Perceptions of Fairness and Trustworthiness Based on Explanations in Human vs. Automated Decision-Making

Jakob Schoeffer
Karlsruhe Institute of Technology
jakob.schoeffer@kit.edu

Yvette Machowski
Karlsruhe Institute of Technology
yvette.machowski@alumni.kit.edu

Niklas Kuehl
Karlsruhe Institute of Technology
niklas.kuehl@kit.edu

Abstract

Automated decision systems (ADS) have become ubiquitous in many high-stakes domains. Those systems typically involve sophisticated yet opaque artificial intelligence (AI) techniques that seldom allow for full comprehension of their inner workings, particularly for affected individuals. As a result, ADS are prone to deficient oversight and calibration, which can lead to undesirable (e.g., unfair) outcomes. In this work, we conduct an online study with 200 participants to examine people's perceptions of fairness and trustworthiness towards ADS in comparison to a scenario where a human instead of an ADS makes a high-stakes decision—and we provide thorough identical explanations regarding decisions in both cases. Surprisingly, we find that people perceive ADS as fairer than human decision-makers. Our analyses also suggest that people's AI literacy affects their perceptions, indicating that people with higher AI literacy favor ADS more strongly over human decision-makers, whereas low-AI-literacy people exhibit no significant differences in their perceptions.

1. Introduction

Automated decision-making has been increasingly adopted in areas such as hiring [1], lending [2], or even policing [3]. As the underlying systems, often referred to as *automated decision systems* (ADS), are informing evermore high-stakes decisions, it is of utmost importance to understand their inner workings—particularly for individuals affected by their decisions, because any malfunction (e.g., unfair decisions) will have staggering consequences for them.

The reasons for adopting ADS are manifold [1, 4]. In fact, if properly designed and deployed, they are a valuable tool to combat stereotyping and thus contribute to overall social equity, e.g., in the fields of recruitment [5, 6], health care [7, 8], or financial inclusion [9]. That said, ADS are typically based on artificial intelligence (AI)—specifically, machine learning (ML)—techniques, which leverage historical data to inform future decisions. Now, if historical data is biased (e.g., because certain socio-demographic groups were systematically disfavored), an ADS will likely pick up and perpetuate existing patterns of unfairness [10]. A significant body of research on algorithmic fairness from the recent past has identified such instances where ADS are causing harmful outcomes, e.g., in job ad delivery [11], facial recognition [12], recidivism prediction [13], or grading [14]. These instances, among others, have likely been contributing to a recent general decline in trust towards AI [15].

In recent years, an extensive body of research has been devoted to detecting and mitigating unfairness in ADS—mainly from a computer science viewpoint [16]. A significant part of this work, however, has focused on formalizing the concept of fairness and modifying ML algorithms to satisfy different statistical equity constraints, without considering the feedback of individuals affected by automated decisions. Among others, Srivastava et al. [17] emphasize this need for better understanding people's attitudes towards fairness of ADS. We argue that this research gap creates the potential for a plethora of high-impact contributions from information systems (IS) research around people's perceptions of fairness and trustworthiness towards ADS. This work is vital not only from a moral perspective but also regarding the effective design and implementation of ADS—with the end goal of creating decision systems that are fair, trustworthy, and, as a result, suitable for wide adoption. To that end, we conduct a study to better understand people's perceptions of fairness and trustworthiness towards ADS in comparison to the (hypothetical) scenario where a human instead of an ADS makes the decision. We furthermore analyze how these perceptions may change depending on people's background and experience with AI.

Another issue of ADS revolves around explaining automated decisions to affected individuals. It is widely

URI: https://hdl.handle.net/10125/79466
978-0-9981331-5-7

understood that opaque (i.e., black box) ML models do not allow for meaningful interpretations as to how or why certain outcomes were arrived at [18, 19]. Prior research has also shown that explanations can be an effective tool for more transparent decision-making [20, 21]. Therefore, in this work, we provide study participants with thorough explanations regarding decisions—identical for both the case of the ADS and the human decision-maker. The context of our study is lending, which is a common high-stakes application of ADS [22].

