Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making

JAKOB SCHOEFFER, Karlsruhe Institute of Technology (KIT), Germany

MARIA DE-ARTEAGA*, University of Texas at Austin, USA

NIKLAS KUEHL*, Universität Bayreuth, Germany

Proponents of explainable AI have often argued that it constitutes an essential path towards algorithmic fairness. Prior works examining these claims have primarily evaluated explanations based on their effects on humans' perceptions, but there is scant research on the relationship between explanations and distributive fairness of AI-assisted decisions. In this paper, we conduct an empirical study to examine the relationship between feature-based explanations and distributive fairness, mediated by human perceptions and reliance on AI recommendations. Our findings show that explanations influence fairness perceptions, which, in turn, relate to humans' tendency to adhere to AI recommendations. However, our findings suggest that such explanations do not enable humans to discern correct and wrong AI recommendations. Instead, we show that they may affect reliance irrespective of the correctness of AI recommendations. Depending on which features an explanation highlights, this can foster *or* hinder distributive fairness: when explanations highlight features that are task-irrelevant and evidently associated with the sensitive attribute, this prompts overrides that *counter* stereotype-aligned AI recommendations. Meanwhile, if explanations appear task-relevant, this induces reliance behavior that *reinforces* stereotype-aligned errors. These results show that feature-based explanations are not a reliable mechanism to improve distributive fairness, as their ability to do so relies on a human-in-the-loop operationalization of the flawed notion of "fairness through unawareness". Finally, our study design provides a blueprint to evaluate the suitability of other explanations as pathways towards improved distributive fairness of AI-assisted decisions.

ACM Reference Format:

Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. 2023. Explanations, Fairness, and Appropriate Reliance in Human-AI Decision-Making. In *CHI TRAIT '23.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

Please also refer to the full version of this manuscript at https://arxiv.org/abs/2209.11812.

1 INTRODUCTION

AI-based systems are commonly used for informing decision-making in consequential areas, where they provide human decision-makers with decision recommendations. The human is then tasked to decide whether to adhere to this recommendation or override it. Researchers, policy makers, and activists have expressed concern over the risk of algorithmic bias resulting in unfair decisions. As a response, many have advocated for the need for explanations, under the assumption that they can enable humans to mitigate algorithmic bias. However, there is often ambiguity regarding what it means for the human to mitigate bias, and a lack of evidence studying whether this is possible. In this paper, we posit that when concerned with distributive fairness, the central mechanism that should be studied is the type of reliance[1] fostered by the explanations and its effect on disparities in AI-assisted decisions.

[*] Both authors contributed equally to this research.

[1] We use *reliance* as an umbrella term for people's behavior of adhering to or overriding AI recommendations [41].

Our work. In this work, we examine the effects of feature-based explanations on people's ability to enhance distributive fairness, mediated by fairness perceptions and reliance on AI recommendations. To empirically study this, we conduct a randomized online experiment and assess differences in perceptions and reliance behavior when participants see and do not see explanations, and when these explanations indicate the use of sensitive features in predictions vs. when they indicate the use of task-relevant features. We operationalize this in the context of occupation prediction, for which we train two AI models with access to different vocabularies. We randomly assign participants to one of two groups and ask them to predict whether bios belong to professors or teachers: for one group, recommendations come from an AI model that uses *gendered* words for predicting occupations, whereas in the other group the AI model uses *task-relevant* words. Both AI models provide the same recommendations, and their distribution of errors is in line with societal stereotypes and the expected risks of bias characterized in previous research [21]. Participants in both conditions are provided with explanations that visually highlight the most predictive words of their respective AI models. We test for differences in perceptions and reliance behavior across conditions, and measure gender disparities for different types of errors.

Findings and implications. **First**, we do not observe any significant differences in decision-making accuracy across conditions, i.e., participants did not make more (or less) accurate decisions in the conditions with explanations compared to the baseline without explanations. Since participants were incentivized to make accurate predictions, this implies that explanations did not enable them to make better decisions with respect to accuracy.

Second, no condition improved participants' likelihood to override mistaken vs. correct AI recommendations, but conditions did affect the likelihood to override recommendations conditioned on the predicted occupation: we see that participants in the *gendered* condition overrode more AI recommendations to *counter* existing societal stereotypes (e.g., by predicting more women to be professors), irrespective of whether the prediction was correct. Simultaneously, when explanations highlight only task-relevant words, reliance behavior *reinforced* stereotype-aligned decisions; e.g., by predicting more men to be professors, even when they are teachers.

This, **third,** has implications for distributive fairness: by prompting reliance behavior that either counters or reinforces societal stereotypes embedded in AI recommendations, (i) explanations that highlight gendered words led to a *decrease* in error rate disparities (i.e., fostering distributive fairness), whereas (ii) explanations that highlight task-relevant words led to an *increase* in error rate disparities (i.e., hindering distributive fairness). These findings emphasize the need to differentiate between improved distributive fairness that is driven by a shift in the types of errors vs. improvements that are driven by humans' ability to override mistaken AI recommendations.

Fourth, we confirm prior works' findings by observing that people's fairness perceptions are significantly lower when explanations highlight gendered words compared to task-relevant words, and empirically show that people override significantly more AI recommendations when their fairness perceptions are low. However, we observe that perceptions solely relate to the quantity of overrides and do *not* correlate with an ability to discern correct and wrong AI recommendations. Hence, fairness perceptions are only a meaningful proxy for distributive fairness when it is desirable to override the AI based on its use of sensitive features. However, prior research has shown that the idea of "fairness through unawareness" is neither a necessary nor sufficient condition for distributive fairness [4, 18, 25, 40, 52, 57].

