit history in this decision and least gender so the order is appropriate.*" At the same time, a few participants rated the order of feature importance as inappropriate, for instance because "*some things are irrelevant yet score higher than loan term.*" In the first of two settings, the counterfactual for *property area* was received negatively by some: "*It shouldn't matter where the property is located.*" Yet, most participants found the counterfactual explanations in the second setting to be appropriate: "*The three scenarios represent plausible changes the individual could perform [. . .]*"

5. Outlook

The potential of automated decision making and its benefits over purely human-made decisions are obvious. However, several instances are known where such automated decision systems (ADS) are having undesirable effects—especially with respect to fairness and transparency. With this work, we aim to contribute novel insights to better understand people's perceptions of fairness and trustworthiness towards ADS, based on the provision of varying degrees of information about such systems and their underlying processes. Moreover, we examine how these perceptions are influenced by people's background and experience in the field of artificial intelligence. As a first step, we have conducted an online pilot study and obtained preliminary results for a subset of conditions. Next, we will initiate our main study with a larger sample size and additional analyses. For instance, we will also explore whether people's perceptions of fairness and trustworthiness change when the decision maker is claimed to be human (as opposed to purely automated). We hope that our contribution will ultimately help in designing more equitable decision systems as well as stimulate future research on this important topic.

References

[1] N. R. Kuncel, D. S. Ones, D. M. Klieger, In hiring, algorithms beat instinct, Harvard Business Review (2014). URL: https://hbr.org/2014/05/in-hiring-algorithms-beat-instinct.

[2] S. Townson, AI can make bank loans more fair, Harvard Business Review (2020). URL: https://hbr.org/2020/11/ai-can-make-bank-loans-more-fair.

[3] A. Satariano, British grading debacle shows pitfalls of automating govern-

ment, The New York Times (2020). URL: https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html.

[4] W. D. Heaven, Predictive policing algorithms are racist. They need to be dismantled, MIT Technology Review (2020). URL: https://www.technologyreview.com/2020/07/17/1005396/predictive-policing-algorithms-racist-dismantled-machine-learning-bias-criminal-justice/.

[5] J. G. Harris, T. H. Davenport, Automated decision making comes of age, MIT Sloan Management Review (2005). URL: https://sloanreview.mit.edu/article/automated-decision-making-comes-of-age/.

[6] S. Barocas, M. Hardt, A. Narayanan, Fairness and machine learning, 2019. URL: http://www.fairmlbook.org.

[7] J. Buolamwini, T. Gebru, Gender shades: Intersectional accuracy disparities in commercial gender classification, in: Conference on Fairness, Accountability and Transparency, 2018, pp. 77–91.

[8] J. Angwin, J. Larson, S. Mattu, L. Kirchner, Machine bias, ProPublica (2016). URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[9] M. Srivastava, H. Heidari, A. Krause, Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2459–2468.

[10] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, N. Shadbolt, 'It's reducing a human being to a percentage'; perceptions of justice in algorithmic decisions, in: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, 2018, pp. 1–14.

[11] J. Dodge, Q. V. Liao, Y. Zhang, R. K. Bellamy, C. Dugan, Explaining models: An empirical study of how explanations impact fairness judgment, in: Proceedings of the 24th International Conference on Intelligent User Interfaces, 2019, pp. 275–285.

[12] M. K. Lee, Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management, Big Data & Society 5 (2018) 1–16.

[13] M. K. Lee, S. Baykal, Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division, in: Proceedings of the 2017 ACM Conference on Computer-Supported Cooperative Work and Social Computing, 2017, pp. 1035–1048.

[14] R. Wang, F. M. Harper, H. Zhu, Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–14.

[15] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, S. Stumpf, P. Kieseberg, A. Holzinger, Explainable AI: The new 42?, in: International Cross-Domain Conference for Machine Learning and Knowledge Extraction, 2018, pp. 295–303.

