"There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making

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

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

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

ABSTRACT

Automated decision systems (ADS) are increasingly used for consequential decision-making. These systems often rely on sophisticated yet opaque machine learning models, which do not allow for understanding how a given decision was arrived at. In this work, we conduct a human subject study to assess people's perceptions of *informational fairness* (i.e., whether people think they are given adequate information on and explanation of the process and its outcomes) and *trustworthiness* of an underlying ADS when provided with varying types of information about the system. More specifically, we instantiate an ADS in the area of automated loan approval and generate different explanations that are commonly used in the literature. We randomize the amount of information that study participants get to see by providing certain groups of people with the same explanations as others *plus* additional explanations. From our quantitative analyses, we observe that different amounts of information as well as people's (self-assessed) AI literacy significantly influence the perceived informational fairness, which, in turn, positively relates to perceived trustworthiness of the ADS. A comprehensive analysis of qualitative feedback sheds light on people's desiderata for explanations, among which are (i) consistency (both with people's expectations and across different explanations), (ii) disclosure of monotonic relationships between features and outcome, and (iii) actionability of recommendations.

CCS CONCEPTS

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Computing methodologies** → **Machine learning**; • **Information systems** → **Decision support systems**.

KEYWORDS

Automated decision-making, explanations, informational fairness, machine learning, perceptions, trustworthiness

ACM Reference Format:
Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 13 pages. https://doi.org/10.1145/3531146.3533218

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06…$15.00
https://doi.org/10.1145/3531146.3533218

1 INTRODUCTION

Automated decision-making has become ubiquitous in many high-stakes domains such as hiring [65], bank lending [105], grading [92], and policing [47], among others. The underlying motives of adopting automated decision systems (ADS)[1] are manifold: they range from cost-cutting to improving performance and enabling more robust and objective decisions [46, 65, 85]. Hopes are also that, if properly designed, ADS can be a valuable tool for breaking out of vicious patterns of human stereotyping and contributing to social equity, e.g., in the realms of recruitment [19, 60], health care [43, 106], or financial inclusion [73]. However, ADS are typically based on ML techniques, which, in turn, rely on historical data. If, e.g., this underlying data is biased (e.g., because certain socio-demographic groups were favored in a disproportionate way), an ADS will learn from and perpetuate existing patterns of unfairness [35]. Prominent examples of such behavior from the recent past are race and gender stereotyping in job ad delivery [51], as well as the discrimination of Latinx and African-American borrowers in algorithmic mortgage loan pricing [7]. These and other cases have put ADS under enhanced scrutiny, justifiably jeopardizing trust in these systems [31].

In recent years, a growing body of AI and ML research has been devoted to detecting, quantifying, and mitigating unfairness in ADS [81]. A significant share of this work has focused on formalizing different concepts of *fairness* through statistical equity constraints, many of which are at odds with each other [23, 58]. As a consequence, there cannot be a one-size-fits-all technical fairness criterion. Moreover, in many cases, these techno-centric works do not explicitly take into account the opinions of people that are (potentially) affected by such automated decisions. While the FAccT community has made a plethora of impactful contributions over the past years, it is still crucial to better understand people's perceptions and attitudes towards ADS—in addition to how researchers may define those systems' fairness in technical terms.

A related issue revolves around explaining automated decisions to affected individuals. As ADS employ ever more sophisticated and "black-box" ML models, several problems arise; one of which is the hampered detectability of adverse behavior of such systems. In order to safeguard transparency and accountability of automated

[1] A summary of our abbreviations is given in Tab. 3 in § A (appendix)

decisions, several laws and regulations demand a "right to explanation". The EU General Data Protection Regulation (GDPR), e.g., requires the disclosure of "the existence of automated decision-making, including [...] meaningful information about the logic involved [...]" [34] to data subjects. In fact, it has been shown, among others, that explanations can enhance people's understanding of certain automated decisions [75]. For most real-world cases, however, those regulations generally remain (too) vague and little actionable—which often results in deficient adoption, as noticed in the context of bank lending [103]. Moreover, research on *explainable AI* (XAI) suggests that there exists no one-size-fits-all approach to explaining ADS either [5, 67].

In this work, we conduct a human subject study to examine the effects of explanations on people's perceptions towards an automated loan approval system, where we randomize the type and amount of information that study participants get to see. The primary dependent variables that we are interested in are perceptions of *informational fairness* of the system (i.e., whether people think they are given adequate information on and explanation of the decision-making process and its outcomes) as well as perceived *trustworthiness*, and the relationship between both. We also assess the influence of people's (self-assessed) *AI literacy* on the outcomes. Finally, we ask multiple open-ended questions w.r.t. people's ability to assess the given system's fairness, as well as regarding the appropriateness of explanations' content.

2 BACKGROUND AND RELATED WORK

Topics of *fairness* and *trustworthiness* have become important pillars of AI and HCI research in recent years. In this section, we provide an overview of relevant literature and highlight our contributions. For brevity, we do not explicitly cover the vast technical literature on algorithmic fairness. While we assume that the FAccT community is familiar with seminal work in this field, we refer interested readers from other disciplines to relevant survey literature: [6, 18, 81].

It is—albeit unsurprisingly—important to note that a "fair" (according to some technical fairness notion) system does not imply that people perceive it as such; either because their personal fairness concepts differ from the employed technical notion or because they are not enabled to assess the system's (un)fairness, to begin with. In fact, it must be questioned whether an ADS that satisfies given statistical notions of fairness (e.g., equitable distribution of outcomes) can ever be *truly* considered fair when at the same time decision-subjects are left in the dark w.r.t. the inner workings of the system. Instead, *fairness* (of ADS) is likely a multi-faceted construct that encompasses different dimensions, similar to dimensions of (organizational) justice [25, 27], which are commonly made up of distributive, procedural, interpersonal, and informational justice [27]. While distributive and procedural aspects have been considered in the context of ADS (e.g., in [42, 71, 78]), work on *informational fairness* of ADS is lacking.

Borrowing from [20], we call a system *informationally fair* if it conveys adequate information on and explanation of the decision-making process and its outcomes; and we define *adequate information* (similar to [27]) as information being *thorough, reasonable, tailored* (to individual needs), as well as helping people *understand* the decision-making process, and enabling them to judge whether this process is fair or unfair. We refer to § B for an overview of our measurement items.[2] *Trustworthiness* is a well-established construct that, according to [16], is defined as "the perception of confidence in the [...] reliability and integrity [of an ADS]." We refer the reader to [53, 68, 109] for survey literature on trust and trustworthiness.

2.1 Related work

Automated decision systems. Harris and Davenport [46] define *automated decision systems* (ADS) as systems that aim to minimize human involvement in decision-making processes. In this work, we assume ADS to be supervised ML models. In many cases, ADS have the potential to make more consistent decisions than humans. Such systems are popular in many industries, such as banking [46, 105] or hiring [17, 19, 60, 65]—and they are emerging in new areas as well, e.g., in health care [43, 106]. With their increasing adoption in different consequential areas, it is important to ensure that ADS reach fair decisions that are transparent, primarily, to affected individuals or auditors. However, there have been multiple cases in the recent past where algorithms made biased decisions that discriminated against certain groups, e.g., based on gender or race [3, 14, 47]. In other instances, ADS have been operating in an opaque ("black-box") fashion, making it, among others, difficult (i) for affected individuals to grasp the rationale behind certain decisions, and (ii) for regulatory agencies and other responsible stakeholders to vet such systems appropriately [88]. On that account, fairness and transparency of ADS have become important topics of interest for the research community. Interestingly, despite known weaknesses of ADS, some prior work has found that human-made decisions are *not* generally perceived as fairer or more trustworthy than automated decisions; primarily for reasons of (alleged) consistency in automated decision-making [95, 98].

Explainable AI. Despite being a popular topic of current research, XAI is a natural consequence of designing ADS and, as such, has been around at least since the 1980s [74]. Its importance, however, keeps rising as increasingly sophisticated (and opaque) AI techniques are used to inform ever more consequential decisions. XAI is not only required by law (e.g., GDPR, ECOA); Eslami et al. [33], e.g., have shown that users' attitudes towards algorithms change when transparency is increased. In general, both quantity and quality of explanations matter: Kulesza et al. [64] explored the effects of soundness and completeness of explanations on end users' mental models and suggest, among others, that oversimplification is problematic. Recent findings from Langer et al. [66], on the other hand, suggest that in the case of automated job interviews it might make sense to withhold certain pieces of information from applicants in order to not evoke negative reactions.