2 BACKGROUND

2.1 Explanations of AI

Goals of explanations. AI systems are becoming increasingly complex and opaque, and researchers and policymakers have called for explanations to make AI systems more understandable to humans [27, 44, 50]. Apart from the central aim of facilitating human understanding, prior research has formulated a wealth of different desiderata that explanations are to provide [44]. Relevant to our work are several desiderata that concern explanations as an alleged means for better and fairer AI-assisted decision-making [1, 24]; we speak to this in more detail in § 2.2 and § 2.3.

Types of explanations. The scientific literature distinguishes explanations that aim at explaining individual predictions (*local* explanations) from those that aim at explaining the general functioning of an AI model (*global* explanations) [35]. So-called *local model-agnostic* explanations, such as LIME [63] or SHAP [48], have gained popularity in the literature [1]. When applied to text data, these methods can generate a highlighting of important words for text classification. In this work, our focus is on these feature-based explanations, and we use LIME in our experiments.

Criticism of explanations. Most desiderata for explanations are insufficiently studied or met with inconclusive or seemingly contradictory empirical findings [14, 22, 44]. A major line of criticism stems from the fact that explanations can mislead people: Chromik et al. [16] discuss situations where system designers may create interfaces or misleading explanations to purposefully deceive more vulnerable stakeholders like auditors or decision-subjects; e.g., through *adversarial attacks* on explanation methods [23, 43, 60, 73]. In the context of fairness, feature-based explanations may or may not highlight the usage of sensitive information (e.g., on gender) by an AI system, which has been shown to be an unreliable indicator of a system's actual fairness [4, 18, 25, 40, 52, 57]. We address this in more detail in § 2.3 due to its importance for our work.

2.2 Explanations and (appropriate) reliance

Effects on accuracy. It has been argued that explanations are an enabler for better AI-assisted decision-making [5, 24, 29, 39, 61]. A recent meta-study [66] on the effectiveness of explanations, however, implies that explanations in most empirical studies did not yield any significant benefits with respect to decision-making accuracy; e.g., in [2, 30, 47, 51, 79]. On the other hand, Lai and Tan [42] find that explanations greatly enhance decision-making accuracy for the case of deception detection. An accuracy increase through explanations may, however, solely be due to (i) an overall increase in adherence to a high-accuracy AI, or (ii) an overall decrease in adherence to a low-accuracy AI.

Effects on reliance. In the context of AI-assisted decision-making, *appropriate reliance* is typically understood as the behavior of humans of overriding wrong AI recommendations and adhering to correct ones [56, 67]. Humans' ability to override mistaken recommendations has also been referred to as *corrective overriding* [28]. When considering the role of explanations in fostering appropriate reliance, it has been claimed that "transparency mechanisms also function to help users learn about how the system works, so they can evaluate the *correctness* of the outputs they experience and identify outputs that are incorrect" [61]. Empirical evidence, however, is less clear: several studies have found that explanations can be detrimental to appropriate reliance [6, 11, 41, 59, 68, 76], when they increase or decrease humans' adherence to AI recommendations regardless of their correctness.

Conflation of reliance and trust. Many studies have treated reliance and trust interchangeably [41], sometimes calling reliance a "behavioral trust measure" [54]. However, definitions of *trust* are often inconsistent [38, 45, 54], which makes empirical findings challenging to compare. More importantly, trust and reliance are different constructs [41]: reliance is the *behavior* of adhering to or overriding AI recommendations, whereas trust is a subjective *attitude* regarding the whole system, which builds up and develops over time [55, 62, 77]. It has been argued that trust may impact reliance [26, 45, 71], but trust is not a sufficient requirement for reliance when other factors, such as time constraints, perceived risk, or self-confidence, impact decision-making [28, 45, 64]. In our work, we directly measure participants' reliance behavior and do not assume an equivalence between reliance and trust.

2.3 Explanations and fairness

Goal of promoting algorithmic fairness. It is known that AI systems can issue predictions that may result in disparate outcomes or other forms of injustices for certain socio-demographic groups—especially those that have been historically marginalized [8, 12, 20, 37]. When AI systems are used to inform consequential decisions, it is important that a human can override problematic recommendations. To that end, the literature has often framed explanations as an important pathway towards improving algorithmic fairness [5, 19, 24, 44]. Grounded on the organizational justice literature [17, 31], researchers distinguish different notions of algorithmic fairness, among which are (*i*) *distributive fairness*, which refers to the fairness of decision outcomes [78], and (*ii*) *procedural fairness*, which refers to the fairness of decision-making procedures [46]. Distributive fairness is typically measured in terms of statistical metrics such as parity in error rates across groups [7, 15]; which is closely related to notions like *equalized odds* or *equal opportunity* [36]. Importantly, there is no conclusive evidence showing that explanations lead to fairer decisions, and it remains unclear *how* explanations may enable this [44].

