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

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

Maria De-Arteaga
The University of Texas at Austin
United States
dearteaga@mccombs.utexas.edu

Niklas Kuehl
Universität Bayreuth
Germany
kuehl@uni-bayreuth.de

ABSTRACT

In this work, we study the effects of feature-based explanations on distributive fairness of AI-assisted decisions. We also investigate how any effects are mediated by humans' fairness perceptions and their 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, we see 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 imply that feature-based explanations are not a reliable mechanism to improve distributive fairness.

1 INTRODUCTION

AI systems are commonly used for assisting 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 such recommendations or override them. 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. For instance, in a recent Forbes article [48], it is claimed that "companies [in financial services and insurance] are using explainable AI to make sure they are making fair decisions about loan rates and premiums." Others have claimed that explanations "provide a more effective interface for the human in-the-loop, enabling people to identify and address fairness and other issues" [25]. 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.

Our work. In this work, we examine the effects of feature-based explanations on people's ability to enhance distributive fairness—and how these effects are 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 study 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 [22]. Participants in both conditions are provided with explanations that visually highlight the most predictive words of their respective AI models. We also include a baseline condition where no explanations are shown. 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 AI 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

[1] We use *reliance* as an umbrella term for people's behavior of adhering to or overriding AI recommendations [52].

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, 19, 26, 50, 69, 76].

2 BACKGROUND

In this section, we provide background on our work and review related literature on explanations, reliance, and fairness.

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 [32, 56, 67]. Apart from the central aim of facilitating human understanding, prior research has formulated a wealth of different desiderata that explanations are to provide, most of which center one or more different types of stakeholders of AI systems [28, 56, 81]. For instance, system designers might be interested in facilitating trust in their systems through explanations, whereas a regulator likely wants to assess a system's compliance with moral and ethical standards [56]. Different goals may sometimes be impossible to accomplish simultaneously [99]. Relevant to our work are several desiderata that concern explanations as an alleged means for better and fairer AI-assisted decision-making [1, 25]; we speak to this in more detail in Section 2.2 and Section 2.3. For a comprehensive overview of different aims of explanations, we refer the reader to Langer et al. [56] and Lipton [60].

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) [40]. However, it has been argued that combining local explanations can also lead to an understanding of global model behavior [62]. So-called *local model-agnostic* explanations, such as LIME [85] or SHAP [63], 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, due to its popularity in the literature as well as in practice [9, 31, 34].

Criticism of explanations. Most desiderata for explanations are insufficiently studied or met with inconclusive or seemingly contradictory empirical findings [15, 23, 56]. A major line of criticism stems from the fact that explanations can mislead people: Chromik et al. [17] 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 [24, 55, 82, 98]. In the extreme case of placebic explanations (i.e., explanations that convey no information about the underlying AI), Eiband et al. [30] find that people may exhibit levels of trust similar to "real explanations". This shows that the sheer presence of explanations can increase people's trust in AI. Even in the absence of any malicious intents, Ehsan and Riedl [29] highlight several challenges arising from unanticipated negative downstream effects of explanations, such as misplaced trust in AI, or over- or underestimating the AI's capabilities. 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, 19, 26, 50, 69, 76]. We address this in more detail in Section 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, 25, 34, 49, 83]. A recent meta-study [88] 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, 35, 61, 68, 106]. On the other hand, Lai and Tan [54] 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 [74, 89]. Humans' ability to override mistaken recommendations has also been referred to as *corrective overriding* [33]. 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" [83]. Empirical evidence, however, is less clear: several studies have found that explanations can be detrimental to appropriate reliance [6, 12, 52, 80, 90, 103], when they increase or decrease humans' adherence to AI recommendations regardless of their correctness. These phenomena are commonly referred to as *over-* or *under-reliance* [89].

Conflation of reliance and trust. Many studies have treated reliance and trust interchangeably [52], sometimes calling reliance a "behavioral trust measure" [72]. However, definitions of *trust* are often inconsistent [46, 58, 72], which makes empirical findings challenging to compare. More importantly, trust and reliance are different constructs [52]: 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 [73, 84, 104]. It has been argued that trust may impact reliance [27, 58, 95], but trust is not a sufficient requirement for reliance when other factors, such as time constraints, perceived risk, or self-confidence, impact decision-making [33, 58, 86]. 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, 13, 21, 45]. 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, 20, 25, 56]. Grounded on the organizational justice literature [18, 36], researchers distinguish different notions of algorithmic fairness, among which are (*i*) *distributive fairness*, which refers to the fairness of decision outcomes [105], and (*ii*) *procedural fairness*, which refers to the fairness of decision-making procedures [59]. Distributive fairness is typically measured in terms of statistical metrics such as parity in error rates across groups [7, 16]; which is closely related to notions like *equalized odds* or *equal opportunity* [41]. Importantly, there is no conclusive evidence showing that explanations lead to fairer decisions, and it remains unclear *how* explanations may enable this [56].

Fairness perceptions. Prior work at the intersection of fairness and explanations has primarily focused on assessing how people *perceive* the fairness of AI systems [52, 100]. Empirical findings are mostly inconclusive, stressing that fairness perceptions depend on many factors, such as the explanation style [10, 25], the amount of information provided [93], the use case [3], user profiles [25], or the decision outcome [96]. 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 [37–39, 69, 79, 102]. 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 typically considered *sensitive* information, e.g., gender or race [19, 37–39, 69, 79, 93]. Interestingly, people's perceptions towards these features change after they learn that "blinding" the AI model to these features can lead to *worse* outcomes for marginalized groups. Similarly, it has been shown that people's perceptions towards the inclusion of sensitive features switch when they are told that this inclusion makes an AI model more accurate [38] or equalizes error rates across demographic groups [42]. 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, 19, 26, 50, 69, 76], and that there exist several real-world examples where the inclusion of sensitive features can make historically disadvantaged groups like Black people or women better off [19, 66, 78, 97]. 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

In this section, we outline our study design. First, we introduce the task and dataset for our study, then we explain the experimental setup and our dependent variables, and, finally, the data collection process.