Even in the presence of explanations, people sometimes rely too heavily on system suggestions [15], a phenomenon commonly referred to as *automation bias* [28, 37]. Ehsan and Riedl [32] have also used the term "explainability pitfalls" for any such unanticipated negative effects of explanations (e.g., unwarranted trust [94]). Eventually, Chromik et al. [24] (inspired by seminal work related to UX design [39]) warn that explanations can be exploited to purposefully deceive users for the benefit of *other* stakeholders. Hence,

[2] The appendix can, e.g., be accessed here: https://github.com/jakobschoeffer/facct22-130-appendix

explanations are by no means the "silver bullet" when it comes to solving problems of opaque AI systems [8]. A comprehensive overview of XAI stakeholders and their distinct desiderata is given by Langer et al. [67]. For instance, people affected by automated decisions may be particularly interested in explanations that enable them to evaluate the fairness and trustworthiness of the underlying systems [67, 96]. This desideratum is closely linked to informational fairness of ADS [25], as introduced earlier. We refer the interested reader to, among others, [2, 4, 5, 38, 44, 67, 82, 83] for more in-depth literature on different XAI techniques and their inner workings. Regarding the effectiveness of explanations, generally speaking, prior research has primarily focused on comparing individual explanation styles head-to-head (e.g., [10, 30]), while little work has been done on evaluating the interplay of different styles, including potential complementarity. Langer et al. [67] emphasize the sparsity of empirical work w.r.t. the effectiveness of explanations overall.

Perceptions towards ADS. 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 et al. [10] and Dodge et al. [30] 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. A different line of research has examined people's moral judgments w.r.t. the use of specific features in ADS [40, 42], also with mixed empirical findings. Lee [69] compares perceptions of fairness and trustworthiness depending on whether the decision maker is a person or an algorithm in the context of managerial decisions. 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 [70] explore how algorithmic decisions are perceived in comparison to group-made decisions. Wang et al. [110] 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. [71] 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. Woodruff et al. [112] conducted workshops with people from traditionally marginalized backgrounds, inferring that awareness of unfairness in ADS can substantially affect trust in companies or products.

Some work has also assessed the impact of people's demographics (including gender [89]), as well as political views and task experience [41] on their perceptions. Saxena et al. [93] examined lay people's perceptions of different technical fairness notions for ADS, suggesting that people prefer notions related to meritocratic fairness [54, 77]. Regarding trustworthiness, Kizilcec [57], e.g., 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. Kästner et al. [55] also examined the relationship between explainability and trust(worthiness), urging system designers to engineer for trustworthiness (as opposed to trust), and indicating that explanations can be a crucial toolbox towards that goal. Regarding perceptions of different social groups, Lee and Rich [72] point out that prior studies have mostly recruited respondents from `Amazon Mechanical Turk` [87], which has predominantly white participants [48]—because of this, among other reasons [90] we have recruited our study participants through `Prolific`[3] [86].

2.2 Research gaps and our contributions

We aim to complement prior work to better understand how much of which information should be provided so that people are optimally enabled to understand the inner workings and appropriately assess the fairness and trustworthiness of ADS. To that end, we conducted a randomized experiment to examine people's perceptions of informational fairness and trustworthiness towards an automated loan approval system, given different combinations of common explanations (relevant factors, factor importance, and counterfactual explanations). While there exists prior work on trustworthiness perceptions for *individual* explanation styles, we see a significant gap w.r.t. assessing *combinations* of different explanations. We argue that this is an important gap to fill because different explanations convey different information and will likely have to be leveraged complementarily (i.e., *not* in isolation) in practice. On a related note, we also set about examining the marginal effects of providing certain explanations *on top of* others—which, to the best of our knowledge, has not been analyzed in depth before. As a consequence, we alter the *amount of information* that different groups of people get to see. We do by no means claim to examine these aspects exhaustively, but we hope that our work will be a stepping stone for further research.

Finally, and perhaps most importantly, we shift focus from examining distributive and procedural fairness perceptions to *informational fairness*. In other words, we do *not* ask people whether they find particular ADS outcomes or procedures fair or not, but—broadly speaking—whether they feel they received sufficient information to *assess* a given system. This is an important distinction. Only very few works have considered the informational fairness dimension when experimentally evaluating effectiveness of ADS explanations: Binns et al. [10] only measure the understandability aspect of informational fairness for individual explanation styles; Schlicker et al. [95] and Schoeffer et al. [98] assess informational fairness perceptions, but with a focus on comparing human with automated decision makers. Uhde et al. [107] and Brown et al. [12] conducted interviews [107] and workshops [12] to infer qualitative statements related to informational fairness; whereby Brown et al. [12] explicitly state that "more research is needed to understand how different elements of algorithmic systems affect perceptions of [...] informational justice." Empirical work on the interplay of informational fairness and trustworthiness perceptions for ADS is, to our knowledge, entirely novel. Finally, we also analyze the relationship between study participants' (self-assessed) AI literacy and their perceptions, and we qualitatively examine their answers to open-ended question regarding (in)appropriateness of explanations as well as what information they feel is missing (if any) to properly vet the given ADS.

[3] `Prolific` is a crowdworking platform for online research: https://www.prolific.co/

3 RESEARCH HYPOTHESES

The conditions of our experiment comprise different amounts of information that study participants get to see w.r.t. an ADS in the realm of automated loan decisioning. Regarding the potential effects of varying amounts of information on our dependent variables of perceived informational fairness and trustworthiness, we formulate two research hypotheses based on preliminary qualitative insights w.r.t. people's desire for transparency and information [12, 107] as well as prior findings from the psychology literature [26, 27, 50, 76, 104, 108]. First, assuming that explanations are not entirely lacking in content, we conjecture (similar to [12, 107]) that more provided information leads to higher informational fairness perceptions. Regarding effects on trustworthiness perceptions, we note that several factors contribute to a system's fairness [27, 71]; among these are *consistency* (of decision-making procedures) as well as *process and outcome control* on behalf of decision-subjects [29, 71]. *Process control* means that decision-subjects have the "ability to influence what [...] data is considered by the decision maker" [71], and *outcome control*, borrowing from [50], refers "to the ability to appeal or modify the outcome [...] once it has been made" [71]. While we do not anticipate our employed explanations to readily increase perceptions of outcome control, we conjecture that certain information may enhance assumed process control, which, in turn, affects procedural fairness perceptions [27, 71] and, ultimately, trust [108].

- **H1** As the amount of information provided increases, perceptions of informational fairness towards the ADS increase.
- **H2** As the amount of information provided increases, perceptions of trustworthiness towards the ADS increase.

While investigating these relationships, we are not only interested in the effects of our conditions on informational fairness and trustworthiness but also in the relationship between the latter two. Some prior work has examined the relationship between informational fairness/justice and trust/trustworthiness (e.g., [26, 36, 113]) in other contexts. Frazier et al. [36] identified a significant positive effect of informational justice on different facets of trustworthiness perceptions in one of their two examined settings in the realm of organizational justice. Similarly, Zhu and Chen [113], in the context of customer satisfaction in internet banking, found that informational fairness (as a component of overall systemic fairness) has a positive effect on trust. Finally, Colquitt and Rodell [26] affirm that "conventional wisdom on the justice-trust connection" implies a causal path from (informational) justice to trust, and not the other way round. While these works address different use cases, we conjecture a positive relationship between informational fairness and trustworthiness perceptions for our ADS setting as well:

- **H3** Perceptions of informational fairness relate positively to perceptions of trustworthiness.

Experts may have a different attitude towards procedures or phenomena that touch on their area of expertise than non-experts. Slovic et al. [101, 102], e.g., found differences in risk perceptions between experts and lay people. Regarding innovative (food) technologies, Siegrist [99] notes that lay people may neither be able to assess risks nor benefits appropriately. For the specific case of ADS, Wang et al. [110] found a significant effect of computer literacy on a mix of procedural and distributive fairness perceptions; specifically, their findings suggest that fairness perceptions are lower for people with lower computer literacy. Pierson [89], along the same lines, found that students' views on algorithmic fairness changed by increasing algorithmic literacy through lecture and discussion: students "became more likely to emphasize transparency, [and] more open to using algorithms rather than using judges." [89] Finally, intuition tells us that AI-literate people may "extract" more information and understanding out of ADS explanations (e.g., because they know how supervised ML in general works).