2. Background and related work

Harris and Davenport [4] define *automated decision systems* (ADS) as systems that aim to minimize human involvement in decision-making processes. In many cases, ADS have the potential to make more consistent decisions than humans [5, 6, 7, 8, 9]. Such systems are popular in several industries, such as banking [2, 4] or hiring [1, 5, 6, 23]—and they are emerging in new areas as well, e.g., in health care [7, 8]. With their increasing adoption in different high-stakes areas, it is vital to ensure that ADS reach fair and transparent decisions. However, there have been multiple cases in the recent past where algorithms made discriminatory decisions, e.g., based on people's gender or race [3, 12, 13]. Additionally, the underlying ML models are increasingly considered black boxes, making their interpretation challenging [18].

2.1. Explainable AI

Despite being a popular topic of current research, *explainable AI* (XAI) is a natural consequence of designing AI-based ADS and, as such, has been around at least since the 1980s [24]. Its importance, however, keeps rising as increasingly sophisticated (and opaque) AI techniques are used to inform evermore high-stakes decisions.

Explanations can be distinguished along different dimensions. Adadi and Berrada [21], e.g., differentiate between model-specific and model-agnostic explanations. *Model-agnostic explanations* refer to methods that are not bound to a single type of ML model and are therefore more generalizable—which is why we employ them in this work. Examples of the model-agnostic (example-based) explanation style, which provide information people can potentially act upon, are counterfactual explanations [25]. In brief, counterfactual explanations provide people with information regarding the minimum changes that would lead to an alternative (generally the desirable) decision. Meske et al. [20], among others, discuss different types of explanations relevant to the IS community, particularly model-agnostic explanations. They argue that explainability is essential for evaluating automated systems. People affected by an automated decision may be particularly interested in explanations to assess the fairness or trustworthiness of the associated ADS. Other popular model-agnostic explanation styles include the provision of the relevant features used by an ML model or (permutation) feature importance [26]—both of which we employ in this work since they could be plausibly provided by both human and automated decision-makers (i.e., ADS). We refer to, e.g., Adadi and Berrada [21] or Goebel et al. [27] for more in-depth literature on the topic of XAI.

2.2. Perceptions of fairness and trustworthiness regarding ADS

A relatively new line of research, primarily in AI and human-computer interaction (HCI), has started focusing on perceptions of fairness and trustworthiness in automated decision-making. Binns et al. [28] and Dodge et al. [29], e.g., compare fairness perceptions in ADS for distinct explanation styles. Their works suggest differences in effectiveness of individual explanation styles—however, they also note that there does not seem to be a single best approach to explaining automated decisions. Lee [30] compares perceptions of fairness and trustworthiness depending on whether the decision-maker is a person or an algorithm in the context of algorithmic management, involving tasks like work scheduling or evaluation. Their findings suggest that, among others, people perceive automated decisions as less fair and trustworthy for tasks that require typical human skills. Lee and Baykal [31] explore how algorithmic decisions are perceived in comparison to group-made decisions. An interesting finding by Lee et al. [32] 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 [33], e.g., concludes that it is essential to provide the right amount of transparency for optimal trust effects. We generally believe that this line of research would benefit significantly from novel contribution of IS scholars.

2.3. Human vs. automated decisions

People encounter algorithms and automation in different settings. Therefore, it is essential to understand how this automation makes people feel and to infer the social force of algorithms. While engineers tend to show optimism in the ability of ADS to trace and

mitigate human biases and stereotypes, laypeople are often worried about AI taking over [34]. However, Castelo et al. [35] found that in case of perceived objective decisions, people favor automated advice, and for subjective decisions they prefer human advice. This is similar to the findings by Lee [30]. Yet, according to the study by Castelo et al. [35], the perceived objectivity of a task can be altered; thus trustworthiness of and reliance on an automated decision can be increased. Perhaps less surprisingly, Kramer et al. [36] found that people's preference for human or AI-based decisions also depends on their prior experiences with ADS.