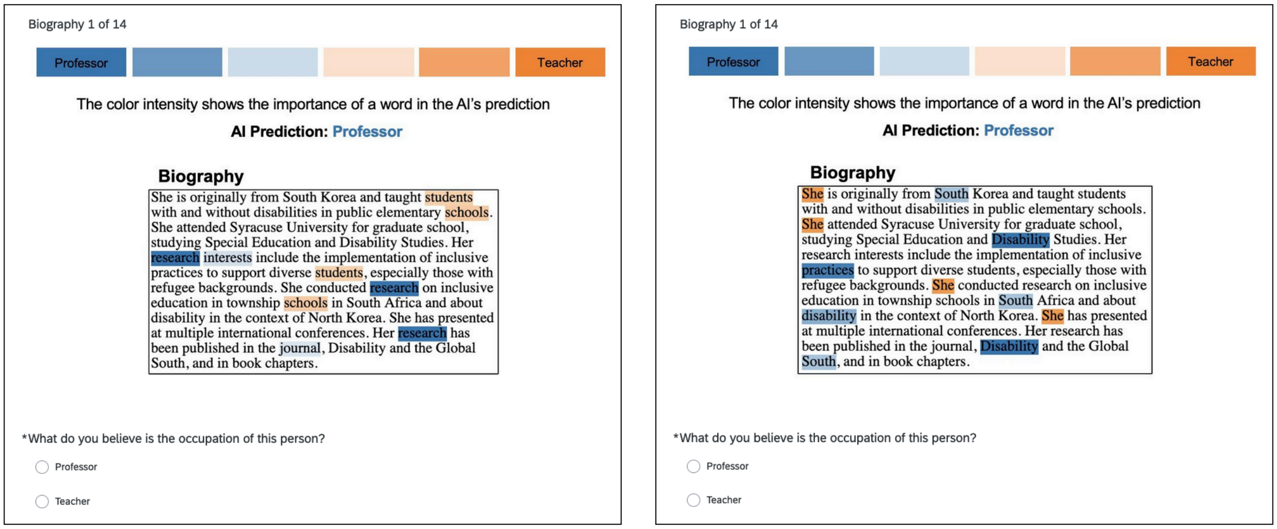
3.1 Task and dataset

Task. Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online [11, 87]. 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 systems, it is susceptible to gender bias and discrimination [11, 22, 87]. De-Arteaga et al. [22] show that these biases can manifest themselves in error rate disparities between genders, and 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) [22]. 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 [67].[2]

Dataset. We use the publicly available BIOS dataset, which contains approximately 400,000 online bios for 28 different occupations from the Common Crawl corpus, initially created by De-Arteaga et al. [22].[3] This data set has been used in other human-AI decision-making studies as well, such as the ones by Liu et al. [61] or Peng et al. [77]. For each bio in the dataset we know the gender of the corresponding person and their true occupation. Gender is based on the pronouns used in the bio, and a limitation of this dataset is that it only contains bios that use "she" or "he" as pronouns, excluding bios of non-binary people. We only consider bios that belong to professors and teachers, which leaves us with 134,436 bios, out of which 118,215 belong to professors and 16,221 to teachers. In line with current demographics and societal stereotypes [67, 107, 108], we have more men (55%) than women (45%) bios of professors and more women (60%) than men (40%) bios of teachers.

[2] See also [107, 108] on current demographic statistics for professors and teachers in the US.

[3] The code that reproduces the dataset can be found at https://github.com/Microsoft/biosbias.



Biography 1 of 14

Professor | | | | | Teacher

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

Biography 1 of 14

Professor | | | | | Teacher

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**

***Gendered* condition**

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

3.2 Experimental setup

General procedure. 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 shown to participants; we outline the construction of these models later in this section as well as, more extensively, in Section A.

Participants in each condition first complete the task of predicting occupations for 14 bios, and—if assigned to a condition with explanations—answer several questions regarding their fairness perceptions. Since the baseline condition does not provide any cues regarding the AI's decision-making procedures, we do not ask about perceptions there. Finally, participants provide some demographic information. A summary of our general setup in illustrated in Figure 2. Note that we ask about fairness perceptions *after* the task is completed, so as to prevent these questions from moderating reliance behavior [14]. Given that distinguishing professors and teachers based on their bios can be at times ambiguous and not everyone may be familiar with the differences, we also ask at the beginning of our questionnaires what participants consider the difference between *professor* and *teacher* to be. Additionally, after completing the task, we ask participants an open-ended question on what information they relied on when differentiating *professor* and *teacher*. This way, we were able to confirm—both quantitatively and qualitatively—that participants thought consistently about this distinction 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. This interface is similar to related studies on AI-assisted text classification [53, 61, 89]. 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 (see Section 3.4).

Task-relevant and gendered classifiers. We provide intuition for how we constructed the AI models that generate recommendations and explanations in the *task-relevant* and *gendered* conditions. We defer a detailed explanation to Section A. The general idea is to 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

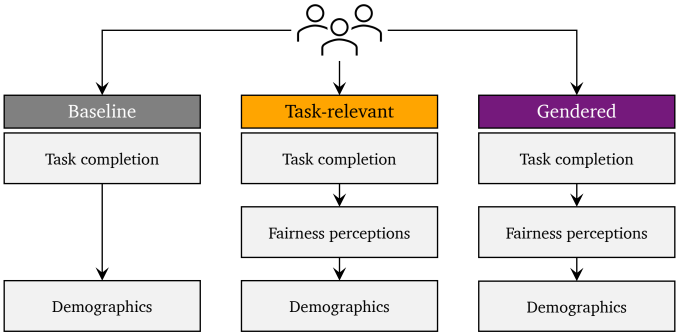


Figure 2: Illustration of our experimental setup. Participants are randomly assigned to one of three conditions. In each condition, they first complete the task of predicting occupations from 14 short bios, and complete a demographic survey. In the conditions with explanations (*Task-relevant* and *Gendered*), participants are also asked about their fairness perceptions after completing the task.

are not evidently gendered but highly correlated with the sensitive attribute. Finally, we train two logistic regression models[4] on a balanced set of professor and teacher bios, and we employ the `TextExplainer` from LIME [85] to generate dynamic explanations with highlighting of predictive words.

3.3 Measuring reliance and fairness

Selection of bios. 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). Note that our focus is on scenarios where the AI recommendations are in line with gender stereotypes. To preempt the misconception that the AI always recommends *teacher* for women and *professor* for men, we also include one case each of correctly recommended woman professor (WPP) and correctly recommended man teacher (MTT). In the light of recent findings from Kim et al. [47], we include the WPP and MTT scenarios early on in our questionnaires. Precisely, we randomize the order in which participants see the 14 bios, with the restriction that the WPP and MTT scenarios are shown among the first five. 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 [22].