- **H4** People with higher AI literacy perceive an automated decision system to be more informationally fair than people with little or no knowledge in the field.
- **H5** People with higher AI literacy perceive an automated decision system to be more trustworthy than people with little or no knowledge in the field.

4 METHODOLOGY

We examine our hypotheses in the context of algorithmic lending. We argue that this is a common context that affects many people at some point in life. It is, furthermore, an area where ADS are typically already utilized within productive settings [1, 52]. Specifically, we confront study participants (SPs) with situations where a person was denied a loan. Similar to [10], we argue that, in practice, explanations are much more likely to be requested by decision-subjects in response to negative outcomes; or, in other words: if someone gets the loan, interest in how and why exactly the decision was arrived at will likely drop. However, we do by no means imply that reactions to positive outcomes are unworthy of being examined—given budget constraints, we defer them to future work.

4.1 Study design

We choose a between-subject design with the following conditions: first, we reveal to SPs some basic information about the lending company. We then explain that a given individual's loan application was rejected by the company, as well as that this decision was communicated to the applying individual electronically and in a timely fashion (see Fig. 1 for the exact wording in our questionnaires). Afterwards, we provide one of four explanations (i.e., conditions) to each SP. Eventually, we measure the effects of assigning different conditions—and by design of the conditions, different amounts of information (AMTIN)—on two dependent variables: perceived informational fairness (INFF) and perceived trustworthiness (TRST) regarding the ADS. (Recall that informational fairness perceptions do *not* involve an actual assessment of the system's fairness w.r.t. its processes or outcomes.) Additionally, we measure the (self-assessed) AI literacy (AILIT) of SPs. We analyze whether differences in SPs' AI literacy affect their perceptions. All measurement items are summarized in § B. Note that for each construct, we measure multiple items; mostly drawn (and partially adapted) from prior work.

ADS Setup. The ADS for our study consists of a random forest classifier which predicts loan approval on unseen data and is able to output different explanations. For training our model, we utilize a publicly available dataset on home loan application decisions [22], which has been used in multiple data science competitions on `Kaggle`. Note that comparable data—reflecting a given finance

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 communicated to the applying individual electronically and in a timely fashion.

Figure 1: Introduction of use case in questionnaires.

company's individual circumstances and approval criteria—might in practice be used to train ADS [52]. 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, 480 observations remain, 332 of which (69.2%) involve the positive label (Y) and 148 (30.8%) the negative label (N). We used 70% of the dataset to train our ADS and use the remaining 30% as a holdout set for the experiment. After encoding and scaling the features, we trained a random forest classifier with bootstrapping [11], which achieves an out-of-bag accuracy estimate of 80.1% on the held-out data. We use this classifier's predictions on the holdout set as a basis for the upcoming conditions/explanations that the SPs are confronted with. Since we are *not* asking to assess the actual (procedural or distributive) fairness of the ADS, it is not critical to quantify how fair the system really is—any such effort would be highly contestable anyhow, for reasons of incompatible fairness notions [23, 58, 84]. The authors still (informally but independently) checked training data as well as output quality for any salient problems that may bias SPs' responses w.r.t. the dependent variables.

Explanations. We impose several requirements on the explanations that we provide to SPs: overall, we employ only model-agnostic explanations [2] in a way that they could plausibly be provided to loan applicants (i.e., lay people) in real-world scenarios. While explanations can be communicated in a wide variety of ways (see, e.g., [2, 4, 44, 82]), we confine ourselves to textual explanations (esp. no visuals) to control for differences in conveyance. We also pick explanations that are immediately understandable semantically—this is important so as to collect meaningful responses. On a related note, we ensure that explanations are not too long, in order to account for known issues around information overload [9]. Finally, and similar to [10], we pick explanations that can plausibly provide insights about a system's "logic involved," as required, e.g., by the GDPR. Based on these preliminaries, we assign SPs to one of four conditions that involve combinations of explanations w.r.t. (i) factors considered by the ADS, (ii) relative importance of these factors, and (iii) counterfactual scenarios where a rejected applicant would have been granted the loan. We acknowledge that additional explanation styles would be equally interesting to consider; however, in order to keep the experiment size manageable, we must defer them to future work.

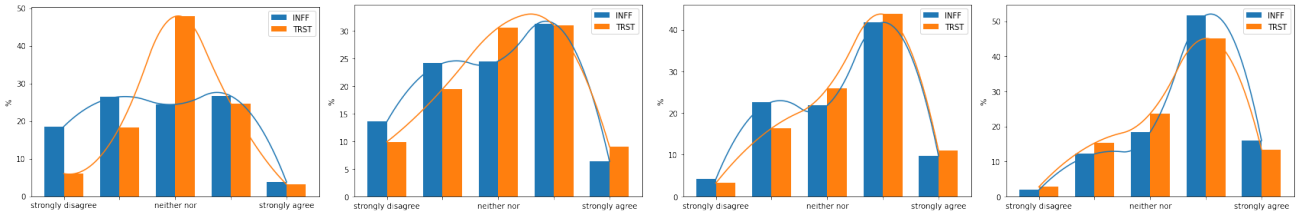
Our first condition, *(Base)*, only reveals to the SPs that the loan decision was communicated to the applying individual electronically and in a timely fashion (as in Fig. 1). Apart from the *(Base)* condition—which might be regarded as a black-box system—all other conditions include the additional information that the loan decision was made by an ADS (i.e., automated). The second condition, *(F)*, consists of disclosing the factors, 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 two different settings in each questionnaire, where settings are chosen at random from the pool of rejected applicants. The authors, again, checked informally that no highly unusual (e.g., extreme outliers) settings were displayed that might distract SPs' perceptions and bias recorded responses. Please refer to § C for an exemplary setting (introduction of use case plus conditions). Next, we computed permutation feature importance [11] from our model and obtained the following hierarchy, using "$\succ$" as a shorthand for "is more important than": *credit history* $\succ$ *loan amount* $\succ$ *applicant income* $\succ$ *co-applicant income* $\succ$ *property area* $\succ$ *marital status* $\succ$ *dependents* $\succ$ *education* $\succ$ *loan amount term* $\succ$ *self-employment* $\succ$ *gender*. Revealing this ordered list in conjunction with *(F)* makes up our third condition, *(FFI)*. To construct our fourth condition, we conducted an online survey with 20 quantitative and qualitative researchers to ascertain which of the aforementioned factors 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 factors are *loan amount, loan amount term, property area, applicant income, co-applicant income*. Our fourth condition *(FFICF)* is then—in conjunction with *(F)* and *(FFI)*—the provision of three counterfactual scenarios where one actionable factor each is (minimally) altered such that our model predicts a loan approval instead of a rejection. Our four conditions are summarized as follows:

(Base)	Baseline without further explanations.
(F)	Disclosure of factors.
(FFI)	Disclosure of factors and factor importance.
(FFICF)	Disclosure of factors, factor importance, and counterfactual scenarios.

Note that the order of provided explanations ($(Base) \rightarrow (F) \rightarrow (FFI) \rightarrow (FFICF)$) is not arbitrary: each subsequent condition provides the exact same information as the previous one *and more*. Since, e.g., factor importances implicitly reveal which factors the ADS considers, this would not necessarily hold true for, e.g., $(FI) \rightarrow (FIF)$.

4.2 Data collection

Study participants (SPs) for our online study were (voluntarily) recruited via `Prolific` [86] and asked to rate their agreement with multiple statements w.r.t. our dependent variables as well as their AI literacy on 5-point Likert scales—where 1 corresponds to "strongly disagree" and 5 denotes "strongly agree". Additionally, we included multiple open-ended questions in the questionnaires to be able to better understand the reasoning behind SPs' quantitative responses.



(Base) *(F)* *(FFI)* *(FFICF)*

Figure 2: Distributions of responses for informational fairness (INFF) and trustworthiness (TRST) per condition.