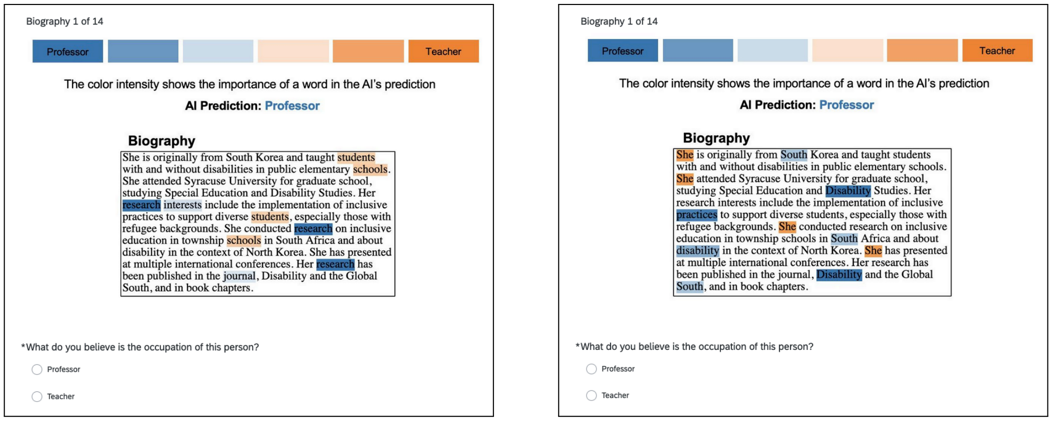
Fairness perceptions. Prior work at the intersection of fairness and explanations has primarily focused on assessing how people *perceive* the fairness of AI systems [41, 74]. Empirical findings are mostly inconclusive, stressing that fairness perceptions depend on many factors, such as the explanation style [9, 24], the amount of information provided [69], the use case [3], user profiles [24], or the decision outcome [72]. Surprisingly, few works have examined downstream effects of fairness perceptions on AI-assisted decisions. Our work complements prior studies by centering distributive fairness and how it relates to fairness perceptions.

Perceptions and sensitive features. A series of prior studies have found that knowledge about the features that an AI model uses influences people's fairness perceptions [32–34, 52, 58, 75]. This type of information is, e.g., conveyed by feature-based explanations like LIME. Specifically, people tend to be averse to the use of what is considered *sensitive* information, e.g., gender or race [18, 32–34, 52, 58, 69]. Interestingly, people's perceptions towards these features change after they learn that "blinding" the AI to these features can lead to *worse* outcomes for marginalized groups. In fact, it is known that prohibiting an AI model from using sensitive information is neither a necessary nor sufficient requirement for fair decision-making [4, 18, 25, 40, 52, 57]. In this work, we build upon these findings on the interplay of fairness perceptions and sensitive features. Concretely, we assess differences in reliance behavior when participants see explanations that highlight task-relevant vs. sensitive features, and derive implications for distributive fairness.

3 STUDY DESIGN

3.1 Task and dataset

Task. Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online [10, 65]. An important task herein is to determine someone's occupation, which is a prerequisite for advertising job openings or recruiting people for adequate positions. This information may not be readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI-based systems, it is susceptible to gender bias and discrimination [10, 21, 65]. De-Arteaga et al. [21] show that these biases can manifest themselves in error rate disparities between genders, and



Biography 1 of 14

The color intensity shows the importance of a word in the AI's prediction

AI Prediction: Professor

Biography

She is originally from South Korea and taught students with and without disabilities in public elementary schools. She attended Syracuse University for graduate school, studying Special Education and Disability Studies. Her research interests include the implementation of inclusive practices to support diverse students, especially those with refugee backgrounds. She conducted research on inclusive education in township schools in South Africa and about disability in the context of North Korea. She has presented at multiple international conferences. Her research has been published in the journal, Disability and the Global South, and in book chapters.

*What do you believe is the occupation of this person?

○ Professor

○ Teacher

Task-relevant condition

Biography 1 of 14

The color intensity shows the importance of a word in the AI's prediction

AI Prediction: Professor

Biography

She is originally from South Korea and taught students with and without disabilities in public elementary schools. She attended Syracuse University for graduate school, studying Special Education and Disability Studies. Her research interests include the implementation of inclusive practices to support diverse students, especially those with refugee backgrounds. She conducted research on inclusive education in township schools in South Africa and about disability in the context of North Korea. She has presented at multiple international conferences. Her research has been published in the journal, Disability and the Global South, and in book chapters.

*What do you believe is the occupation of this person?

○ Professor

○ Teacher

Gendered condition

Fig. 1. **Exemplary bio.** A bio of a woman professor, both in the *task-relevant* (left) and the *gendered* (right) condition.

that error rate disparities are correlated with gender imbalances in occupations. For instance, women surgeons are significantly more often misclassified than men surgeons because the occupation *surgeon* is heavily men-dominated. Similar disparities occur, among others, for professors and teachers. Interestingly, the disparate impact on people persists when the AI model does *not* consider explicit gender indicators (e.g., pronouns) [21]. Such misclassifications in hiring have tremendous repercussions for affected people because they may be systematically excluded from exposure to relevant opportunities. In our study, we instantiate an AI-assisted decision-making setup where participants see short textual bios and are asked—with the help of an AI recommendation—to predict whether a given bio belongs to a professor or a teacher. Professors are historically a men-dominated occupation, whereas teachers have been mostly associated with women [50, 80, 81].

Dataset. We use the publicly available BIOS dataset, which contains approximately 400,000 online bios for 28 different occupations from the Common Crawl corpus [21]. For each bio in the dataset we know the gender of the corresponding person and their true occupation. In line with current demographics and societal stereotypes [50, 80, 81], we have more men (55%) than women (45%) bios of professors and more women (60%) than men (40%) bios of teachers.

3.2 Experimental setup

General setup. Participants see 14 bios one by one, each including the AI recommendation as well as an explanation highlighting the most predictive words. We also include a baseline condition without explanations. The crux of our experimental design is that we assign participants to conditions where they see recommendations and explanations either from (i) an AI model that uses *task-relevant* features, or (ii) an AI model that uses *gendered* (i.e., sensitive) features. An exemplary bio including explanations is depicted in Figure 1. Note that the AI predictions and explanations stem from actual AI models that agree in their predictions for the 14 bios. Participants in each condition first complete the task of predicting occupations for 14 bios, and—if assigned to a condition with explanations—answer questions regarding their fairness perceptions, similar to other human-AI studies [9, 49, 70]. We ask about fairness perceptions

Table 1. **Overview of the six types of scenarios employed in our study.** Our study includes 14 bios of different scenarios.