All bios shown to participants are taken from a random holdout set of BIOS 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. The whole selection process is described in more detail in Section B.

Measuring reliance behavior. In our assessment of reliance behavior, we distinguish four cases, as depicted in Table 2. We refer to cases where humans adhere to correct AI recommendations as *correct adherence*, to cases where humans adhere to wrong recommendations as *detrimental adherence*, to cases where humans override correct recommendations as *detrimental overriding*, and to cases where humans override wrong recommendations as *corrective overriding*. Note that the sum of shares of correct adherence and corrective overriding make up the final decision-making accuracy [92]. This taxonomy is similar to the one proposed by Liu et al. [61] for trust; however, we want to stress the difference between trust and reliance (see Section 2.2). When comparing participants' reliance behavior across conditions, we compute and report the relative shares of any of these four types of reliance behavior on the 14 bios that participants see.

Measuring distributive fairness. To evaluate distributive fairness of decisions, we measure disparities in error rates across gender [7, 16], which is closely linked to the ideas of *equalized odds* and *equal opportunity* [41]. From a fairness perspective, the goal is to minimize such disparities, so as to equalize the burden of being misclassified and, as a result, being excluded from exposure to relevant opportunities between men and women. We formalize these disparities as follows: let FP_W be the share of wrongly predicted women professors, i.e., women professors that are predicted to be teachers, and FT_W the share of wrongly predicted woman teachers. Similarly define FP_M and FT_M for men. We can then quantify disparities in error rates as follows:

$$\text{Error rate disparity (Teacher} \longrightarrow \text{Professor)} = |FT_W - FT_M|$$
$$\text{Error rate disparity (Professor} \longrightarrow \text{Teacher)} = |FP_W - FP_M|,$$

where we use the notation of "Teacher → Professor" to indicate teachers that are wrongly predicted as professors, and vice versa for "Professor → Teacher". If we assume that the occupation of

[4] We use logistic regression to ensure that explanations are faithful to the underlying model.

Table 1: Overview of the six types of scenarios employed in our study. Our study includes 14 bios, consisting of three scenarios of types WTT, WPT, MTP, and MPP, respectively, and one scenario each of types WPP and MTT.

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: humans can adhere to or override correct AI recommendations, or they can adhere to or override wrong AI recommendations.

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

professor is associated with a higher societal status than *teacher*, we may also refer to cases of "Teacher ⟶ Professor" as *promotions*, and to "Professor ⟶ Teacher" as *demotions*. This will be important in the discussion of our findings.

Measuring fairness perceptions. To measure fairness perceptions, we provide a brief introduction and then ask participants' agreement with three statements, measured on 5-point Likert scales from 1 ("Fully disagree") to 5 ("Fully agree"). We operationalize this in our questionnaires similar to Colquitt and Rodell [18] as follows:

> *The questions below refer to the procedures the AI uses to predict a person's occupation. Please rate your agreement with the following statements.*
>
> (1) *The AI's procedures are free of bias.*
> (2) *The AI's procedures uphold ethical and moral standards.*
> (3) *It is fair that the AI considers the highlighted words for predicting a person's occupation.*

Note that items (1) and (2) are taken from the *procedural justice* construct of Colquitt and Rodell [18] and slightly rephrased to fit our case of AI-assisted decision-making. These items have been frequently used in other human-AI studies, e.g., [10, 65, 91, 94]. Note that prior work has often measured fairness perceptions through single items only [100]. Colquitt and Rodell [18] propose up to eight measurement items for procedural justice in the organizational psychology context; however, several of these items are not applicable here. Instead, we amend our questionnaires by a third item (3) that is more tailored to our experimental setup. Since item (3) is more explicit and we want to avoid priming, we ask (3) last and without possibility to modify responses for (1) and (2) retroactively. To obtain a single measure of fairness perceptions per participant, we eventually average ratings across the three items per participant; and we also confirm scale reliability in Section 4.3.

3.4 Data collection

Our study has received clearance from an institutional ethics committee. Participants were recruited via `Prolific`—a crowdworking platform for online research [71]. 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 prescreeners 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`. At the time of taking the survey, 13.5% of participants were 18–24 years old, 32.6% were 25–34 years old, 21.3% between 35–44, 13.8% between 45–54, 11.3% between 55–64, and 7.6% were older than 65. Regarding gender, 49.2% identified as women, 48.0% as men, and 1.8% identified as non-binary / third gender, or preferred not to say. 8.0% of participants are of Spanish, Hispanic, or Latinx ethnicity; and the majority (78.4%) considered their race to be White or Caucasian, followed by Black or African American (7.0%) and Asian (6.1%). For their participation, participants were paid on average £10.58 (approximately $12.70 at the time the study was conducted) per hour, excluding individual bonus payments of £0.05 per correctly predicted occupation. Participants took on average 10:12min (baseline), 12:51min (*task-relevant*), and 12:27min (*gendered*) to complete the survey.

4 ANALYSIS AND RESULTS

We first present results on the effects of explanations on accuracy as well as overriding behavior. Then, we examine how reliance behavior translates to distributive fairness. Finally, we assess the role of fairness perceptions. For all statistical comparisons, we conduct nonparametric tests because we cannot confirm the prerequisites (normal distribution and equal variance) of their parametric counterparts. Specifically, we conduct Kruskal-Wallis omnibus tests [51] whenever applicable, and two-tailed Mann-Whitney U tests [64] for

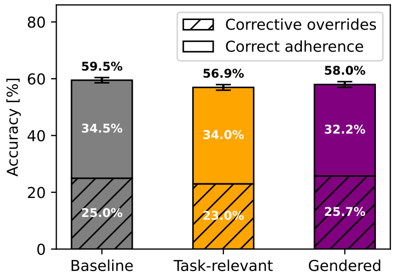


Figure 3: Accuracy by condition. Accuracy is not higher when explanations are provided, compared to the baseline.