The SPs were randomly and in equal proportions assigned to one of the four conditions, and each SP was provided with two consecutive questionnaires associated with two different settings. We collected 401 responses, of which 4 had to be eliminated due to failure to pass one or more attention checks. Thus, we obtained 397 analyzable responses. Among the SPs, 60% indicated to be male, 39% female, and the remaining SPs either responded with "non-binary" or chose not to disclose their gender; 46% were students, 27% employed full-time, 8% employed part-time, 7% self-employed, 11% unemployed, less than 1% retired, and 1% chose not to disclose their profession. The reported average age of SPs was 25.7. SPs were monetarily compensated above the recommended min. pay of $6.50 per hour.

5 QUANTITATIVE ANALYSES AND RESULTS

We now examine the effects of our conditions and people's (self-assessed) AI literacy on perceived informational fairness and trustworthiness of our ADS. For our measurement model, describing a confirmatory factor analysis and reporting correlations and factor loadings, we refer the reader to § D. In this section, we first present the results of group difference analyses for our conditions with tests for pairwise comparison. After that, we report our findings on the validation of our hypotheses **H1** to **H5** with a full structural equation model.

5.1 Analysis of group differences

Since we cannot confirm the assumption of normality for all variables, we conduct multiple non-parametric Kruskal-Wallis H tests for multiple group comparisons [61]. Afterwards, we carry out pairwise comparisons using Bonferroni-corrected Mann-Whitney U tests [80]. With these tests, we initially assess the effects of our four conditions revealing different amount of information (AMTIN) on the constructs of informational fairness (INFF) and trustworthiness (TRST). Overall, we find a significant effect between different conditions on perceptions of informational fairness ($p < 0.001$) as well as on perceptions of trustworthiness ($p < 0.001$). A Mann-Whitney U test for pairwise comparisons shows that the effect for informational fairness is significant ($p < 0.05$) between all conditions except *(Base)* and *(F)*. The effect for trustworthiness is significant between *(Base)* and *(FFI)*, *(Base)* and *(FFICF)*, as well as *(F)* and *(FFICF)*, and marginally significant between *(F)* and *(FFI)* ($p = 0.052$). Looking at the mean response values for (INFF) and (TRST) by condition (see Tab. 1), we note that they are increasing as more information is shown to SPs. Please refer to Fig. 2 for the distribution of responses by condition, and to Tab. 2 for a detailed summary of the results of the Mann-Whitney U tests.

Table 1: Means and standard deviations of response values for informational fairness (INFF) and trustworthiness (TRST) by condition. All items were measured on 5-point Likert scales.

Condition	M(INFF)	SD(INFF)	M(TRST)	SD(TRST)
(Base)	2.71	1.16	3.01	0.89
(F)	2.93	1.16	3.10	1.12
(FFI)	3.30	1.05	3.43	0.99
(FFICF)	3.68	0.94	3.51	0.99

Notes: M = Mean; SD = Standard deviation

5.2 Hypotheses testing

We estimate a full structural equation model (SEM), the results of which are depicted in Fig. 3. We also report more exhaustive information, including standard errors, z-values, p-values, and standardized path estimates in Tab. 6 in § E. Consistent with using Kruskal-Wallis H tests for group comparisons, we estimate our SEM using unweighted least squares (ULS) because this estimator makes no distributional assumptions. We assess the fit of our model with multiple common measures: the comparative fit index (CFI) as well as Tucker-Lewis index (TLI) should be above 0.9 [59], root mean square error of approximation (RMSEA) below 0.05 [13], and standardized root mean squared residual (SRMR) below 0.08 [45] to indicate good model fit. Our model's values are

$$CFI = 0.997;\ TLI = 0.997;\ RMSEA = 0.024;\ SRMR = 0.051.$$

Hence, all considered fit measures meet the required thresholds. Note that the chi-square test is not a meaningful measure of model fit in our case because variables are not normally distributed, and because we apply the ULS method to estimate our model [56].

In the following, we use a shorthand for our variables: AMTIN, AILIT, INFF, TRST (as introduced in § 4.1 and summarized in Tab. 3 of § A). To investigate our hypotheses, we first examine the effect of AMTIN on INFF. As expected, and previously supported by the Kruskal-Wallis H test as well as the comparison of means between different conditions, increasing AMTIN has a significant positive effect on INFF (0.37^{***}). Hence, **H1** is supported.

Table 2: Pairwise differences in perceptions of informational fairness (INFF) and trustworthiness (TRST) between conditions.

INFF			TRST		
Condition 1	**Condition 2**	**Difference**	**Condition 1**	**Condition 2**	**Difference**
(Base)	*(F)*	n/s	*(Base)*	*(F)*	n/s
(Base)	*(FFI)*	***	*(Base)*	*(FFI)*	***
(Base)	*(FFICF)*	***	*(Base)*	*(FFICF)*	***
(F)	*(FFI)*	*	*(F)*	*(FFI)*	n/s
(F)	*(FFICF)*	***	*(F)*	*(FFICF)*	**
(FFI)	*(FFICF)*	**	*(FFI)*	*(FFICF)*	n/s

Notes: $^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$; n/s: not significant

Next, we examine the influence of AMTIN on TRST. The results of the Kruskal-Wallis H test from § 5.1 indicate that there is a significant positive relationship between AMTIN and TRST. However, a mediation analysis within the SEM reveals that this effect is mediated by INFF. When assessing this mediating effect more closely in the context of our SEM, a small direct effect of AMTIN on TRST persists. Interestingly, in the context of the model, the stronger effect of AMTIN on TRST through INFF is positive, while the smaller but significant remaining direct effect is negative (-0.09*). We discuss this in more detail in § 7. Overall, **H2**, which conjectures a positive *total* (i.e., direct plus indirect) effect of AMTIN on TRST, is supported in our study.

The SEM's path coefficient concerning **H3** (0.78***) confirms that there is a statistically significant positive relationship between INFF and TRST—which confirms **H3**. This result provides a crucial individual piece of information in the context of the analysis of INFF as a mediator between AMTIN and TRST. As presumed in **H4**, the path coefficient between AILIT and INFF (0.59***) confirms the conjecture of a significant positive relationship between these two variables—therefore, **H4** is supported by our results. Similar to our findings w.r.t. the effect of AMTIN on TRST, the relationship between AILIT and TRST is also mediated by INFF. The analysis of effects within the full SEM confirms a strong indirect effect of AILIT on TRST through INFF, but the remaining direct effect of AILIT on TRST is not significant. Hence, the effect of AILIT on TRST is completely mediated by INFF. In conclusion, **H5**, which assumes a positive relationship between AILIT and TRST, is supported.

6 QUALITATIVE ANALYSIS

In this section, we aim to understand people's perceptions in more detail. To that end, we collected responses to open-ended questions regarding (i) what information SPs think they are missing (if any) to be able to judge whether the system behaves fairly, and (ii) SPs' perceptions of (in)appropriateness of the given explanations. These questions were part of each condition. The first and second author jointly coded the qualitative data according to grounded theory [21], i.e., codes evolved as we analyzed the data. In total, 982 text passages were coded over five coding sessions with MAXQDA [62]. The emerging themes from the collected responses are summarized in the following subsections. Every direct quote is provided with a unique identifier, introduced with the "#" symbol. Some responses contain statements w.r.t. multiple themes; hence, percentages do not always add up to 100%.

6.1 What information is missing?

For this question, we coded 421 text passages from SPs' responses to the open-ended question: *If you don't feel you received sufficient information to judge whether the decision-making procedures are fair or unfair, what information is missing?* We distinguish responses by condition and examine how many SPs felt that they received sufficient information (either by saying so explicitly or by not answering this question altogether). The latter is visually summarized in Figure 4.

(Base). Most SPs (79%) assigned to this condition felt that they did not receive sufficient information; 17% did not answer the question, and 4% explicitly stated that they are not missing any information. Little surprisingly, when asked which information they are missing, SPs were interested in knowing why the system made particular decisions; 37% of all responses contained statements substantially similar to this: *"All I know is that the loan was denied and not the reason why"* (#1315). Similarly, 30% of responses inquired about decision criteria that underlie the rejected loans: *"I have no way to know what references the company may or may not use to consolidate a decision about the eligibility of an individual for a particular loan, and therefore I might or might not find the procedures to be truly fair"* (#1260). 16% of responses also thought that decision-making procedures in general must be explained more thoroughly, arguing that *"everything to do with how they made their decision of whether to accept the loan or not [is missing]"* (#1234). Some SPs were more specific as to what explanations they need: 18% indicated that relevant factors of applicants would be helpful to know (#1259: *"To decide whether the decision-making procedures are fair or unfair, I probably would need to know how the client was economically and other factors such as criminal records"*); and 6% of responses requested counterfactual-type insights related to recourse, e.g., *"what he can do to try again"* (#1265).