A major issue with ADS is that people are often unaware of their existence. Eslami et al. [37], e.g., uncovered people's ignorance towards the algorithm behind Facebook's news feed. More than half of the participants in their study were unaware of the algorithm's manipulations, and some responded with anger and dissatisfaction. This unawareness—apart from negative experiences [12, 14], among others—might be part of the reason why many people have such a profound aversion against algorithms [15, 38]. Increasing people's awareness of ADS, e.g., by proactively disclosing the nature of the decision-maker, and considering their perceptions of these systems, may help raise acceptance in situations where ADS can make better (e.g., fairer) decisions than humans.

2.4. Our contribution

We aim to complement prior research (e.g., [28, 29, 30, 31, 32, 33, 35, 36]) to better understand people's perceptions of fairness and trustworthiness towards ADS vs. human decision-makers in high-stakes settings. Specifically, our goal is to add novel insights in the following ways: First, we integrate different model-agnostic explanations and provide them to study participants to enable them to assess the decision-making procedures. This contrasts with most existing work, which have typically employed distinct individual explanation styles only. Second, we provide identical model-agnostic explanations to study participants for both the case of ADS and the human decision-maker to not bias the collected responses. Third, we examine how perceptions may change for people with high vs. low AI literacy [39]. To the best of our knowledge, the combination of the previous aspects has not been examined before. Fourth, we consider the provider-customer context of lending, which differentiates our work from, e.g., Lee [30], who has analyzed the perceptions of human vs. automated decisions in algorithmic management. Finally, we aim to bring (back) to the IS community pressing relevant issues of societal relevance, which have experienced seminal contributions mostly from other communities, such as computer science and HCI.

3. Research hypotheses

Drawing on Chan [40], *informational fairness* is about "people's expectation that they should receive adequate information on and explanation of the process and its outcomes." In accordance with Bélanger et al. [41], we define *trustworthiness* as the perception of confidence in the reliability and integrity of the ADS. People often tend to avoid algorithms and prefer a human decision-maker over an automated one, even in situations where the algorithm outperforms the person. This phenomenon is called *algorithm aversion* [38]. Based on this theory, as well as recent developments regarding a decline in trust towards AI [15], we formulate our first two hypotheses, which conjecture higher perceptions of informational fairness and trustworthiness towards human decision-makers as compared to ADS:

H1 People's perceptions of informational fairness are higher when they are told the decision-maker is a human as compared to an ADS.

H2 People's perceptions of trustworthiness are higher when they are told the decision-maker is a human as compared to an ADS.

Experts of a certain type of decision procedure may have different attitudes towards a decision that touches on their area of expertise than laypeople. Wang et al. [42], e.g., found a significant effect of general computer literacy on fairness evaluations in automated decision-making. In the work at hand, we measure a construct that applies more directly to our context: We measure people's *AI literacy*, i.e., their "set of competencies that enables individuals to critically evaluate AI technologies; communicate and collaborate effectively with AI; and use AI as a tool online, at home, and in the workplace" [39]. We are interested in whether differences in people's AI literacy change their perceptions of informational fairness and trustworthiness towards human vs. automated decision-makers. Thus, we formulate the following additional hypotheses:

H3 People's AI literacy moderates the effect of the nature of the decision-maker (human vs. ADS) on people's perceptions of informational fairness.

H4 People's AI literacy moderates the effect of the nature of the decision-maker (human vs. ADS) on people's perceptions of trustworthiness.

4. Methodology

We evaluate our hypotheses 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 argue that this is a common context that affects many people at some point in life. According to, e.g., Atico [22], this is also an area where ADS are commonly employed for high-stakes decision-making.