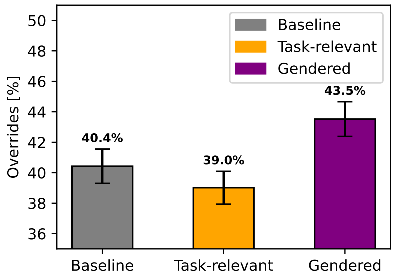


Figure 4: Overrides by condition. Overrides are highest in the *gendered* condition.

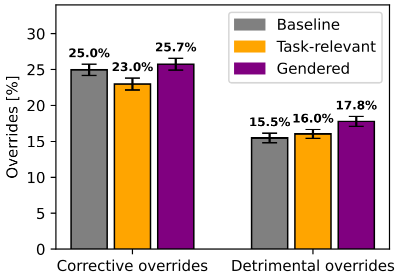


Figure 5: Overriding behavior. Explanations do not enable corrective vs. detrimental overrides.

pairwise comparisons. We report p-values for pairwise comparison tests in Table 3 in Section C.

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*. Mean accuracies[5] per condition are $M_{base} = 59.49\%$ ($SD_{base} = 13.11$), $M_{rel} = 56.94\%$ ($SD_{rel} = 13.86$), and $M_{gen} = 57.96$ ($SD_{gen} = 14.30$), as shown in Figure 3.[6] The Kruskal-Wallis omnibus test further suggests that there are no significant differences between the three means ($p = 0.260$). Recall that participants were incentivized through bonus payments to accurately predict occupations. This suggests that **explanations did not aid AI-assisted decision-making when measured in terms of accuracy.**

Effects on overriding behavior. In Figures 4 and 5, we see that participants in the *gendered* condition overrode more AI recommendations than in the *task-relevant* condition ($p = 0.005$). From Figure 5 we further conclude that *both* corrective *and* detrimental overrides are highest in the *gendered* condition, with detrimental overrides being significantly higher than the baseline ($p = 0.012$). We interpret this increase in overrides further in Section 4.2. In the *task-relevant* condition, we see that overall overrides are lowest across conditions (Figure 4), with corrective overrides being marginally[7] lower ($p = 0.097$) and detrimental overrides not significantly different ($p = 0.374$) compared to the baseline (Figure 5). Overall, we conclude that people's reliance behavior is affected by how the AI explains its recommendations; notably, people overrode AI recommendations more often when explanations highlight features that are evidently associated with gender. Across conditions, we also infer from Figure 5 that participants generally performed more corrective than detrimental overrides, and that **the ability to perform corrective vs. detrimental overrides (i.e., the ratio of corrective to 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 3), we do not observe any accuracy changes through explanations over the baseline in Figure 6, neither for men ($p = 0.199$) nor women ($p = 0.151$) bios. 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 7 and 8, no intervention improved participants' ability to perform corrective vs. detrimental overrides of AI recommendations compared to the baseline—i.e., the ratio of corrective to detrimental overrides did not improve—neither for men nor women bios. This is consistent with our findings at the aggregate level (see Figure 5). Notably, we see that detrimental overrides in the *gendered* condition marginally increase for men bios (Figure 7) over the baseline ($p = 0.078$), and in the *task-relevant* condition they significantly increase for women bios (Figure 8) compared to the baseline ($p = 0.013$). At the same time, corrective overrides remain unchanged in either case.

From comparing Figures 7 and 8 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 (see Figures 14 and 15 in Section D). Together, these results suggest that **people were overall more prone to do promoting[8] overrides**; which means that participants overrode AI recommendations more often when someone was suggested to be a teacher vs. a professor.

[5] We use M as a shorthand for *mean*, and SD for *standard deviation*. We also use the subscripts $base$, rel, and gen to refer to the baseline, task-relevant, and gendered conditions, respectively.

[6] In figures we provide standard errors as error bars, where we compute the measure of interest (e.g., accuracy) for each individual participant in a given condition, then compute the standard deviation across all participants in that condition, and divide the result by the square root of the number of participants in that condition.

[7] We report *marginal* significance for $0.05 < p \leq 0.10$ in line with prior work [70], noting that such p-values have lower evidential value.

[8] 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.

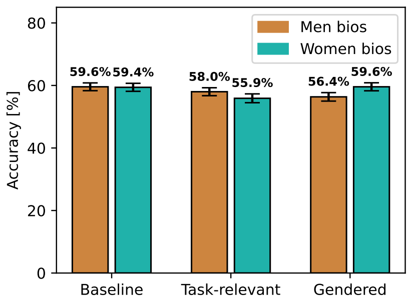


Figure 6: Accuracy by condition and gender of bio. Explanations (middle and right) do not increase accuracy over the baseline, neither for men nor women bios.

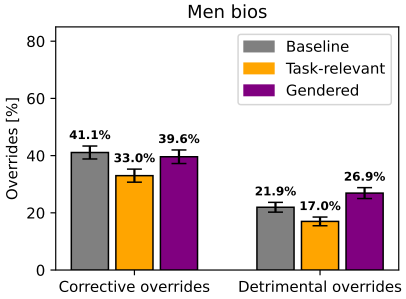


Figure 7: Overriding behavior for men bios. *Task-relevant* explanations decrease both corrective and detrimental overrides for men bios, compared to the baseline; whereas *gendered* explanations marginally increase detrimental overrides.

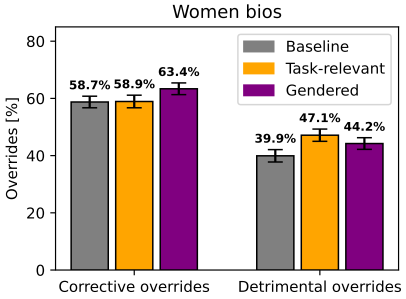


Figure 8: Overriding behavior for women bios. *Gendered* explanations increase both corrective and detrimental overrides over the baseline; *task-relevant* explanations increase detrimental overrides.

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[9] corrective overrides, and both stereotype-aligned and stereotype-countering detrimental overrides. As explained in Section 3.2, the motivation for this design is our focus on studying whether explanations allow humans to correct for stereotype-aligned wrong AI predictions, which would be the most frequent errors of an occupation prediction model that exhibits gender bias [22]. We see that in the *task-relevant* condition, people perform fewer corrective overrides for men ($p = 0.011$) and the same amount for women ($p = 0.834$) in comparison to the baseline, as shown in Figures 7 and 8. Meanwhile, in the *gendered* condition participants perform marginally more corrective overrides for women ($p = 0.083$) and the same amount of such overrides for men ($p = 0.588$). 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 marginally increase in the *gendered* condition for men bios ($p = 0.078$), compared to the baseline (Figure 7). We also see that they are higher for women bios in the *gendered* condition (Figure 8), even though not statistically significant ($p = 0.110$). Considering that we do not observe differences in stereotype-aligned detrimental overrides between conditions (Figures 14 and 15 in Section D), 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.** We elaborate on the implications of this for distributive fairness below.