(F). In the *(Factors)* condition, already 54% of SPs indicated that they received sufficient information. Of those who indicated that more information is needed, 15% are still interested in the "why" behind the rejections (#587: *"I think clearly spelled reason is missing instead of numbers"*). 15% still thought that more information w.r.t. decision criteria is needed. Interestingly, knowing what factors are used by the ADS raises further, more specific, questions as to why (i) these given factors are considered (#731: *"There needs to be more in depth explanations given as to why these factors are taken into consideration"*), and (ii) not others, e.g., *"how many loans have they*

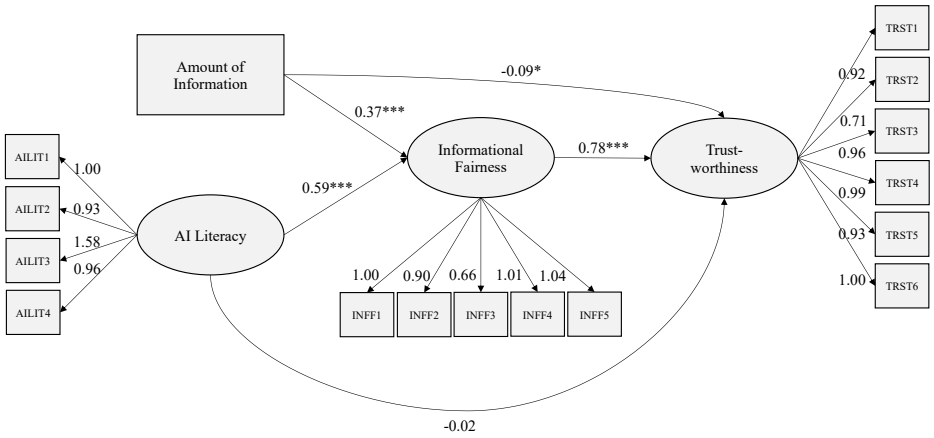


Figure 3: Full structural equation model (SEM) including measurement model; $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

taken out in the past, what is the money going to be spent on etc" (#663). Overall, 23% of SPs requested these justifications. Another 10% of responses indicated that it would be necessary to know how each factor impacts the final decision—both in terms of weighting (#474: *"What kind of value does each factor hold?"*) and monotonic relationships with the outcome (#602: *"The factors are told, but not which ones influenced the response positively of negatively."*) Finally, 3% are interested in counterfactual explanations, e.g., *"how the factors should differ for the application to be approved"* (#474).

(FFI). In this condition, only 37% of SPs requested further information. Among these, 15% still requested more information w.r.t. reasons why the ADS rejected the applications; and 17% felt that they still had not received sufficient information regarding decision-criteria (#677: *"There is not enough information about what thresholds have to be met to qualify for a loan."*) On a related note, 6% of SPs wanted to see more explanation as to why *"the [factor importance] ranking is the way it is"* (#764). Similar to the *(F)* condition, some SPs (10%) wanted to know why certain factors of the applicants are not being considered by the ADS. 3% of SPs still needed to know how exactly specific factors impact the final decision (#684: *"I don't know the significance level/weight assigned to [the factors]"*); and another 3% specifically requested counterfactual-type explanations. A newly occurring theme is w.r.t. communication of the explanations, as 3% requested *"less formal descriptions"* (#714).

(FFICF). In our condition with the highest amount of provided information, only 22% requested additional information. Generally speaking, responses are more dispersed compared to other conditions. Some SPs still alluded to missing justification w.r.t. the given selection and importance of relevant factors (overall 14%), and others (7%) still asked for more information on the relationship between certain input factors and the outcome (#796: *"Since I think gender being a factor is unfair, not knowing the degree to which it affects the outcome seems to be a deficiency."*) 6% of SPs were interested in the rationale behind providing given counterfactuals: *"The factors that could have changed the outcome [are revealed], but not the reason why those [...] factors would be needed. Ex.: Why would a rural area be more easily accepted?"* (#856) Interestingly, no SP requested additional information as to why the ADS rejected the applicants—as opposed to the other conditions. Yet, 11% still requested more information w.r.t. decision criteria, e.g., *"the thresholds that are required for a loan to be accepted"* (#800). 6% stated that processes were generally still not fully clear; however, some acknowledged that this might not necessarily be expedient, to begin with (#863: *"It's not clear how practically the priority system works, but I can understand it would be too hard to explain, and probably most of the people wouldn't understand it anyway."*)

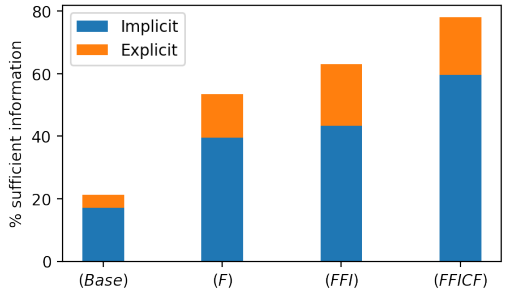


Figure 4: Percentage of responses indicating that study participants received sufficient information to judge whether the system's procedures are fair or unfair; either indicated explicitly in their responses, or implicitly by not answering the respective question.

6.2 (In)Appropriateness of individual explanations

We also asked SPs about their feelings of (in)appropriateness of isolated explanations, specific to the condition they were assigned to: *Why do you think {some factors, the order of factor importance, some counterfactual scenarios} are appropriate or inappropriate?* For that, we coded 561 text passages and summarized the main themes for each type of explanation.

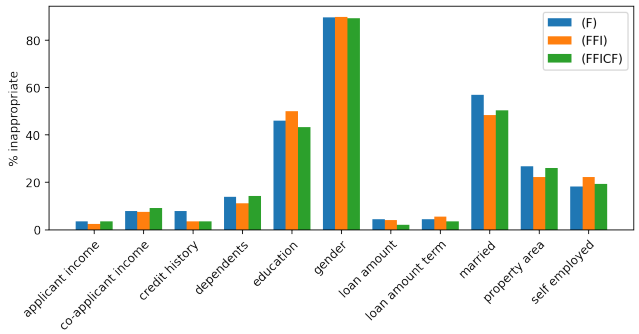


Figure 5: Inappropriate factors according to responses from study participants, broken down by condition.

Factors. Only 14% of responses explicitly stated that (at least a subset of) the factors considered by the ADS were appropriate—mostly those related to an applicant's financial situation (#602: *"Economic factors seem apropriate [sic] to me. Self employment sometimes involves risks and it is a relevant factor also."*) We also asked SPs to check specific factors they deem inappropriate—this is visualized (by condition) in Figure 5. Among responses w.r.t. inappropriate factors, two general themes emerged: 72% indicated that some factors are (causally) irrelevant for deciding on creditworthiness (#632: *"Some of the more social-oriented factors (ie education, gender, dependents) aren't necessarily indicative of someone's ability to pay back a loan"*), and 28% found the usage of certain factors (primarily *gender, education*, and *married*) morally wrong (#561: *"In the world we live, i dont [sic] think gender is something to even be at question, neither marriage."*) Interestingly, SPs often assumed that the sheer presence of a factor like *gender* means that it is being used with malicious intent: *"Gender can be somewhat problematic because all people deserve to have the right to the loan and not only men"* (#637), or, *"some factors like gender are plain racist to make a financial decision"* (#647).

Factor importance. Generally speaking, most SPs found the order of factor importance reasonably appropriate. Many responses resembled this: *"I may not agree with the placement of every single factor, but overall i think they are ranked appropriately"* (#695). Yet, 35% still suggested concrete changes w.r.t. the order of importance; particularly around assigning less weight to *education* and *marital status*. 14% were still entirely put off by the fact that *gender* or *marital status* were used in the decision-making process. However, learning that *gender* is the least important factor made many SPs feel better w.r.t. appropriateness of procedures (#510: *"It is appropriate. Gender should be considered the least and credit history is most important."*) One SP even suggested that *"gender could play a part in the decision making, but not a big one so it's good as it is"* (#751). (Recall that *gender* was ranked last in our explanation (see § 4.1).)