4.1. Study design

Overall setup We choose a between-subject design with the following conditions: First, we reveal to study participants some basic information about the lending company—similarly to the study setup introduced in our earlier work [43]. We then explain that the company rejected a given individual's loan application. Afterwards, we randomly allocate study participants to one of two conditions: 50% of participants are provided the information that an ADS made the decision, and the other 50% are told that the decision-maker was a human being. We then provide identical explanations regarding a decision to study participants in either condition, the exact specifications of which will be derived and explained in more detail shortly. Finally, we measure perceptions of informational fairness (INFF) and trustworthiness (TRST) through multiple measurement items, drawn (and partially adapted) from previous studies (INFF: Colquitt et al. [44]; TRST: Carter and Bélanger [45], Chiu et al. [46], Lee [30]). Additionally, we measure AI literacy (AILIT) of study participants, with items partially derived from Long and Magerko [39] as well as Wilkinson et al. [47].

Data and ADS We design and implement a functional ADS for our study—similarly to earlier work by the authors [43]. The ADS consists of an ML model that predicts loan approval on unseen data and can output different explanations. For training our model, we utilize a publicly available dataset[1] on home loan application decisions, which has been used in multiple data science competitions on the platform `Kaggle`.[2] The dataset at hand consists of 614 labeled (loan Y/N) observations. It 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*. Note that comparable data—reflecting a given finance company's circumstances and approval criteria—might, in practice, be used to train ADS [48]. 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). As it is common in ML-based applications, we use 70% of the dataset to train our ADS and use the remaining 30% as a holdout set for the experiment. As groundwork for the design of our ADS, after encoding and scaling the features, we train a random forest classifier [26]. The classifier is then able to predict the (unseen) held-out labels—which it achieves with an out-of-bag accuracy of 80.1%. We use this classifier as a basis for the scenarios and explanations that participants are confronted with.

Explanations Recall that 50% of study participants are assigned the *ADS* condition and 50% the *human* condition. Both conditions are provided with identical explanations regarding the decisions—the only difference is that study participants in the ADS condition are told that the ADS provides the explanatory information. In contrast, participants in the human condition are told that a company representative (i.e., a human) provides this information.

We now explain in more detail the provided explanations. As noted earlier, we employ only model-agnostic explanations [21] in a way that they could plausibly be provided by humans and ADS alike. First, we disclose all *features* (applicant income, co-applicant income, etc., as mentioned earlier), including corresponding values (e.g., *applicant income: $3,069 per month*) for an observation (i.e., an applicant) from the holdout set whom our ADS denied the loan. We refer to such an observation as a *setting*. In our study, we employ different settings to ensure generalizability.

We also explain to study participants the *importance* of these features in the decision-making process. For that, we compute permutation feature importances [26] from our model and obtain the following hierarchy, ordered from most to least important: *credit history* $>$ *loan amount* $>$ *applicant income* $>$ *co-applicant income* $>$ *property area* $>$ *marital status* $>$ *dependents* $>$ *education* $>$ *loan amount term* $>$ *self-employment* $>$ *gender*. Note that feature importance is a global explanation style, meaning that this ordered list will be identical for any setting (i.e., applicant).

For each setting, we finally provide three *counterfactual* scenarios where one actionable feature each is minimally altered such that our model predicts a loan approval instead of a rejection (e.g., *the individual would have been granted the loan if, everything else*

[1] `https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset` (last accessed: August 24, 2021)

[2] `Kaggle` is the world's largest data science community (`https://www.kaggle.com/`)

unchanged, the co-applicant income had been at least $800 per month). To ascertain which of the features are actionable—in a sense that people can (hypothetically) act on them to increase their chances of being granted a loan—we conducted an online survey with 20 quantitative and qualitative researchers. According to this survey, the top-5 actionable features are *loan amount, loan amount term, property area, applicant income, co-applicant income*. We finally provide counterfactual explanations for a random subset of three of these features per setting.