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 9), and erroneously demote women more than men (41.3% vs. 21.9% in Figure 10). 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 11. 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 11, left). Similarly, demotions of men decreased much more than demotions of women, leading to increased disparities in demotions over the baseline (Figure 11, 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 11, left). The most drastic reduction in disparities happens for demotions (Figure 11, right), since demotions *increased* for men and *decreased* for women (Figure 10). 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**

[9] Recall that societal stereotypes typically associate men with being professors and women with being teachers [67].

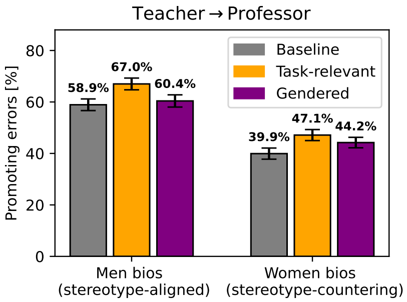


Figure 9: Bios wrongly classified by humans as professor. Promoting errors increase for both men and women bios in the *task-relevant* condition, compared to the baseline; and they marginally increase for women bios in the *gendered* condition.

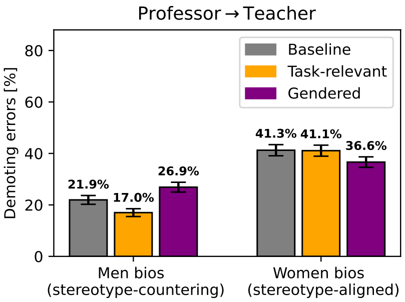


Figure 10: Bios wrongly classified by humans as teacher. In the *gendered* condition, demoting errors increase for men bios (left) and decrease for women bios (right), compared to the baseline; and they only decrease for men bios in the *task-relevant* condition.

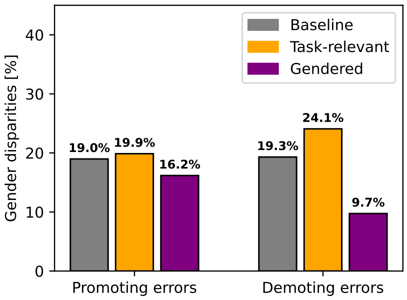


Figure 11: Absolute error differences between men and women. *Gendered* explanations decrease both disparities in promoting (teacher → professor) and demoting (professor → teacher) errors between genders, compared to the baseline; *task-relevant* explanations increase disparities.

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 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. Recall that we measure three items regarding fairness perceptions on 5-point Likert scales, ranging from 1 (unfair) to 5 (fair), see Section 3.2. We confirm good scale reliability at a Cronbach's alpha [101] value of 0.77. We then take the average of the three item ratings for each participant to obtain a single measure of fairness perceptions. From the distribution in Figure 12, we see 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. This means that people who are shown a highlighting of task-relevant words perceived the underlying AI as fairer than people who were shown gendered words as being important for given AI recommendations. 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 observe a strong negative relationship ($p = 1.10 \times 10^{-11}$) between fairness perceptions and overriding of AI recommendations, i.e., participants overrode the AI more often when their fairness perceptions were lower. Concretely, we see that people overrode on average 52% of AI recommendations when their fairness perceptions were lowest, and only 31% when their fairness perceptions were highest. 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 13 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. All slopes in Figure 13 are significantly negative (total: $p = 1.97{\times}10^{-7}$; corrective: $p = 9.18{\times}10^{-5}$; detrimental: $p = 1.53 \times 10^{-4}$). 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—in other words, the ratio of corrective to detrimental overrides is constant across perceptions. Overall, 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.

5 DISCUSSION AND CONCLUSION

Summary of findings. In this work, we conducted a first holistic analysis of the effects of feature-based explanations on distributive fairness in AI-assisted decision-making. We also studied the mediating roles of reliance behavior and fairness perceptions, which have been the focus of prior work. Our findings suggest that feature-based explanations can have different effects on people's perceptions, their reliance behavior, and distributive fairness—depending on whether they highlight the use of task-relevant words or words that are proxies for sensitive attributes. Specifically, we observe that for the task of occupation classification, a highlighting of gendered words led to lower fairness perceptions, which are associated with more overrides of AI recommendations. On the other hand, when

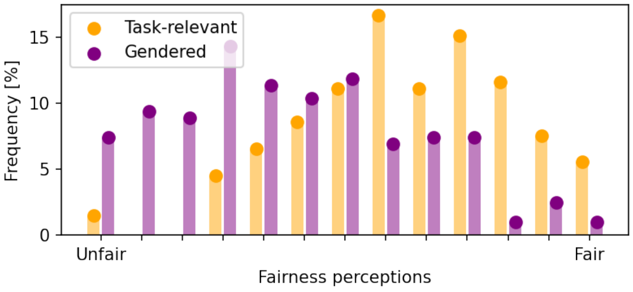


Figure 12: Distribution of fairness perceptions. Fairness perceptions are higher in the *task-relevant* condition compared to the *gendered* condition. Fairness perceptions are averages of three items measured on 5-point Likert scales, resulting in values between 1 ("unfair") and 5 ("fair") with 0.33 increments.