Gender of bio	True occupation	AI recommendation	AI correct?	Acronym	#Bios
Woman	**T**eacher	**T**eacher	✓	WTT	3
Woman	**P**rofessor	**T**eacher	✗	WPT	3
Woman	**P**rofessor	**P**rofessor	✓	WPP	1
Man	**T**eacher	**T**eacher	✓	MTT	1
Man	**T**eacher	**P**rofessor	✗	MTP	3
Man	**P**rofessor	**P**rofessor	✓	MPP	3

Table 2. **Different types of reliance on AI recommendations.** We distinguish four types of reliance in AI-assisted decision-making.

	Human adherence to AI	Human overriding of AI
AI correct	Correct adherence	Detrimental overriding
AI wrong	Detrimental adherence	Corrective overriding

after the task is completed, so as to prevent these questions from moderating reliance behavior [13]. Finally, we measure and confirm that participants thought consistently about the distinction *professor vs. teacher* between conditions.

Task completion. Figure 1 shows the interface that participants in the *task-relevant* as well as the *gendered* condition see during the completion of the task. Explanations involve a dynamic highlighting of important words for either AI model (*task-relevant* and *gendered*); and they also indicate whether certain words are indicative of *professor* (blue) or *teacher* (orange). Lastly, the color intensity shows the importance of a given word in the AI's prediction. Participants in the *task-relevant* and the *gendered* condition are confronted with 14 bios similar to the one in Figure 1, whereas participants in the baseline condition are shown the same set of bios without highlighting of words, and the AI prediction without color coding. Recall that the AI recommendations are identical across conditions. For each instance, participants are asked to make a binary prediction about whether they believe that a given bio belongs to a professor or a teacher. We incentivize accurate predictions through bonus payments.

In order to be able to assess differences in reliance behavior across conditions, participants see a mix of cases where the AI is correct and where it is wrong. More specifically, we distinguish six types of scenarios that make up the 14 bios that participants see—they are summarized in Table 1. We distinguish these scenarios based on three dimensions: (*i*) gender of the person associated with a bio; (*ii*) true occupation of that person; (*iii*) AI recommended occupation. We show 3 cases each of correctly recommended women teachers (WTT) and men professors (MPP), as well as 3 cases of wrongly recommended women professors (WPT) and men teachers (MTP). To preempt the misconception that the AI always recommends *teacher* for women and *professor* for men, we also include one case each of WPP and MTT. We do not consider scenarios where women teachers are classified as professors, or where men professors are classified as teachers, because our focus is on the errors that are more likely to occur in practice [21]. In our assessment of reliance behavior, we distinguish four cases, as depicted in Table 2.

Task-relevant and gendered classifiers. We train two classifiers with access to mutually disjoint vocabularies as predictors. The *task-relevant* vocabulary consists of words that appear on average—for both men and women—more often in professor or teacher bios than in any of the 26 remaining occupations in the BIOS dataset. The resulting vocabulary consists of words such as *faculty*, *kindergarten*, or *phd*. The *gendered* vocabulary, on the other hand, consists of words that are most predictive of gender, which includes, apart from gender pronouns and words such as *husband* and *wife*, words like *dance*, *art*, or *engineering*, which are not evidently gendered but highly correlated with the

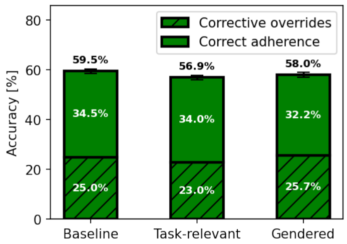


Fig. 2. **Accuracy by condition.** Accuracy is slightly lower when explanations are provided, compared to the baseline. Error bars represent standard errors.

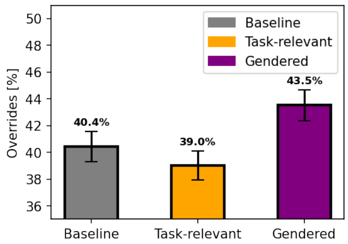


Fig. 3. **Overrides by condition.** Overrides are higher in the *gendered* condition vs. *task-relevant* and the baseline.

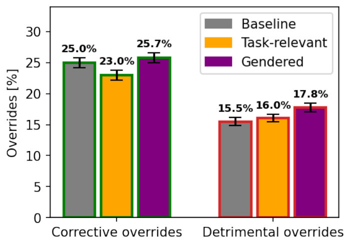


Fig. 4. **Overriding behavior.** Both corrective (left) and detrimental (right) overrides are highest in the *gendered* condition.

sensitive attribute. Finally, we train two classifiers on a balanced set of professor and teacher bios, and we employ the `TextExplainer` from LIME [63] to generate dynamic explanations with highlighting of predictive words.

Selection of bios. Recall that participants see 14 bios as outlined in Table 1. These bios are taken from a random holdout set that our two classifiers make predictions on. Specifically, we choose bios that are reasonably similar in length and where both classifiers yield the same predicted occupation as well as similar prediction probabilities. We also require that these predictions probabilities for a bio must not be too high, which aims at eliminating bios that are "too easy" to classify. The authors then manually screened the remaining contenders to settle on the final 14 bios.