[16] M. Eslami, K. Vaccaro, M. K. Lee, A. Elazari Bar On, E. Gilbert, K. Karahalios, User attitudes towards algorithmic opacity and transparency in online reviewing platforms, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[17] A. Bussone, S. Stumpf, D. O'Sullivan, The role of explanations on trust and reliance in clinical decision support systems, in: IEEE International Conference on Healthcare Informatics, 2015, pp. 160–169.

[18] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? Ways explanations impact end users' mental models, in: 2013 IEEE Symposium on Visual Languages and Human-Centric Computing, 2013, pp. 3–10.

[19] C. Molnar, Interpretable machine learning, 2020. URL: https://christophm.github.io/interpretable-ml-book/.

[20] A. Adadi, M. Berrada, Peeking inside the black-box: A survey on explainable artificial intelligence (XAI), IEEE Access 6 (2018) 52138–52160.

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

[22] R. F. Kizilcec, How much information? Effects of transparency on trust in an algorithmic interface, in: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, 2016, pp. 2390–2395.

[23] D. Long, B. Magerko, What is AI literacy? Competencies and design considerations, in: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, 2020, pp. 1–16.

[24] J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. Porter, K. Y. Ng, Justice at the millennium: A meta-analytic review of 25 years of organizational justice research, Journal of Applied Psychology 86 (2001) 425–445.

[25] L. Breiman, Random forests, Machine Learning 45 (2001) 5–32.

[26] S. Palan, C. Schitter, Prolific.ac—a subject pool for online experiments, Journal of Behavioral and Experimental Finance 17 (2018) 22–27.

[27] J. M. Cortina, What is coefficient alpha? An examination of theory and applications, Journal of Applied Psychology 78 (1993) 98–104.

[28] V. McKinney, K. Yoon, F. M. Zahedi, The measurement of web-customer satisfaction: An expectation and disconfirmation approach, Information Systems Research 13 (2002) 296–315.

[29] J. A. Colquitt, J. B. Rodell, Measuring justice and fairness, in: R. S. Cropanzano, M. L. Ambrose (Eds.), The Oxford Handbook of Justice in the Workplace, Oxford University Press, 2015, pp. 187–202.

[30] C.-M. Chiu, H.-Y. Lin, S.-Y. Sun, M.-H. Hsu, Understanding customers' loyalty intentions towards online shopping: An integration of technology acceptance model and fairness theory, Behaviour & Information Technology 28 (2009) 347–360.

[31] L. Carter, F. Bélanger, The utilization of e-government services: Citizen trust, innovation and acceptance factors, Information Systems Journal 15 (2005) 5–25.

[32] A. Wilkinson, J. Roberts, A. E. While, Construction of an instrument to measure student information and communication technology skills, experience and attitudes to e-learning, Computers in Human Behavior 26 (2010) 1369–1376.

A. Constructs and Items for Automated Decisions

All items within the following constructs were measured on a 5-point Likert scale and mostly drawn (and adapted) from previous studies.

1. **Understandability**
   Please rate your agreement with the following statements:
   - The explanations provided by the automated decision system are clear in meaning. [28]
   - The explanations provided by the automated decision system are easy to comprehend. [28]
   - In general, the explanations provided by the automated decision system are understandable for me. [28]
2. **Procedural Fairness**
   The statements below refer to the *procedures* the automated decision system uses to make decisions about loan applications. Please rate your agreement with the following statements:
   - Those procedures are free of bias. [29]
   - Those procedures uphold ethical and moral standards. [29]
   - Those procedures are fair.
   - Those procedures ensure that decisions are based on facts, not personal biases and opinions. [29]
   - Overall, the applying individual is treated fairly by the automated decision system. [29]
3. **Informational Fairness**
   The statements below refer to the *explanations* the automated decision system offers with respect to the decision-making procedures. Please rate your agreement with the following statements:
   - The automated decision system explains decision-making procedures thoroughly. [29]
   - The automated decision system's explanations regarding procedures are reasonable. [29]
   - The automated decision system tailors communications to meet the applying individual's needs. [29]
   - I understand the process by which the decision was made. [10]
   - I received sufficient information to judge whether the decision-making procedures are fair or unfair.
4. **Trustworthiness**
   The statements below refer to the *automated decision system.* Please rate your agreement with the following statements:

- Given the provided explanations, I trust that the automated decision system makes good-quality decisions. [12]
- Based on my understanding of the decision-making procedures, I know the automated decision system is not opportunistic. [30]
- Based on my understanding of the decision-making procedures, I know the automated decision system is trustworthy. [30]
- I think I can trust the automated decision system. [31]
- The automated decision system can be trusted to carry out the loan application decision faithfully. [31]
- In my opinion, the automated decision system is trustworthy. [31]

5. **AI Literacy**

   - How would you describe your knowledge in the field of artificial intelligence?
   - Does your current employment include working with artificial intelligence?

   Please rate your agreement with the following statements:

   - I am confident interacting with artificial intelligence. [32]
   - I understand what the term *artificial intelligence* means.

B. Explanation Styles for Automated Decisions and One Exemplary Setting (Male Applicant)

Explanation Style *(F)*

A finance company offers loans on real estate in urban, semi-urban and rural areas. A potential customer first applies online for a specific loan, and afterwards the company assesses the customer's eligibility for that loan.
An individual applied online for a loan at this company. The company denied the loan application. The decision to deny the loan was made by an automated decision system and communicated to the applying individual electronically and in a timely fashion.

The automated decision system explains that the following factors (in alphabetical order) on the individual were taken into account when making the loan application decision:

- Applicant Income: $3,069 per month
- Co-Applicant Income: $0 per month
- Credit History: Good

- Dependents: 0
- Education: Graduate
- Gender: Male
- Loan Amount: $71,000
- Loan Amount Term: 480 months
- Married: No
- Property Area: Urban
- Self-Employed: No

Explanation Style *(FFI)*

A finance company offers loans on real estate in urban, semi-urban and rural areas. A potential customer first applies online for a specific loan, and afterwards the company assesses the customer's eligibility for that loan.
An individual applied online for a loan at this company. The company denied the loan application. The decision to deny the loan was made by an automated decision system and communicated to the applying individual electronically and in a timely fashion.

The automated decision system explains …

- …that the following factors (in alphabetical order) on the individual were taken into account when making the loan application decision:
  - Applicant Income: $3,069 per month
  - Co-Applicant Income: $0 per month
  - Credit History: Good
  - Dependents: 0
  - Education: Graduate
  - Gender: Male
  - Loan Amount: $71,000
  - Loan Amount Term: 480 months
  - Married: No
  - Property Area: Urban

  - Self-Employed: No

- …that different factors are of different importance in the decision. The following list shows the order of factor importance, from most important to least important: Credit History > Loan Amount > Applicant Income > Co-Applicant Income > Property Area > Married > Dependents > Education > Loan Amount Term > Self-Employed > Gender

Explanation Style *(FFICF)*

> A finance company offers loans on real estate in urban, semi-urban and rural areas. A potential customer first applies online for a specific loan, and afterwards the company assesses the customer's eligibility for that loan.
> An individual applied online for a loan at this company. The company denied the loan application. The decision to deny the loan was made by an automated decision system and communicated to the applying individual electronically and in a timely fashion.

The automated decision system explains …

- …that the following factors (in alphabetical order) on the individual were taken into account when making the loan application decision:
  - Applicant Income: $3,069 per month
  - Co-Applicant Income: $0 per month
  - Credit History: Good
  - Dependents: 0
  - Education: Graduate
  - Gender: Male
  - Loan Amount: $71,000
  - Loan Amount Term: 480 months
  - Married: No
  - Property Area: Urban
  - Self-Employed: No
- …that different factors are of different importance in the decision. The following list shows the order of factor importance, from most important to least important: Credit History > Loan Amount > Applicant Income > Co-Applicant Income > Property Area > Married > Dependents > Education > Loan Amount Term > Self-Employed > Gender

- …that the individual would have been granted the loan if—everything else unchanged—one of the following hypothetical scenarios had been true:
  - The Co-Applicant Income had been at least \$800 per month
  - The Loan Amount Term had been 408 months or less
  - The Property Area had been Rural