4.2. Data collection

We conducted a between-subjects online study to test our hypotheses. Participants for this study were recruited via `Prolific`[3] [49] and randomly assigned to either the human decision scenario or the ADS decision scenario. Every participant was provided with two questionnaires associated with two different settings. In each questionnaire, we asked participants to rate their agreement with multiple statements per construct on 5-point Likert scales [50]. A score of 1 corresponds to "strongly disagree" and a score of 5 to "strongly agree". To be able to understand participants' quantitative responses better, we included multiple open-ended questions as well. We had to eliminate 4 of the 200 collected responses due to failure to pass an attention check—therefore, we analyzed 196 responses. Among our participants, 62% were male, 36% female, and the remaining 2% referred to themselves as non-binary or did not disclose their gender at all; 42% were students, 29% employed full-time, 11% employed part-time, 7% self-employed, 10% unemployed, and 1% chose not to disclose their profession. The average age of participants was 26.4 years.

5. Quantitative and qualitative results

Before conducting our tests, we assess the validity and reliability of our latent constructs (INFF, TRST, AILIT), each of which is measured through multiple items. We note that average variance extracted (AVE) is above or equal to the recommended threshold of 0.5 for INFF and TRST, while the AVE of AILIT is 0.39. According to Fornell and Larcker [51], if the AVE value of a construct is low, its convergent validity can still be sufficient if composite reliability (CR) is above 0.6. The CR of all our three constructs, INFF (0.83), TRST (0.94), and AILIT (0.72) is, in fact, above the threshold of 0.7, which is recommended by Barclay et al. [52]. Therefore, our convergent validity is sufficient for AILIT as well. Values for Cronbach's alpha (CA) are larger than the recommended threshold of 0.7 for our three constructs, indicating good reliability [53]. Validity and reliability measures are summarized in Table 1.

5.1. Comparison of perceptions

We conduct two Mann-Whitney U tests [54] to examine the differences in perceptions between ADS and human decision-makers. The Mann-Whitney U test for informational fairness is statistically significant ($p = 0.017$), suggesting a significant difference between participants' perceptions of informational fairness. Comparing the means of perceptions of informational fairness for both conditions reveals that the ADS condition ($M = 3.68$) is perceived to be significantly fairer than the human condition ($M = 3.47$). For perceptions of trustworthiness, however, there is no significant difference between the conditions ($p = 0.113$). Hence, neither **H1** nor **H2** are supported by our analyses. In fact, **H1** is reversely supported, eventually suggesting that for our study setup, informational fairness perceptions tend to be *higher towards the ADS* compared to the human decision-maker. Based on qualitative responses from study participants, we conjecture that this might be due to the perceived absence of emotions and subjectivity in automation. Other potential reasons for this based on qualitative feedback are given in Section 5.2. Note that this finding seems contradictory to some prior works' results (e.g., Castelo et al. [35]), which raises doubts about the generalizability of such findings beyond specific domains.

Interestingly, when considering people's AI literacy, these results change. For this analysis, we split our data into two (approximately equal-sized) sub-samples along the median value of AI literacy. We refer to one sample as *high AI literacy* participants and the other as *low AI literacy* participants. We then conduct separate Mann-Whitney U tests for the two sub-samples. Participants with high AI literacy perceive the ADS as significantly more informationally fair ($p = 0.021$) and more trustworthy ($p = 0.042$) than the human decision-maker. For participants with low AI literacy, we do not find a significant difference for perceptions of informational fairness ($p = 0.312$) or trustworthiness ($p = 0.995$) between the human and the ADS condition. Hence, we conclude that AI literacy has a moderating effect, which supports **H3** and **H4**. As stated in Section 3, we expected the moderating effect of AI

[3] `Prolific` is an online platform for recruiting high-quality research participants.

Table 1. Correlations and measurement information for latent factors.