Data collection. Our study has received clearance from an institutional ethics committee. Participants were recruited via `Prolific` [53]. We required participants to be at least 18 years of age, and to be fluent in English. We also sampled approximately equal amounts of men and women; no other pre-screeners were applied. After consenting to the terms of our study, participants were then randomly and in equal proportions assigned to one of our three conditions and asked to complete the respective questionnaire. Overall, we recruited 600 lay people through `Prolific`.

4 RESULTS AND ANALYSIS

4.1 Effects of explanations on accuracy and overriding behavior

Effects on accuracy. First, we examine how accuracy may be different between the baseline and the conditions with explanations, *task-relevant* and *gendered*. Recall that participants were incentivized through bonus payments to accurately predict occupations. Figure 2 suggests that **explanations did not aid AI-assisted decision-making when measured in terms of accuracy.**

Effects on overriding behavior. In Figures 3 and 4, we see that participants in the *gendered* condition overrode more AI recommendations than in the *task-relevant* condition and the baseline. From Figure 4 we further conclude that *both* corrective *and* detrimental overrides are highest in the *gendered* condition, with detrimental overrides being significantly higher than in the *task-relevant* condition and the baseline. We interpret this increase in overrides further in § 4.2. In the *task-relevant* condition, we see that overall overrides are lowest across conditions (Figure 3), with corrective overrides being significantly lower than the baseline (Figure 4). Overall, we conclude that people's reliance behavior is affected by how the AI explains its recommendations; notably, people overrode AI recommendations significantly more often when explanations highlight features that are evidently associated with gender. Across conditions, we also infer

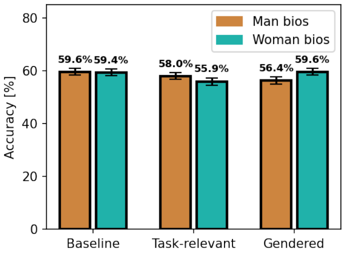


Fig. 5. **Accuracy by condition and gender of bio.** Explanations do not increase accuracy for either men or women bios. *Task-relevant* explanations decrease accuracy for women; *gendered* explanations decrease accuracy for men.

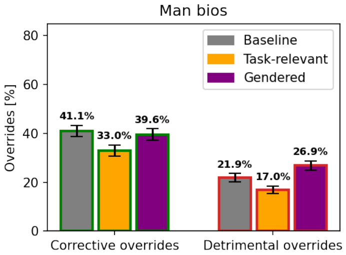


Fig. 6. **Overriding behavior for men bios.** *Task-relevant* explanations decrease both corrective and detrimental overrides for men bios, compared to the baseline; whereas *gendered* explanations only increase detrimental overrides.

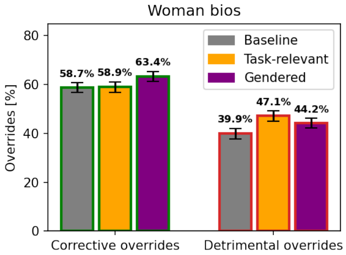


Fig. 7. **Overriding behavior for women bios.** *Gendered* explanations increase both corrective and detrimental overrides over the baseline; *task-relevant* explanations increase detrimental overrides.

from Figure 4 that participants generally performed more corrective than detrimental overrides, and that **the ability to perform corrective vs. detrimental overrides did not improve through the provision of explanations.**

4.2 Interplay between explanations, reliance, and distributive fairness

Accuracy by gender. Consistent with our findings at the aggregated level (see Figure 2), we do not observe any accuracy improvements through explanations over the baseline in Figure 5, neither for men nor women bios. We see that accuracy for men and women bios is approximately equal in the baseline condition, and that accuracy in the *task-relevant* condition is relatively lower for women bios, whereas in the *gendered* condition it is relatively lower for men, compared to the baseline. This means that **both in the *task-relevant* and the *gendered* condition, explanations did not enable people to improve decision-making accuracy, neither for men nor women bios.**

Types of overrides by gender and occupation. When looking at effects of explanations on overriding behavior by gender in Figures 6 and 7, no intervention improved participants' ability to perform corrective vs. detrimental overrides of AI recommendations, neither for men nor women bios. This is consistent with our findings at the aggregate level (see Figure 4). Notably, we see that in the *gendered* (Figure 6) and the *task-relevant* (Figure 7) condition detrimental overrides increase over the baseline, whereas corrective overrides remain unchanged.

From Figures 6 and 7 we also see that participants generally overrode more recommendations for women than men bios. However, this is not due to gender: we show that there are more overrides for men teachers predicted by the AI model as teachers than for women professors predicted as professors. Together, these results suggest that **people were overall more prone to do promoting[2] overrides**; which means that participants overrode AI recommendations more often when someone was suggested to be a teacher vs. a professor.

Importantly, people's likelihood to override conditioned on gender and predicted occupation did vary across conditions. By virtue of our study design, we are able to observe stereotype-countering[3] corrective overrides, and both stereotype-aligned and stereotype-countering detrimental overrides. As explained in § 3.2, the motivation for this design is our focus on studying whether explanations allow humans to correct for stereotype-aligned wrong AI predictions,

[2] We assume here that the occupation of *professor* is associated with a higher societal status than that of *teacher*. Hence, *promoting* refers to predicting someone to be a professor, whereas *demoting* means to predict someone to be a teacher.

[3] Recall that societal stereotypes typically associate men with being professors and women with being teachers [50].

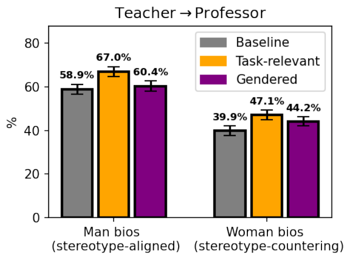


Fig. 8. **Bios wrongly classified by humans as professor.** Promotions increase for both men and women bios in the *task-relevant* condition, compared to the baseline; and they only increase for women bios in the *gendered* condition.