Counterfactual scenarios. 47% of coded responses indicated that the provided counterfactual scenarios are appropriate, e.g., endorsing that they *"are all financial and based on the ability of the loan to be paid back"* (#448). However, 20% questioned the effectiveness of adhering to some of the counterfactual recommendations; especially regarding suggested changes to *co-applicant income* or *property area*: *"These factors do not change the fact that an applicant can or can not pay his/her debt"* (#454). *Actionability* of counterfactual scenarios was another important theme: 9% overall addressed this, being appreciative that some counterfactual scenarios are explicitly actionable (#836: *"Changing the loan term is possible immediately"*) and disenchanted when not (#462: *"Some hardly achivable [sic] scenarios must be met to ensure the bank [will] be repayed [sic]."*) Some themes were addressed by fewer SPs but are highly interesting: one SP was, e.g., confused by the "direction" of suggested changes: *"Instead of a short loan amount term, it could be a bit longer"* (#778). Others were seemingly distracted by suggested changes that are (too) small: *"The incomes are so close to the required that it shouldn't matter"* (#447). Finally, some SPs hinted at potential inconsistencies between individual explanations: *"It seems odd that loan amount term is placed so low when it was one of the areas the individual could change to obtain the loan"* (#435).

7 DISCUSSION AND IMPLICATIONS

In this section, we link our quantitative results to qualitative insights to get a better understanding as to why certain effects were observed, and we analyze and discuss in more detail the findings from the fitted SEM. Finally, we allude to several implications of our work.

Connecting quantitative and qualitative findings. As observed in Tab. 1 (§ 5), both perceptions of informational fairness and trustworthiness increase as more explanations are provided to SPs—however, INFF at a much higher rate than TRST. Interestingly, many SPs in the *(Base)* condition, who do not receive any further explanations w.r.t. the inner workings of the ADS, do *not* find this "black-box" system to be overly problematic w.r.t. informational fairness: as can be seen in Fig. 2 (§ 5), SPs' responses for INFF are approx. equally distributed across ratings 1–4. This might be due to people's expectations; one SP simply stated that this *"seems to be standard practice"* (#1212) in terms of explaining ADS. From Tab. 2 (§ 5) we infer that providing relevant factors *(F)* to SPs does not significantly increase INFF. A likely reason for this observation is that SPs asked for significant follow-up information w.r.t. *how* the factors are used for decision-making. Both the differences for $(F) \rightarrow (FFI)$ and $(FFI) \rightarrow (FFICF)$ are significant for INFF. Considering the qualitative findings (§ 6.1), this seems little surprising as the complementary explanations (e.g., factor importance in *(FFI)* over *(F)*) were specifically requested by SPs.

While some explanations clearly helped SPs understand the given ADS better, they also reveal certain aspects that might be detrimental to people's trust. Similar to INFF, one might have expected to see lower ratings for TRST in the *(Base)* condition. Instead, SPs' responses for TRST are symmetrically distributed around the mean of 3 (see Fig. 2, § 5). Regarding marginal effects of explanations on TRST, we note that none of $(Base) \rightarrow (F)$, $(F) \rightarrow (FFI)$, or $(FFI) \rightarrow (FFICF)$ lead to statistically significant changes in SPs' perceptions. As for $(Base) \rightarrow (F)$, SPs' trust appears to be hampered by the experience that certain (presumably) inappropriate factors (e.g., *gender*) are being considered by the ADS. While the change $(F) \rightarrow (FFI)$ is marginally significant ($p = 0.052$) for TRST, we still suspect a certain attenuation due to SPs' disagreement with the relative importance ranking of certain factors like *education* and

married. On the other hand, from analyzing the qualitative statements, we might assume *gender* playing the least important role in the decision-making process had a positive effect on SPs' trust. As for $(FFI) \rightarrow (FFICF)$, we suspect that a potential positive effect of counterfactual explanations on perceived outcome control [50] might have been overshadowed by the fact that several SPs found some of the provided scenarios incomprehensible, ineffective, or unactionable.

Interpreting SEM results. In addition to confirming significant *total* effects (see Fig. 3, § 5) of the amount of information (AMTIN) on INFF (0.37^{***}) and TRST ($0.37 \cdot 0.78 - 0.09 = 0.20^{***}$), we also learn that SPs' (self-assessed) AI literacy (AILIT) is strongly related to INFF (0.59^{***}) and TRST (0.44^{***}), implying that we observe higher INFF and TRST ratings for higher AI-literacy people—given our study setup. Additionally, we see a strong positive relationship between INFF and TRST (0.78^{***}). The SEM also lets us decompose total effects of AMTIN and AILIT on TRST into direct and indirect (through the mediator INFF) effects (see Tab. 7, § E). We see, e.g., that the direct effect of AILIT on TRST (-0.02) is not significantly different from zero when INFF is acting as a mediator. Since the indirect effect AILIT→INFF→TRST is significantly positive (0.46^{***}), we observe a complete mediation of the effect of AILIT on TRST through INFF. A similar observation can be made for the effect of AMTIN on TRST: the total effect consists of a significantly *positive* indirect effect through INFF (0.29^{***}) as well as a small *negative* direct effect (-0.09^{*}). Hence, we conclude that increasing AMTIN does *not directly* increase TRST, but that the positive total effect stems from the strong indirect effect through INFF. This phenomenon is sometimes also referred to as *inconsistent mediation* [56, 79]. Future work should further investigate the link between INFF and TRST for other scenarios.

Implications. Our work has several implications for the design of automated decision systems and explanations thereof. Revealing to (potential) decision-subjects *what* information about them is used and *how* exactly individual factors affect the outcome is something that appears to go a long way towards facilitating informational fairness. We have also seen that many people require an understanding of (assumed) monotonic relationships between individual features and outcome (#856: *"We don't know if being married is a good or bad thing in this case."*) However, these types of global monotonic relationships cannot generally be derived from nonlinear ML models—something that has been discussed, e.g., in [91, 97, 111]. Employing inherently interpretable (e.g., linear) models might be a potential remedy.

We made a similar observation w.r.t. monotonicity for counterfactual explanations: people are put off when the "direction" of suggested change(s) contradicts commonly-held assumptions (e.g., if a *decrease* in income were suggested in order to get the loan). System designers must therefore pay close attention that counterfactual scenarios or general recommendations on recourse are intuitive, meaningful, and actionable. Regarding the latter, we have observed that certain factors are deemed actionable by some SPs and immutable by others. This poses further challenges w.r.t. individualizing explanations [63]; this is also relevant for people with different AI backgrounds as their perceptions differ. In general, however, counterfactual explanations appear to be effective in a way that they help people understand *"where [an] applicant fell short"* (#731). From the analysis of qualitative data (also confirmed quantitatively), we learned that SPs in the *(Base)* condition specifically requested explanations related to both factor importance and recourse / why the ADS decided negatively. This suggests the employment of both explanation types in a complementary fashion. Designers will have to ensure, however, that they are *consistent* with one another. For instance, people seem to expect that recommendations for recourse (e.g., that income should be increased) apply to the factors that are most important in the decision-making process. Since individual explanations are often automatically and independently generated, this poses a significant technical challenge. Our findings also suggest that informational fairness might be further increased by providing rejected loan applicants with a crisp statement in lay people's terms as to why they were denied. Finally, regarding the usage of sensitive information like *gender*, it should be clearly justified why and how (if at all) this information is used, and that this is not automatically to the disadvantage of marginalized groups; e.g., in the case of affirmative action [49].

8 LIMITATIONS AND OUTLOOK

We acknowledge limitations of our work that open up avenues for future studies. Firstly, we investigated only one setting where ADS are currently used to inform consequential decisions: lending. Our study design should be replicated and the results should be compared in different settings, e.g., hiring or university admissions, where the relevant factors will be significantly different. It would also be interesting to work with domain experts, as opposed to crowdworkers. Future work should further examine the complementarity and interplay of other explanation styles (e.g., case-based or demographic explanations [10]). Furthermore, our quantitative results (including SEM) are contingent upon the concrete instantiation of our ADS including the employed explanations, which limits our ability to generalize findings.