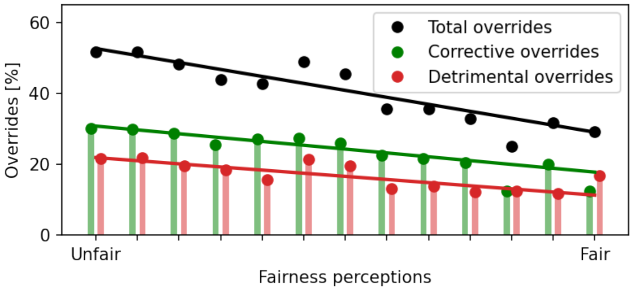


Figure 13: Overrides over perceptions (*gendered*). Significant negative relationship between fairness perceptions and overrides, both corrective and detrimental, as well as overall. Ratio of corrective to detrimental overrides is independent of fairness perceptions.

task-relevant words are highlighted this led to higher fairness perceptions, which translate to fewer overrides. In no case, however, do we observe that explanations improve people's ability to perform corrective vs. detrimental overrides, compared to a scenario with no explanations. Finally, we show that feature-based explanations can improve *or* hinder distributive fairness by fostering shifts in errors that counter or reinforce stereotypes: in the *gendered* condition, participants displayed stereotype-countering reliance behavior, while in the *task-relevant* condition, they displayed stereotype-aligned behavior. In both these cases, the respective reliance behavior affected both corrective and detrimental overrides. This means that the conditions affected the likelihood to perform an override conditioned on the predicted occupation and a bio's associated gender, but with no relationship to the true occupation. For instance, the *gendered* condition fostered *more* overrides of AI recommendations when a woman was predicted to be a teacher, irrespective of whether this prediction was correct; meanwhile, in the *task-relevant* condition participants were *less* likely to override recommendations where a man was predicted to be a professor, irrespective of his true occupation.

Limitations. Our study setup assigned participants to either the *gendered* or the *task-relevant* condition; i.e., participants saw either only explanations with highlighting of gendered words or task-relevant words. We made this choice because we wanted to measure perceptions of fairness, but eliciting perceptions at an instance level could lead people to anchor their decisions to their expressed perceptions (or vice versa), which would compromise external validity. Assigning people to different conditions enabled us to measure perceptions at the aggregate level. In practice, an AI model might sometimes highlight only sensitive features, sometimes only task-relevant features, and at other times a mix of both. Future work that studies how instance-level perceptions relate to aggregate-level perceptions, and how these interdependencies shape reliance behavior could complement our findings. While our study design does not explicitly account for this, even if perceptions vary at the instance level, our findings suggest that reliance would depend on the inclusion of sensitive features, which research has shown to be an unreliable signal for assessing algorithmic fairness [4, 26, 50, 55, 69, 76, 82]. In particular, previous research has shown that "fairness through unawareness", i.e., the exclusion of information that is evidently indicative of a person's demographics, is neither necessary nor sufficient for an algorithm to be procedurally fair [55, 69, 82] or to not display bias in terms of distributive fairness [4, 26, 50, 76]. Our work complements these works by showing that feature-based explanations may foster stereotype-aligned reliance behavior, therefore *hindering* distributive fairness in AI-assisted decisions.

Importantly, our study does not claim that the observed effects will necessarily generalize beyond the given setup. Instead, with this work, we aim to provide an important example that shows how unreliable feature-based explanations are when it comes to effects on humans' reliance behavior and distributive fairness. Our hope is that this work will inform improved assessment and design of explanability techniques, leading to a nuanced understanding of when and how certain types of explanations can enable humans to improve fairness properties of a system.

Implications and outlook. A main argument of our work is 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 should be used to evaluate other types of explanations and tasks. 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 [32, 33].

These findings also have implications for the interpretation of studies focused on perceptions of fairness [100]. 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 [100].

Previous work has emphasized that interpretability is not a monolithic concept, and the design of explanations should always be grounded on a concrete objective that it helps advance [60]. Our work emphasizes the importance of designing explanations with the explicit purpose of enabling people to rely on AI recommendations in a way that enhances distributive fairness, and it casts doubt over the reliability of popular explainability approaches to advance this goal. To this point, novel findings from ethnographic work studying the use of AI have the potential to inform alternative designs of explanations. For instance, Lebovitz et al. [57] study the adoption of AI in three healthcare domains and emphasize the importance of *interrogation practices*, which are practices used by humans to relate their own knowledge to AI's predictions. Other works have studied interventions that help humans reason over the information that is and is not available to the algorithm [43, 44]. Future studies should explore whether explanations of the broader socio-technical system better enable humans to perform corrective overrides that foster distributive fairness.

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] Umang Bhatt, Alice Xiang, Shubham Sharma, Adrian Weller, Ankur Taly, Yunhan Jia, Joydeep Ghosh, Ruchir Puri, José MF Moura, and Peter Eckersley. 2020. Explainable machine learning in deployment. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 648–657.

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

[12] 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.

[13] 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.

[14] 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.

[15] 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).

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

[17] 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.

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

[19] 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).

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

[21] 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).

[22] 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.

[23] 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.

[24] 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*.

[25] 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.

[26] 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.

[27] 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.

[28] Upol Ehsan and Mark O Riedl. 2020. Human-centered explainable AI: Towards a reflective sociotechnical approach. In *International Conference on Human-Computer Interaction*. Springer, 449–466.

[29] Upol Ehsan and Mark O Riedl. 2021. Explainability pitfalls: Beyond dark patterns in explainable AI. *arXiv preprint arXiv:2109.12480* (2021).

[30] Malin Eiband, Daniel Buschek, Alexander Kremer, and Heinrich Hussmann. 2019. The impact of placebic explanations on trust in intelligent systems. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.

[31] Radwa ElShawi, Youssef Sherif, Mouaz Al-Mallah, and Sherif Sakr. 2021. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence* 37, 4 (2021), 1633–1650.

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

[33] 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).

[34] 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.

[35] 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.

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

[37] 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.

[38] 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.

[39] 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.

[40] 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.

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

[42] Galen Harrison, Julia Hanson, Christine Jacinto, Julio Ramirez, and Blase Ur. 2020. An empirical study on the perceived fairness of realistic, imperfect machine learning models. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 392–402.

[43] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. 2022. On the effect of information asymmetry in human-AI teams. *ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI)* (2022).

[44] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. 2022. Toward Supporting Perceptual Complementarity in Human-AI Collaboration via Reflection on Unobservables. *arXiv preprint arXiv:2207.13834* (2022).

[45] 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.

[46] 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.

[47] Antino Kim, Mochen Yang, and Jingjing Zhang. 2022. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Trans. Comput.-Hum. Interact.* (2022). https://doi.org/10.1145/3557889

[48] Jennifer Kite-Powell. 2022. Explainable AI is trending and here's why. *Forbes* (2022).

[49] 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.

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

[51] William H Kruskal and W Allen Wallis. 1952. Use of ranks in one-criterion variance analysis. *J. Amer. Statist. Assoc.* 47, 260 (1952), 583–621.