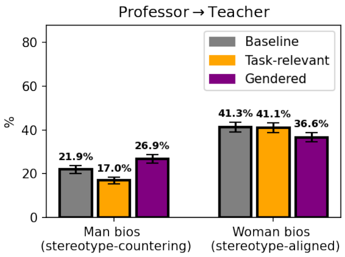


Fig. 9. **Bios wrongly classified by humans as teacher.** In the *gendered* condition, demotions increase for men bios and decrease for women bios, compared to the baseline; and they only decrease for men bios in the *task-relevant* condition.

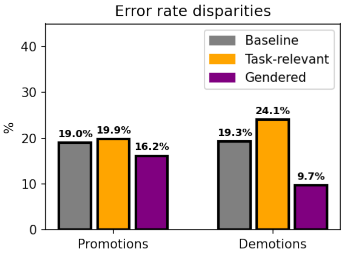


Fig. 10. **Disparities in error rates across gender.** *Gendered* explanations decrease both disparities in promotions and demotions between genders, compared to the baseline; *task-relevant* explanations increase disparities.

which would be the most frequent errors of an occupation prediction model that exhibits gender bias [21]. We see that in the *task-relevant* condition, people perform fewer corrective overrides for men and the same amount for women in comparison to the baseline, as shown in Figures 6 and 7. Meanwhile, in the *gendered* condition participants perform more corrective overrides for women and the same amount of such overrides for men. This means that **participants in the *gendered* condition were more likely to perform stereotype-countering corrective overrides than in the baseline, while participants in the *task-relevant* condition were less likely to do so.**

As for detrimental overrides, we see that they increase in the *gendered* condition for both men and women bios, compared to the baseline (Figures 6 and 7). Considering that we do not observe differences in stereotype-aligned detrimental overrides between conditions, we infer that people in the *gendered* condition performed more stereotype-countering detrimental overrides, by predicting more men to be teachers and women to be professors. It is noteworthy that when contrasting corrective and detrimental overrides, we observe that **no condition improved participants' ability to make stereotype-countering *corrective* overrides vs. stereotype-countering *detrimental* overrides**. In the *gendered* condition, this means that participants became more likely to override an AI recommendation when it predicted that a woman is a teacher, irrespective of her true occupation. **Overall, we observe reliance behavior in the *gendered* condition that counters societal stereotypes, whereas in the *task-relevant* condition people tend to rely on AI recommendations in a way that reinforces stereotypes.**

Implications for distributive fairness. We now examine how the observed reliance behavior relates to distributive fairness with respect to disparities in errors between men and women. First, we note that in the baseline condition, people tend to make more errors that promote men vs. women (58.9% vs. 39.9% in Figure 8), and demote women more than men (41.3% vs. 21.9% in Figure 9). Note that in the case of men, promoting behavior is stereotype-aligned, whereas in the case of women such behavior is stereotype-countering; and vice versa for demoting behavior. The resulting absolute error rate disparities between men and women for the baseline are, hence, 19.0% (promotions) and 19.3% (demotions), as depicted in Figure 10. From the previous paragraph we know that people in the *task-relevant* condition showed a tendency of reinforcing stereotypes, meaning that promotions of men increased more than those of women, which increased disparities in promotions even further over the baseline (Figure 10, left). Similarly, demotions of men decreased much more than demotions of women, leading to increased disparities in demotions over the baseline

(Figure 10, right). In conclusion, we note that **people's stereotype-aligned reliance behavior in the *task-relevant* condition exacerbated existing disparities in the baseline condition and, hence, hindered distributive fairness.**

In the *gendered* condition, on the other hand, people countered stereotypes, meaning that promotions of women increased more than for men, reducing existing disparities (Figure 10, left). The most drastic reduction in disparities happens for demotions (Figure 10, right), since demotions *increased* for men and *decreased* for women (Figure 9). This results in a reduction of disparities in demotions from 19.3% (baseline) to 9.7% (*gendered* condition). Hence, **people's stereotype-countering reliance behavior in the *gendered* condition mitigated existing disparities and, hence, fostered distributive fairness.** It is important to stress that while disparities in error types decreased in the *gendered* condition compared to the baseline, this was mostly due to a shift in the types of errors, as opposed to an increased ability to override mistaken AI recommendations.

4.3 The role of fairness perceptions

Effects of explanations on fairness perceptions. We measure three items regarding fairness perceptions on 5-point Likert scales, ranging from 1 (unfair) to 5 (fair). We then take the average of the three item ratings for each participant to obtain a single measure of fairness perceptions. We find that participants in the *task-relevant* and *gendered* conditions have significantly different perceptions of fairness towards the AI model. Concretely, we observe $M_{rel} = 3.53$ ($SD_{rel} = 0.85$) in the *task-relevant* condition, and $M_{gen} = 2.54$ ($SD_{gen} = 0.98$) in the *gendered* condition. Overall, we confirm prior works' findings and conclude that **the AI system was perceived as significantly less fair when explanations point at the use of sensitive features compared to cases where explanations point at task-relevant features.**

Relationship of fairness perceptions with overriding behavior. When we look at people's overriding behavior as a function of their fairness perceptions, we find an overall strong negative relationship between fairness perceptions and overriding of AI recommendations, i.e., participants overrode the AI more often when their fairness perceptions were lower. This negative relationship is consistent in both the *task-relevant* and the *gendered* condition, and it also persists when we disentangle corrective and detrimental overrides at the aggregate level. Figure 11 shows the relationship of overrides—both corrective, detrimental, and total—as a function of fairness perceptions for the *gendered* condition. Dots represent mean values of overrides for a given level of perceptions, and lines are OLS regressions fitted on the original data. We observe that as participants overrode more AI recommendations in the *gendered* condition, the rates at which corrective and detrimental overrides increase are approximately equal. We conclude that **people's fairness perceptions are associated with their reliance behavior in a way that low perceptions relate to more overrides than high perceptions. However, both corrective *and* detrimental overrides increased as fairness perceptions decreased.** This implies that perceptions are not an indicator of people's ability to perform corrective vs. detrimental overrides, but tend to only be associated with the quantity of overrides.