Factor	M	SD	CA	CR	AVE	INFF	TRST	AILIT
INFF	3.57	0.62	0.83	0.83	0.50	1.00		
TRST	3.45	0.72	0.94	0.94	0.72	0.69	1.00	
AILIT	2.87	0.61	0.71	0.72	0.39	0.30	0.27	1.00

Notes: M = Mean; SD = Standard deviation

literacy. However, the finding that people with high AI literacy tend to perceive ADS as both fairer and more trustworthy than human decision-makers is not obvious to us. On the one hand, we might think that people with high AI literacy understand such systems better and are thus less skeptical; on the other hand, it might well be the case that the same type of people are more aware of the shortcomings of ADS (e.g., [3, 12, 13]).

5.2. Qualitative insights based on open-ended questions

We also collected unstructured textual data based on open-ended questions embedded in our questionnaires. An in-depth analysis reveals that many study participants are convinced that automation is precisely the reason why decisions are fair ("Automated system is fair by design"). They perceive the ADS as fair because, in their opinion, its decisions are objective: "it [the ADS] states the criteria and follows [them], there is no room for subjectivity and the data is used to make an objective decision." This is likely one of the reasons why our hypotheses **H1** and **H2** are not supported. While some participants allude to underlying issues of automated decisions ("AI can be programmed to be unfair" and "I do not believe an Automated Decision System can replace a human. We can't expect it to not make mistakes"), most view the ADS as fair because the system is "purely looking at numbers [therefore] its [*sic*] completely fair." Finally, one person points out that the situation "is fair because the consumer knows that he has been judged using an algorithm."

On the other hand, an interesting comment states that "[t]he decision may have been made by a machine, but someone decided to program it that way," which raises questions around accountability of ADS. Some issues are equally criticized in the human and the ADS condition: "I don't think it is fair to take education, gender or marital status into account," or "[s]ome factors are indifferent to the decision of the loan and are personal information." Even though overall the human condition is perceived as significantly less informationally fair than the ADS condition and people believe the ADS "can help eliminate [...] bias," there are still participants who "hope bots wont [*sic*] have to decide crucial life decisions for [them]."

6. Conclusion and outlook

We conducted an online study with 200 participants to evaluate differences in people's perceptions of informational fairness and trustworthiness towards human vs. automated decision-making in the high-stakes context of lending. We provided thorough explanations to study participants, identical in both conditions (human and automated), to facilitate meaningful and unbiased responses. Our findings suggest that within the scope of our study setup—contrary to some prior work as well as our own hypothesis—automated decisions are perceived as more informationally fair than human-made decisions. In contrast, no significant differences were measured for trustworthiness in our case. Based on qualitative responses, it appears that people particularly appreciate the absence of subjectivity in ADS as well as their data-driven approach. Interestingly, our analyses also imply that people's AI literacy affects their perceptions, given the provided explanations. Specifically, we found that people with high AI literacy tend to perceive ADS as both fairer and more trustworthy than a human decision-maker, whereas no significant differences for either construct were detected for people with low AI literacy.

Based on our findings, we may conjecture that providing thorough explanations can enhance perceptions of fairness and trustworthiness towards ADS over human decision-makers—particularly for people with higher AI literacy. This hypothesis will have to be tested in follow-up work. However, we must be cognizant of the dangers of wrongful persuasion and automation biases, i.e., the tendency of people to over-rely on ADS—which might become a problem if too many (compelling) explanations about the inner workings of ADS are provided. Future work should also account for this by examining how perceptions change when the quality of the ADS changes for the worse (e.g., by making unfair decisions) [55]. Other

natural extensions include the consideration of domains other than lending, as well as the adoption of different explanation styles. We hope that our work will stimulate multifaceted future research on this topic of utmost societal relevance.

References

[1] N. R. Kuncel, D. M. Klieger, and D. S. Ones, "In hiring, algorithms beat instinct," *Harvard Business Review*, 2014.

[2] S. Townson, "AI can make bank loans more fair," *Harvard Business Review*, 2020.

[3] W. D. Heaven, "Predictive policing algorithms are racist. They need to be dismantled.," *MIT Technology Review*, 2020.