[52] 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).

[53] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards building model-driven tutorials for humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[54] 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.

[55] 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.

[56] 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.

[57] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 2022. To engage or not to engage with AI for critical judgments: How professionals deal with opacity when using AI for medical diagnosis. *Organization Science* 33, 1 (2022), 126–148.

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

[59] 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.

[60] Zachary C Lipton. 2018. The mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 3 (2018), 31–57.

[61] 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.

[62] Scott M Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2, 1 (2020), 56–67.

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

[64] 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.

[65] 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.

[66] Sandra G Mayson. 2018. Bias in, bias out. *The Yale Law Journal* 128 (2018).

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

[68] 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).

[69] 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.

[70] Anton Olsson-Collentine, Marcel ALM Van Assen, and Chris HJ Hartgerink. 2019. The prevalence of marginally significant results in psychology over time. *Psychological Science* 30, 4 (2019), 576–586.

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

[72] 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.

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

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

[75] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[76] 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.

[77] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar. 2022. Investigations of performance and bias in human-AI teamwork in hiring. *arXiv preprint arXiv:2202.11812* (2022).

[78] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Daniel Jenson, Amy Shoemaker, Vignesh Ramachandran, Phoebe Barghouty, Cheryl Phillips, Ravi Shroff, et al. 2020. A large-scale analysis of racial disparities in police stops across the United States. *Nature Human Behaviour* 4, 7 (2020), 736–745.

[79] 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.

[80] 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.

[81] Alun Preece, Dan Harborne, Dave Braines, Richard Tomsett, and Supriyo Chakraborty. 2018. Stakeholders in explainable AI. *arXiv preprint arXiv:1810.00184* (2018).

[82] 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).

[83] 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.

[84] 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.

[85] 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.

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

[87] 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.

[88] 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.

[89] 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).

[90] 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).

[91] Nadine Schlicker, Markus Langer, Sonja K 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* 122 (2021), 106837.

[92] Jakob Schoeffer, Johannes Jakubik, Michael Voessing, Niklas Kuehl, and Gerhard Satzger. 2023. On the Interdependence of Reliance Behavior and Accuracy in AI-Assisted Decision-Making. *arXiv preprint arXiv:2304.08804* (2023).

[93] 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

[94] 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*.

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

[96] 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.

[97] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. 2016. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and Human Behavior* 40, 5 (2016), 580.

[98] 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.

[99] Aaron Springer and Steve Whittaker. 2019. Making transparency clear. In *Algorithmic Transparency for Emerging Technologies Workshop*, Vol. 5.

[100] 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).

[101] Keith S Taber. 2018. The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education* 48 (2018), 1273–1296.

[102] 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.

[103] 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.

[104] 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.

[105] 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.

[106] 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.

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

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

A CONSTRUCTION OF TASK-RELEVANT AND GENDERED CLASSIFIERS

Here, we explain in more detail how we constructed the AI models that we use for generating recommendations and explanations in the *task-relevant* and *gendered* conditions.

Let $\mathcal{W} := \{w_1, \ldots, w_n\}$ be the set of n words that occur most often across the set of all bios. We chose $n = 5000$, i.e., $\mathcal{W}$ contains the top-5000 most occurring words, after removal of (manually defined) stop words. We inferred $\mathcal{W}$ from applying a `CountVectorizer` [75]. In trial runs, we found that increasing n beyond 5000 does not significantly change the classifiers' predictions. We then constructed two logistic regression classifiers, $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$, with access to mutually disjoint vocabularies: *task-relevant words* ($\mathcal{W}_{rel} \subset \mathcal{W}$) and *gendered words* ($\mathcal{W}_{gen} \subset \mathcal{W}$).

Task-relevant vocabulary. We performed the following steps to construct the task-relevant vocabulary $\mathcal{W}_{rel}$:

(1) For all $i \in \{1, \ldots, n\}$, compute the average occurrence of word $w_i \in \mathcal{W}$ in bios of men and women professors and teachers. We call the results $\widehat{w_i^{P,m}}$, $\widehat{w_i^{P,w}}$, $\widehat{w_i^{T,m}}$, and $\widehat{w_i^{T,w}}$, where we use P, T and m, w as a shorthand for the respective occupations and genders. We also compute $\widehat{w_i^{\bullet}}$ as the average occurrence of w_i for any other occupation $\bullet$ that is *not* professor or teacher.

(2) For given gender $g \in \{m, w\}$, check whether $\widehat{w_i^{P,g}} > \widehat{w_i^{\bullet}}$ *or* $\widehat{w_i^{T,g}} > \widehat{w_i^{\bullet}}$ for all other occupations $\bullet$, i.e., whether the average of word w_i in professor or teacher bios of gender g is greater to the average in *any* other occupation. If this condition is met, add w_i to $\mathcal{W}_{rel}^{g}$, the set of task-relevant words for gender g.

(3) Compute $\mathcal{W}_{rel}^{m} \cap \mathcal{W}_{rel}^{w} = \mathcal{W}_{rel}$ as the set of words that are task-relevant for *both* genders.

After completing steps (1)–(3), we obtain the task-relevant vocabulary $\mathcal{W}_{rel}$ of 543 words, including *faculty*, *kindergarten*, or *phd*, among others.

Gendered vocabulary. Denote $|\mathcal{B}^{o,g}|$ the amount of bios of occupation $o \in \{P, T\}$ and gender $g \in \{m, w\}$. We perform the following steps to construct the gendered vocabulary $\mathcal{W}_{gen}$:

(1) Sample equal amounts of bios for men and women professors and teachers. Since $\min\{|\mathcal{B}^{o,g}|\} = |\mathcal{B}^{T,m}| = 6440$, randomly sample 6440 bios for each combination of occupation and gender.

(2) Extract features from bios by applying a `CountVectorizer` with `TF-IDF` weighting [75].

(3) Train a logistic regression to predict *gender* from the extracted features.

(4) Compute the importance of each (weighted) feature based on the absolute magnitude of their corresponding regression coefficient, and sort the resulting list of words by importance.

(5) Include the top-5% most important words in $\mathcal{W}_{gen}$ as the set of words that are highly predictive of gender. We choose the threshold of 5% so as to exclude words that are spuriously correlated with gender (e.g., *towards*).