While we informally checked the model as well as the underlying data and all derived explanations so as to ensure behavior that might be representative of many real-world applications, it would be insightful to randomize different aspects about the model's quality and compare the results. More specifically, if we managed to construct—broadly speaking—a trustworthy ADS and an untrustworthy ADS, we would be able to contrast people's perceptions for either system. This would allow to derive insights w.r.t. *(un)warranted* perceptions, i.e., (i) are people *actually* able to spot problematic behavior of ADS, and (ii) do they trust the system if and only if the system is trustworthy? In fact, for an untrustworthy ADS, we would ideally expect that more explanations lead to *higher* informational fairness perceptions but to *lower* trust. If perceptions of trustworthiness increase *regardless* of the actual trustworthiness of the ADS, this would indicate serious issues around over-reliance [100] or automation bias [28, 37], and must be avoided by system designers at all costs.

We also acknowledge that our work does not explicitly take into account potential issues around information overload [9]: while we specifically examine situations where selected explanations convey complementary information, unsystematic provision of more and more explanations will likely have undesirable effects.

The authors suggest by no means that more information is always better. Finally, we hope that this work can serve as a stepping stone for further empirical research on the complementarity and interplay of different explanations and their effects on people's perceptions towards ADS.

ACKNOWLEDGMENTS

We thank our study participants as well as our anonymous reviewers, who helped improve this manuscript.

REFERENCES

[1] ACTICO. 2021. Automated credit decisioning for enhanced efficiency. https://www.actico.com/blog-en/automated-credit-decisioning-for-enhanced-efficiency/
[2] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.
[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. *ProPublica* (2016).
[4] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador Garcia, Sergio Gil-Lopez, Daniel Molina, Benjamins, et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.
[5] Vijay Arya, Rachel KE Bellamy, Pin-Yu Chen, Amit Dhurandhar, Michael Hind, Samuel C Hoffman, Stephanie Houde, Q Vera Liao, Ronny Luss, Aleksandra Mojsilović, et al. 2019. One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. *arXiv preprint arXiv:1909.03012* (2019).
[6] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2018. Fairness and machine learning. (2018). http://www.fairmlbook.org
[7] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2021. Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics* (2021).
[8] Kevin Bauer, Oliver Hinz, and Moritz von Zahn. 2021. Expl(AI)ned: The impact of explainable artificial intelligence on cognitive processes. (2021).
[9] David Bawden and Lyn Robinson. 2009. The dark side of information: Overload, anxiety and other paradoxes and pathologies. *Journal of Information Science* 35, 2 (2009), 180–191.
[10] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. '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*. 1–14.
[11] Leo Breiman. 2001. Random forests. *Machine Learning* (2001).
[12] Anna Brown, Alexandra Chouldechova, Emily Putnam-Hornstein, Andrew Tobin, and Rhema Vaithianathan. 2019. Toward algorithmic accountability in public services: A qualitative study of affected community perspectives on algorithmic decision-making in child welfare services. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–12.
[13] Michael W Browne and Robert Cudeck. 1992. Alternative ways of assessing model fit. *Sociological Methods & Research* 21, 2 (1992), 230–258.
[14] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.
[15] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*. IEEE, 160–169.
[16] France Bélanger, Janine S Hiller, and Wanda J Smith. 2002. Trustworthiness in electronic commerce: The role of privacy, security, and site attributes. *The Journal of Strategic Information Systems* 11, 3-4 (2002), 245–270.
[17] Dennis Carey and Matt Smith. 2016. How companies are using simulations, competitions, and analytics to hire. *Harvard Business Review* (2016).
[18] Simon Caton and Christian Haas. 2020. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053* (2020).
[19] Aaron Chalfin, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig, and Sendhil Mullainathan. 2016. Productivity and selection of human capital with machine learning. *American Economic Review* 106, 5 (2016), 124–127.
[20] David Chan. 2011. Perceptions of fairness. (2011).
[21] Kathy Charmaz and J Smith. 2003. Grounded theory. *Qualitative Psychology: A Practical Guide to Research Methods* 2 (2003), 81–110.
[22] Debdatta Chatterjee. 2019. Loan Prediction Problem Dataset. (2019). https://www.kaggle.com/altruistdelhite04/loan-prediction-problem-dataset
[23] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.
[24] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. 2019. Dark patterns of explainability, transparency, and user control for intelligent systems. In *IUI Workshops*, Vol. 2327.
[25] Jason A Colquitt, Donald E Conlon, Michael J Wesson, Christopher O L H Porter, and K Yee Ng. 2001. Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology* 86, 3 (2001), 425.
[26] Jason A Colquitt and Jessica B Rodell. 2011. Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives. *Academy of Management Journal* 54, 6 (2011), 1183–1206.
[27] Jason A Colquitt and Jessica B Rodell. 2015. Measuring justice and fairness. (2015).
[28] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.
[29] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2018. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64, 3 (2018), 1155–1170.
[30] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 275–285.
[31] Edelman. 2021. 2021 Edelman Trust Barometer: Trust in Technology. (2021). https://www.edelman.com/trust/2021-trust-barometer/trust-technology
[32] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).
[33] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
[34] European Union. 2016. General Data Protection Regulation. (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj
[35] Stefan Feuerriegel, Mateusz Dolata, and Gerhard Schwabe. 2020. Fair AI: Challenges and opportunities. *Business & Information Systems Engineering* 62 (2020), 379–384.
[36] M Lance Frazier, Paul D Johnson, Mark Gavin, Janaki Gooty, and D Bradley Snow. 2010. Organizational justice, trustworthiness, and trust: A multifoci examination. *Group & Organization Management* 35, 1 (2010), 39–76.
[37] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2014. Automation bias: Empirical results assessing influencing factors. *International Journal of Medical Informatics* 83, 5 (2014), 368–375.
[38] Randy Goebel, Ajay Chander, Katharina Holzinger, Freddy Lecue, Zeynep Akata, Simone Stumpf, Peter Kieseberg, and Andreas Holzinger. 2018. Explainable AI: The new 42?. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*. Springer, 295–303.
[39] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. 2018. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
[40] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. 2018. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*. 903–912.
[41] Nina Grgić-Hlača, Adrian Weller, and Elissa M Redmiles. 2020. Dimensions of diversity in human perceptions of algorithmic fairness. *arXiv preprint arXiv:2005.00808* (2020).
[42] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2018. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.
[43] Thomas Grote and Philipp Berens. 2020. On the ethics of algorithmic decision-making in healthcare. *Journal of Medical Ethics* 46, 3 (2020), 205–211.
[44] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)* 51, 5 (2018), 1–42.
[45] Joseph F Hair Jr, G Tomas M Hult, Christian Ringle, and Marko Sarstedt. 2016. *A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM)*. Sage Publications.
[46] Jeanne G Harris and Thomas H Davenport. 2005. Automated decision making comes of age. *MIT Sloan Management Review* 46, 4 (2005), 2–10.
[47] Will Douglas Heaven. 2020. Predictive policing algorithms are racist. They need to be dismantled. *MIT Technology Review* (2020).
[48] Paul Hitlin. 2016. Research in the crowdsourcing age: A case study. (2016).
[49] Harry Holzer and David Neumark. 2000. Assessing affirmative action. *Journal of Economic literature* 38, 3 (2000), 483–568.
[50] Pauline Houlden, Stephen LaTour, Laurens Walker, and John Thibaut. 1978. Preference for modes of dispute resolution as a function of process and decision control. *Journal of Experimental Social Psychology* 14, 1 (1978), 13–30.

[51] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for discrimination in algorithms delivering job ads. In *Proceedings of the Web Conference 2021*. 3767–3778.

[52] Infosys. 2019. How FinTechs can enable better support to FIs' credit decisioning? (2019). https://www.infosys.com/industries/financial-services/insights/documents/fintechs-fi-partners-credit-decision.pdf

[53] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 624–635.

[54] Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. 2016. Fairness in learning: Classic and contextual bandits. *arXiv preprint arXiv:1605.07139* (2016).

[55] Lena Kästner, Markus Langer, Veronika Lazar, Astrid Schomäcker, Timo Speith, and Sarah Sterz. 2021. On the relation of trust and explainability: Why to engineer for trustworthiness. *arXiv preprint arXiv:2108.05379* (2021).