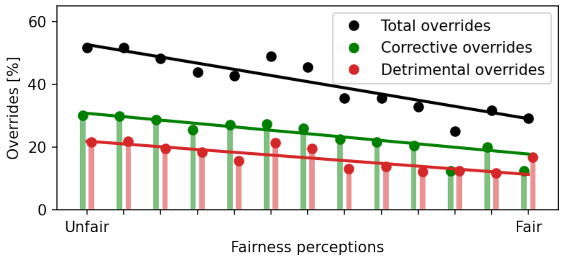


Fig. 11. **Overrides over perceptions (*gendered*).** Significant negative relationship between fairness perceptions and overrides, both corrective and detrimental, as well as overall. Ratio of corrective to detrimental overrides is independent of fairness perceptions.

5 CONCLUSION

In this work, we argue that claims around explanations fostering distributive fairness must directly measure the impact of explanations on fairness metrics of AI-assisted decisions, which depend on humans' reliance behavior. To this end, our study constitutes a blueprint that can be used to evaluate other types of explanations. Crucially, our research shows that the mechanism through which reliance behavior affects metrics of fairness matters. In particular, we show that distributive fairness may improve even in the absence of an enhanced ability to perform corrective overrides. In other words, the presence of explanations may drive a change in fairness metrics by fostering over- or under-reliance for certain types of cases. This finding may be particularly important from a design and a policy perspective, since a common motivation when providing humans with discretionary power to override decisions is an expectation that they will be able to correct for an AI system's mistakes [27, 28].

These findings also have implications for the interpretation of studies focused on perceptions of fairness [74]. Our work shows that fairness perceptions have no bearing on people's ability to correctively override AI recommendations. Instead, our study results suggest that low fairness perceptions are associated with more overrides of AI recommendations, irrespective of their correctness. This may still lead to improvements in distributive fairness but does not indicate that humans differentiate between correct and wrong AI recommendations. This is important as perceptions are often used as proxies for trust and reliance [74].

REFERENCES

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

[2] Yasmeen Alufaisan, Laura R Marusich, Jonathan Z Bakdash, Yan Zhou, and Murat Kantarcioglu. 2021. Does explainable artificial intelligence improve human decision-making?. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 6618–6626.

[3] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. 2022. Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction* 4, 2 (2022), 556–579.

[4] Evan P Apfelbaum, Kristin Pauker, Samuel R Sommers, and Nalini Ambady. 2010. In blind pursuit of racial equality? *Psychological Science* 21, 11 (2010), 1587–1592.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. 2020. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (2020), 82–115.

[6] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[7] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning*. fairmlbook.org. http://www.fairmlbook.org.

[8] Robert Bartlett, Adair Morse, Richard Stanton, and Nancy Wallace. 2022. Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics* 143, 1 (2022), 30–56.

[9] 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.

[10] Miranda Bogen and Aaron Rieke. 2018. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn* 7 (2018).

[11] 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.

[12] Maarten Buyl, Christina Cociancig, Cristina Frattone, and Nele Roekens. 2022. Tackling algorithmic disability discrimination in the hiring process: An ethical, legal and technical analysis. In *2022 ACM Conference on Fairness, Accountability, and Transparency*. 1071–1082.

[13] Stephen Chaudoin, Brian J Gaines, and Avital Livny. 2021. Survey design, order effects, and causal mediation analysis. *The Journal of Politics* 83, 4 (2021), 1851–1856.

[14] Valerie Chen, Q Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the role of human intuition on reliance in human-AI decision-making with explanations. *arXiv preprint arXiv:2301.07255* (2023).

[15] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[16] 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.

[17] Jason A Colquitt and Jessica B Rodell. 2015. Measuring justice and fairness. (2015).

[18] Sam Corbett-Davies and Sharad Goel. 2018. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* (2018).

[19] Arun Das and Paul Rad. 2020. Opportunities and challenges in explainable artificial intelligence (XAI): A survey. *arXiv preprint arXiv:2006.11371* (2020).

[20] Maria De-Arteaga, Stefan Feuerriegel, and Maytal Saar-Tsechansky. 2022. Algorithmic fairness in business analytics: Directions for research and practice. *Production and Operations Management* (2022).

[21] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 120–128.

[22] Hans de Bruijn, Martijn Warnier, and Marijn Janssen. 2022. The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. *Government Information Quarterly* 39, 2 (2022), 101666.

[23] Botty Dimanov, Umang Bhatt, Mateja Jamnik, and Adrian Weller. 2020. You shouldn't trust me: Learning models which conceal unfairness from multiple explanation methods. In *SafeAI @ AAAI*.

[24] 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.

[25] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. 214–226.

[26] Mary T Dzindolet, Scott A Peterson, Regina A Pomranky, Linda G Pierce, and Hall P Beck. 2003. The role of trust in automation reliance. *International Journal of Human-Computer Studies* 58, 6 (2003), 697–718.