[4] J. G. Harris and T. H. Davenport, "Automated decision making comes of age," *MIT Sloan Management Review*, vol. 46, no. 4, pp. 2–10, 2005.

[5] A. Chalfin, O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan, "Productivity and selection of human capital with machine learning," *American Economic Review*, vol. 106, no. 5, pp. 124–127, 2016.

[6] S. Koivunen, T. Olsson, E. Olshannikova, and A. Lindberg, "Understanding decision-making in recruitment: Opportunities and challenges for information technology," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. GROUP, pp. 1–22, 2019.

[7] T. Grote and P. Berens, "On the ethics of algorithmic decision-making in healthcare," *Journal of Medical Ethics*, vol. 46, no. 3, pp. 205–211, 2020.

[8] S. Triberti, I. Durosini, and G. Pravettoni, "A "Third Wheel" effect in health decision making involving artificial entities: A psychological perspective," *Frontiers in Public Health*, vol. 8, 2020.

[9] B. Lepri, J. Staiano, D. Sangokoya, E. Letouzé, and N. Oliver, "The tyranny of data? The bright and dark sides of data-driven decision-making for social good," in *Transparent Data Mining for Big and Small Data*, pp. 3–24, Springer, 2017.

[10] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI: Challenges and Opportunities," *Business & Information Systems Engineering*, vol. 62, pp. 379–384, 2020.

[11] B. Imana, A. Korolova, and J. Heidemann, "Auditing for discrimination in algorithms delivering job ads," in *WWW*, 2021.

[12] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *FAccT*, pp. 77–91, 2018.

[13] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine bias," *ProPublica*, 2016.

[14] A. Satariano, "British grading debacle shows pitfalls of automating government," *The New York Times*, 2020.

[15] Edelman, "Edelman Trust Barometer 2021," 2021.

[16] S. Barocas, M. Hardt, and A. Narayanan, "Fairness and machine learning," 2018.

[17] M. Srivastava, H. Heidari, and A. Krause, "Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning," in *KDD*, pp. 2459–2468, 2019.

[18] J. Wanner, L.-V. Herm, and C. Janiesch, "How much is the black box? The value of explainability in machine learning models," *ECIS*, pp. 1–14, 2020.

[19] F. Peters, L. Pumplun, and P. Buxmann, "Opening the black box: Consumer's willingness to pay for transparency of intelligent systems," *ECIS*, 2020.

[20] C. Meske, E. Bunde, J. Schneider, and M. Gersch, "Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities," *Information Systems Management*, pp. 1–11, 2020.

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

[22] Atico, "Automated credit decisioning for enhanced efficiency," 2021.

[23] D. Carey and M. Smith, "How companies are using simulations, competitions, and analytics to hire," *Harvard Business Review*, 2016.

[24] C. Lewis and R. Mack, "The role of abduction in learning to use a computer system," 1982.

[25] C. Fernandez, F. Provost, and X. Han, "Counterfactual explanations for data-driven decisions," *ICIS*, 2019.

[26] L. Breiman, "Random forests," *Machine Learning*, 2001.

[27] R. Goebel, A. Chander, K. Holzinger, F. Lecue, Z. Akata, and Others, "Explainable AI: The new 42?," in *CD-MAKE*, pp. 295–303, 2018.

[28] R. Binns, M. Van Kleek, M. Veale, U. Lyngs, J. Zhao, and Others, "It's reducing a human being to a percentage; Perceptions of justice in algorithmic decisions," in *CHI*, pp. 1–14, 2018.

[29] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, "Explaining models: An empirical study of how explanations impact fairness judgment," in *IUI*, pp. 275–285, 2019.

[30] M. K. Lee, "Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management," *Big Data & Society*, vol. 5, no. 1, pp. 1–16, 2018.

[31] M. K. Lee and S. Baykal, "Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division," in *CSCW*, pp. 1035–1048, 2017.

[32] 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, no. CSCW, pp. 182:1–182:26, 2019.