After completing steps (1)–(5), we obtain the gendered vocabulary $\mathcal{W}_{gen}$ of 214 words, which include—apart from gender pronouns and words such as *husband* and *wife*—words like *dance*, *art*, or *engineering*, which are not evidently gendered.

Deploying the classifiers. Having established our vocabularies $\mathcal{W}_{rel}$ and $\mathcal{W}_{gen}$, we proceed by training two logistic regression models on a balanced set of bios containing 50% professors and 50% teachers. Denote $|\mathcal{B}^P|$ and $|\mathcal{B}^T|$ the amounts of bios of occupations P and T. Since $|\mathcal{B}^T| = 16,221 < |\mathcal{B}^P|$, we randomly sample 16,221 bios of professors, while preserving the gender distribution from the original data. This yields a dataset of 32,442 bios, 50% of which we use as a holdout set. We separate a relatively large holdout set because we will eventually use a specific subset of these bios in our questionnaires (see Section B). The resulting classifiers achieve F_1 scores of 0.87 ($\mathbf{AI}_{rel}$) and 0.77 ($\mathbf{AI}_{gen}$). For generating dynamic explanations with highlighting of predictive words, we employ the `TextExplainer` from LIME [85].

B SELECTION OF BIOS

Pre-selection. As outlined in Section 3.2, participants are confronted with 14 bios of professors and teachers. We impose a series of constraints to select which bios from the holdout set we include in the questionnaires. In particular, for a given bio to be included in our questionnaires, we require it to satisfy the following:

- Both models $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ must yield the same predicted occupation for the bio.
- The prediction probabilities of $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ towards either occupation must be *at most* 20% different. This ensures that both models are comparably certain in their predictions for the given bio.
- The prediction probabilities of $\mathbf{AI}_{rel}$ and $\mathbf{AI}_{gen}$ towards either occupation must be *at most* 80%. This aims at eliminating a large share of bios that are "too easy" to classify.
- To avoid any confounding effects of bios' length on people's behavior, we only consider bios of length between 50 and 100 words.

Enforcing these constraints on bios from the holdout set leaves us with 690 eligible bios (out of 16,221). In a next step, we decide on the final set for our questionnaires.

Final selection. The authors jointly screened these 690 bios and ruled out those that are trivial (e.g., because humans would easily be able to tell the occupation) or otherwise not suitable (e.g., because of misspellings or excessive use of jargon). We also discarded bios where explanations would highlight too few or too many words, or where the number of highlighted words was significantly different between the *task-relevant* and the *gendered* condition. This filtering narrows down the set of eligible bios to 38. The authors then independently screened the resulting 38 bios including the corresponding explanations, and assigned a rating of green ("in favor of using it"), yellow ("indifferent"), or red ("in favor of discarding it"), based on both a bio's content as well as the associated explanation, favoring bios that were non-trivial but that contained enough information to make a correct prediction. We then decided on the final set of 14 bios based on majority vote, taking into account the required composition of scenarios, as outlined in Table 1 in Section 3.2.

C STATISTICAL RESULTS OF PAIRWISE COMPARISONS

Table 3 contains results of pairwise comparison tests.

D OVERRIDES OF CORRECT ANTI-STEREOTYPICAL AI RECOMMENDATIONS

Figures 14 and 15 show participants' overriding behavior for cases where AI recommendations are correct and anti-stereotypical; i.e., correctly suggesting men to be teachers (MTT) and women to be professors (WPP). We see that across conditions, overrides are much higher for the MTT case than for the WPP case. Together with the findings from Section 4 this suggests that participants were more prone to override AI recommendations whenever they suggested someone to be a teacher vs. a professor.

Table 3: Results of pairwise comparisons. We report p-values of two-tailed nonparametric Mann-Whitney U tests for pairwise comparisons. We provide p-values for both the comparison between baseline and *task-relevant* as well as baseline and *gendered* condition. Column names refer to the corresponding figures in the main body of the paper.

Comparison	Fig. 4	5 (Corr.)	5 (Detr.)	7 (Corr.)	7 (Detr.)	8 (Corr.)	8 (Detr.)	9 (M)	9 (W)	10 (M)	10 (W)
base – rel	0.307	0.097	0.374	0.011	0.047	0.834	0.013	0.011	0.013	0.047	0.834
base – gen	0.082	0.485	0.012	0.588	0.078	0.083	0.110	0.588	0.110	0.078	0.083

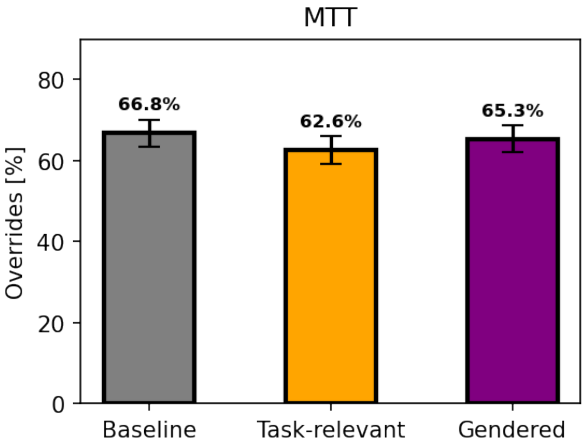


Figure 14: Overrides for MTT. Overrides of AI recommendations that correctly predict a man teacher to be a teacher.

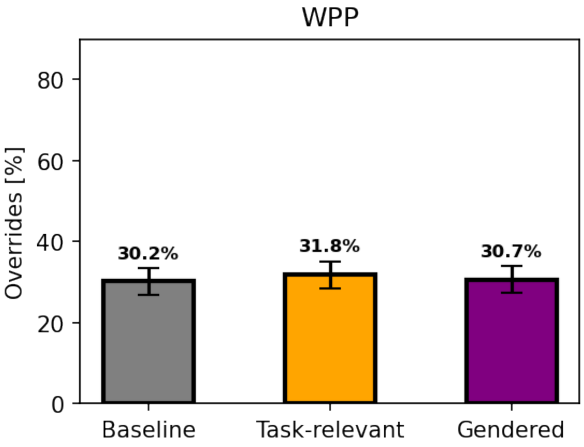


Figure 15: Overrides for WPP. Overrides of AI recommendations that correctly predict a woman professor to be a professor.