[27] European Union. 2016. General Data Protection Regulation. (2016). https://eur-lex.europa.eu/eli/reg/2016/679/oj

[28] Riccardo Fogliato, Maria De-Arteaga, and Alexandra Chouldechova. 2022. A case for humans-in-the-loop: Decisions in the presence of misestimated algorithmic scores. *Available at SSRN 4050125* (2022).

[29] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 80–89.

[30] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[31] Jerald Greenberg. 1987. A taxonomy of organizational justice theories. *Academy of Management Review* 12, 1 (1987), 9–22.

[32] 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.

[33] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. 2016. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, Vol. 1. Barcelona, Spain.

[34] 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.

[35] 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.

[36] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems* 29 (2016).

[37] 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.

[38] 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.

[39] 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.

[40] Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. 2018. Algorithmic fairness. In *AEA Papers and Proceedings*, Vol. 108. 22–27.

[41] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a science of human-AI decision making: A survey of empirical studies. *arXiv preprint arXiv:2112.11471* (2021).

[42] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[43] Himabindu Lakkaraju and Osbert Bastani. 2020. "How do I fool you?" Manipulating user trust via misleading black box explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 79–85.

[44] 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.

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

[46] 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), 1–26.

[47] Han Liu, Vivian Lai, and Chenhao Tan. 2021. Understanding the effect of out-of-distribution examples and interactive explanations on human-AI decision making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–45.

[48] Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems* 30 (2017).

[49] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 2020. Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 122–130.

[50] JoAnn Miller and Marilyn Chamberlin. 2000. Women are teachers, men are professors: A study of student perceptions. *Teaching Sociology* (2000), 283–298.

[51] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. *arXiv preprint arXiv:1802.00682* (2018).

[52] Julian Nyarko, Sharad Goel, and Roseanna Sommers. 2021. Breaking taboos in fair machine learning: An experimental study. In *Equity and Access in Algorithms, Mechanisms, and Optimization*. 1–11.

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

[54] Andrea Papenmeier, Dagmar Kern, Gwenn Englebienne, and Christin Seifert. 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. *ACM Transactions on Computer-Human Interaction (TOCHI)* 29, 4 (2022), 1–33.

[55] Raja Parasuraman and Victor Riley. 1997. Humans and automation: Use, misuse, disuse, abuse. *Human Factors* 39, 2 (1997), 230–253.

[56] Samir Passi and Mihaela Vorvoreanu. 2022. *Overreliance on AI: Literature review*. Technical Report. Microsoft Research.

[57] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 560–568.

[58] Angelisa C Plane, Elissa M Redmiles, Michelle L Mazurek, and Michael Carl Tschantz. 2017. Exploring user perceptions of discrimination in online targeted advertising. In *Proceedings of the 26th USENIX Security Symposium*. 935–951.

[59] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[60] Danish Pruthi, Mansi Gupta, Bhuwan Dhingra, Graham Neubig, and Zachary C Lipton. 2019. Learning to deceive with attention-based explanations. *arXiv preprint arXiv:1909.07913* (2019).

[61] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations as mechanisms for supporting algorithmic transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[62] John K Rempel, John G Holmes, and Mark P Zanna. 1985. Trust in close relationships. *Journal of Personality and Social Psychology* 49, 1 (1985), 95.

[63] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why should I trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 1135–1144.

[64] Victor Riley. 2018. Operator reliance on automation: Theory and data. In *Automation and Human Performance: Theory and Applications*. CRC Press, 19–35.

[65] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 2020. What does it mean to 'solve' the problem of discrimination in hiring? Social, technical and legal perspectives from the UK on automated hiring systems. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 458–468.

[66] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. 2022. A Meta-Analysis of the Utility of Explainable Artificial Intelligence in Human-AI Decision-Making. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 617–626.

[67] Max Schemmer, Niklas Kühl, Carina Benz, Andrea Bartos, and Gerhard Satzger. 2023. Appropriate Reliance on AI Advice: Conceptualization and the Effect of Explanations. *arXiv preprint arXiv:2302.02187* (2023).

[68] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. 2022. On the influence of explainable AI on automation bias. *30th European Conference on Information Systems (ECIS 2022)* (2022).

[69] 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*. Association for Computing Machinery, New York, NY, USA, 1616–1628. https://doi.org/10.1145/3531146.3533218

[70] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 2021. A Study on Fairness and Trust Perceptions in Automated Decision Making. In *Joint Proceedings of the ACM IUI 2021 Workshops, April 13–17, 2021, College Station, USA*.

[71] Donghee Shin and Yong Jin Park. 2019. Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98 (2019), 277–284.

[72] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 2022. Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24, 1 (2022), 1–13.

[73] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial attacks on post hoc explanation methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 180–186.

[74] Christopher Starke, Janine Baleis, Birte Keller, and Frank Marcinkowski. 2021. Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *arXiv preprint arXiv:2103.12016* (2021).

[75] Niels Van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M Kelly, and Vassilis Kostakos. 2019. Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–21.

[76] Jasper van der Waa, Elisabeth Nieuwburg, Anita Cremers, and Mark Neerincx. 2021. Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence* 291 (2021), 103404.

[77] Kun Yu, Shlomo Berkovsky, Ronnie Taib, Dan Conway, Jianlong Zhou, and Fang Chen. 2017. User trust dynamics: An investigation driven by differences in system performance. In *Proceedings of the 22nd International Conference on Intelligent User Interfaces*. 307–317.

[78] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*. 1171–1180.

[79] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.

[80] Zippia. 2022. Professor demographics and statistics in the US. https://www.zippia.com/professor-jobs/demographics/.

[81] Zippia. 2022. Teacher demographics and statistics in the US. https://www.zippia.com/teacher-jobs/demographics/.