[33] R. F. Kizilcec, "How much information? Effects of transparency on trust in an algorithmic interface," in *CHI*, pp. 2390–2395, 2016.

[34] K. Crawford and R. Calo, "There is a blind spot in AI research," *Nature*, vol. 538, no. 7625, pp. 311–313, 2016.

[35] N. Castelo, M. W. Bos, and D. R. Lehmann, "Task-dependent algorithm aversion," *Journal of Marketing Research*, vol. 56, no. 5, pp. 809–825, 2019.

[36] M. F. Kramer, J. Schaich Borg, V. Conitzer, and W. Sinnott-Armstrong, "When do people want AI to make decisions?," in *AIES*, pp. 204–209, 2018.

[37] M. Eslami, A. Rickman, K. Vaccaro, A. Aleyasen, A. Vuong, K. Karahalios, K. Hamilton, and C. Sandvig, ""I always assumed that I wasn't really that close to [her]": Reasoning about invisible algorithms in news feeds," in *CHI*, pp. 153–162, 2015.

[38] B. J. Dietvorst, J. P. Simmons, and C. Massey, "Algorithm aversion: People erroneously avoid algorithms after seeing them err," *Journal of Experimental Psychology: General*, vol. 144, no. 1, p. 114, 2015.

[39] D. Long and B. Magerko, "What is AI literacy? Competencies and design considerations," in *CHI*, 2020.

[40] D. Chan, "Perceptions of fairness," 2011.

[41] F. Bélanger, J. S. Hiller, and W. J. Smith, "Trustworthiness in electronic commerce: The role of privacy, security, and site attributes," *The Journal of Strategic Information Systems*, vol. 11, no. 3-4, pp. 245–270, 2002.

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

[43] J. Schoeffer, Y. Machowski, and N. Kuehl, "A study on fairness and trust perceptions in automated decision making," in *Joint Proceedings of the ACM IUI 2021 Workshops*, 2021.

[44] J. A. Colquitt, D. E. Conlon, M. J. Wesson, C. O. L. H. Porter, and K. Y. Ng, "Justice at the millennium: A meta-analytic review of 25 years of organizational justice research," *Journal of Applied Psychology*, vol. 86, no. 3, p. 425, 2001.

[45] L. Carter and F. Bélanger, "The utilization of e-government services: Citizen trust, innovation and acceptance factors," *Information Systems Journal*, vol. 15, no. 1, pp. 5–25, 2005.

[46] C.-M. Chiu, H.-Y. Lin, S.-Y. Sun, and M.-H. Hsu, "Understanding customers' loyalty intentions towards online shopping: An integration of technology acceptance model and fairness theory," *Behaviour & Information Technology*, vol. 28, no. 4, pp. 347–360, 2009.

[47] A. Wilkinson, J. Roberts, and 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*, vol. 26, no. 6, pp. 1369–1376, 2010.

[48] Infosys, "How FinTechs can enable better support to FIs' credit decisioning?." 2019.

[49] S. Palan and C. Schitter, "Prolific.ac—A subject pool for online experiments," *Journal of Behavioral and Experimental Finance*, vol. 17, pp. 22–27, 2018.

[50] A. Joshi, S. Kale, S. Chandel, and D. K. Pal, "Likert scale: Explored and explained," *British Journal of Applied Science & Technology*, vol. 7, no. 4, pp. 396–403, 2015.

[51] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981.

[52] D. Barclay, C. Higgins, and R. Thompson, *The partial least squares (PLS) approach to casual modeling: Personal computer adoption and use as an illustration*. 1995.

[53] J. M. Cortina, "What is coefficient alpha? An examination of theory and applications," *Journal of Applied Psychology*, vol. 78, no. 1, pp. 98–104, 1993.

[54] P. E. McKnight and J. Najab, "Mann-Whitney U Test," *The Corsini Encyclopedia of Psychology*, 2010.

[55] J. Schoeffer and N. Kuehl, "Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems," in *CSCW '21 Companion*, 2021.