[56] David A Kenny. 2015. Measuring model fit.

[57] René F Kizilcec. 2016. 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*. 2390–2395.

[58] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[59] Rex B Kline. 2015. *Principles and Practice of Structural Equation Modeling*. Guilford Publications.

[60] Sami Koivunen, Thomas Olsson, Ekaterina Olshannikova, and Aki Lindberg. 2019. Understanding decision-making in recruitment: Opportunities and challenges for information technology. *Proceedings of the ACM on Human-Computer Interaction* 3, GROUP (2019), 1–22.

[61] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American statistical Association* 47, 260 (1952), 583–621.

[62] Udo Kuckartz and Stefan Rädiker. 2019. *Analyzing Qualitative Data with MAXQDA*. Springer.

[63] Niklas Kuehl, Jodie Lobana, and Christian Meske. 2020. Do you comply with AI? – Personalized explanations of learning algorithms and their impact on employees' compliance behavior. *arXiv preprint arXiv:2002.08777* (2020).

[64] Todd Kulesza, Simone Stumpf, Margaret Burnett, Sherry Yang, Irwin Kwan, and Weng-Keen Wong. 2013. 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*. IEEE, 3–10.

[65] Nathan R Kuncel, David M Klieger, and Deniz S Ones. 2014. In hiring, algorithms beat instinct. *Harvard Business Review* (2014).

[66] Markus Langer, Kevin Baum, Cornelius J König, Viviane Hähne, Daniel Oster, and Timo Speith. 2021. Spare me the details: How the type of information about automated interviews influences applicant reactions. *International Journal of Selection and Assessment* 29, 2 (2021), 154–169.

[67] Markus Langer, Daniel Oster, Timo Speith, Holger Hermanns, Lena Kästner, Eva Schmidt, Andreas Sesing, and Kevin Baum. 2021. What do we want from explainable artificial intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence* 296 (2021), 103473.

[68] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human Factors* 46, 1 (2004), 50–80.

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

[70] Min Kyung Lee and Su Baykal. 2017. 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*. 1035–1048.

[71] Min Kyung Lee, Anuraag Jain, Hea Jin Cha, Shashank Ojha, and Daniel Kusbit. 2019. Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 182:1–182:26.

[72] Min Kyung Lee and Katherine Rich. 2021. Who is included in human perceptions of AI? Trust and perceived fairness around healthcare AI and cultural mistrust. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[73] Bruno Lepri, Jacopo Staiano, David Sangokoya, Emmanuel Letouzé, and Nuria Oliver. 2017. 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*. Springer, 3–24.

[74] Clayton Lewis and Robert Mack. 1982. The role of abduction in learning to use a computer system. (1982).

[75] Brian Y Lim, Anind K Dey, and Daniel Avrahami. 2009. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2119–2128.

[76] E Allan Lind, Robin I Lissak, and Donald E Conlon. 1983. Decision control and process control effects on procedural fairness judgments. *Journal of Applied Social Psychology* 13, 4 (1983), 338–350.

[77] Yang Liu, Goran Radanovic, Christos Dimitrakakis, Debmalya Mandal, and David C Parkes. 2017. Calibrated fairness in bandits. *arXiv preprint arXiv:1707.01875* (2017).

[78] Robert Long. 2021. Fairness in machine learning: Against false positive rate equality as a measure of fairness. *Journal of Moral Philosophy* 1 (2021), 1–30.

[79] David P MacKinnon, Amanda J Fairchild, and Matthew S Fritz. 2007. Mediation analysis. *Annu. Rev. Psychol.* 58 (2007), 593–614.

[80] Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics* (1947), 50–60.

[81] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635* (2019).

[82] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[83] Christoph Molnar. 2020. Interpretable machine learning. https://christophm.github.io/interpretable-ml-book/

[84] Deirdre K Mulligan, Joshua A Kroll, Nitin Kohli, and Richmond Y Wong. 2019. This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–36.

[85] Sue Newell and Marco Marabelli. 2015. Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of 'datification'. *The Journal of Strategic Information Systems* 24, 1 (2015), 3–14.

[86] Stefan Palan and Christian Schitter. 2018. Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance* 17 (2018), 22–27.

[87] Gabriele Paolacci, Jesse Chandler, and Panagiotis G Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision making* 5, 5 (2010), 411–419.

[88] Frank Pasquale. 2015. *The Black Box Society*. Harvard University Press.

[89] Emma Pierson. 2017. Demographics and discussion influence views on algorithmic fairness. *arXiv preprint arXiv:1712.09124* (2017).

[90] Prolific. 2022. Prolific vs. MTurk. (2022). https://prolific.co/prolific-vs-mturk/

[91] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1, 5 (2019), 206–215.

[92] Adam Satariano. 2020. British grading debacle shows pitfalls of automating government. *The New York Times* (2020). https://www.nytimes.com/2020/08/20/world/europe/uk-england-grading-algorithm.html

[93] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 99–106.

[94] Nadine Schlicker and Markus Langer. 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. In *Mensch und Computer 2021*. 325–329.

[95] Nadine Schlicker, Markus Langer, Sonja Ötting, Kevin Baum, Cornelius J König, and Dieter Wallach. 2021. What to expect from opening up 'black boxes'? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior* (2021), 106837.

[96] Jakob Schoeffer and Niklas Kuehl. 2021. Appropriate fairness perceptions? On the effectiveness of explanations in enabling people to assess the fairness of automated decision systems. In *Companion Publication of the 2021 Conference on Computer Supported Cooperative Work and Social Computing*. 153–157.

[97] Jakob Schoeffer, Niklas Kuehl, and Isabel Valera. 2021. A ranking approach to fair classification. In *ACM SIGCAS Conference on Computing and Sustainable Societies*. 115–125.

[98] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 2021. Perceptions of fairness and trustworthiness based on explanations in human vs. automated decision-making. *arXiv preprint arXiv:2109.05792* (2021).

[99] Michael Siegrist. 2008. Factors influencing public acceptance of innovative food technologies and products. *Trends in Food Science & Technology* 19, 11 (2008), 603–608.

[100] Linda J. Skitka, Kathleen Mosier, and Mark D. Burdick. 2000. Accountability and automation bias. *International Journal of Human Computer Studies* 52, 4 (2000), 701–717. https://doi.org/10.1006/ijhc.1999.0349

[101] Paul Slovic. 1987. Perception of risk. *Science* 236, 4799 (1987), 280–285.

[102] Paul Slovic, Baruch Fischhoff, and Sarah Lichtenstein. 1981. Perceived risk: Psychological factors and social implications. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences* 376, 1764 (1981), 17–34.

[103] Konrad Szczygieł. 2022. In Poland, a law made loan algorithms transparent. Implementation is nonexistent. (2022). https://algorithmwatch.org/en/poland-credit-loan-transparency/

[104] John W Thibaut and Laurens Walker. 1975. *Procedural Justice: A Psychological Analysis*. L. Erlbaum Associates.
[105] Sian Townson. 2020. AI can make bank loans more fair. *Harvard Business Review* (2020).
[106] Stefano Triberti, Ilaria Durosini, and Gabriella Pravettoni. 2020. A "third wheel" effect in health decision making involving artificial entities: A psychological perspective. *Frontiers in Public Health* 8 (2020).
[107] Alarith Uhde, Nadine Schlicker, Dieter P Wallach, and Marc Hassenzahl. 2020. Fairness and decision-making in collaborative shift scheduling systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.
[108] Kees Van den Bos, Henk AM Wilke, and E Allan Lind. 1998. When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology* 75, 6 (1998), 1449.
[109] Oleksandra Vereschak, Gilles Bailly, and Baptiste Caramiaux. 2021. How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–39.
[110] Ruotong Wang, F Maxwell Harper, and Haiyi Zhu. 2020. 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*. 1–14.
[111] Serena Wang and Maya Gupta. 2020. Deontological ethics by monotonicity shape constraints. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 2043–2054.
[112] Allison Woodruff, Sarah E Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A qualitative exploration of perceptions of algorithmic fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
[113] Yu-Qian Zhu and Houn-Gee Chen. 2012. Service fairness and customer satisfaction in internet banking: Exploring the mediating effects of trust and customer value. *Internet Research* (2012).