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

Jakob Schoeffer[1], **Maria De-Arteaga**[2,*], and **Niklas Kuehl**[3,*]

[1]Karlsruhe Institute of Technology, `jakob.schoeffer@kit.edu`
[2]University of Texas at Austin, `dearteaga@mccombs.utexas.edu`
[3]Karlsruhe Institute of Technology, `niklas.kuehl@kit.edu`
[*]Equal contribution

Abstract

Explanations have been framed as an essential feature for better and fairer human-AI decision-making. In the context of fairness, this has not been appropriately studied, as prior works have mostly evaluated explanations based on their effects on people's perceptions. We argue, however, that for explanations to promote fairer decisions, they must enable humans to discern correct and wrong AI recommendations. To validate our conceptual arguments, we conduct an empirical study to examine the relationship between explanations, fairness perceptions, and reliance behavior. Our findings show that explanations influence people's fairness perceptions, which, in turn, affect reliance. However, we observe that low fairness perceptions lead to more overrides of AI recommendations, regardless of whether they are correct or wrong. This (i) raises doubts about the usefulness of existing explanations for enhancing distributive fairness, and, (ii) makes an important case for why perceptions must not be confused as a proxy for appropriate reliance.

1 Introduction

Artificial intelligence (AI)-based systems are commonly used for informing decision-making in consequential areas such as hiring [65], lending [121], or clinical care [120], where they provide human decision-makers with decision recommendations. The human is then tasked to decide whether to adhere to this recommendation or override it. However, discerning correct and wrong AI recommendations can be difficult; and the explainable AI (XAI) community has proposed a plethora of explanation techniques as a potential aid. Adadi and Berrada [1], for instance, note, "[...] XAI is essential if users are to understand, appropriately trust, and effectively manage AI results." The running assumption is that explanations should enable the human to complement the AI by overriding mistaken recommendations. In reality, however, it remains unclear as to whether existing explainability techniques can live up to these promises [70, 106].

Research gap In particular, explanations are often framed as an essential pathway towards improving fairness of AI-informed decisions. For instance, in a recent Forbes article [61], 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" [30]. Empirical evidence on explanations' ability to enhance fairness is, however, inconclusive [70]. Prior work has found that people's perceptions towards an AI system are influenced by the features that a system is considering in its decision-making process [48, 69, 112]. For instance, if explanations were to highlight the importance of sensitive features (e.g., gender or race), it is likely that people will perceive such a system as unfair [48, 73, 122]. However, researchers have challenged the assumption that "unawareness" of an AI with regard to sensitive information will generally lead to fairer outcomes, and shown that simple interventions can shift people's preference in favor of *including* such information [85]. Moreover, the relationship between people's perceptions and their ability to override wrong AI recommendations and adhere to correct ones—i.e., to *appropriately rely* on the AI—is not well understood.

Our work In this work we examine the interplay of explanations, perceptions, and appropriate reliance on AI recommendations; and we argue that claims regarding explanations' ability to improve algorithmic fairness should, first and foremost, be evaluated against their ability to foster appropriate reliance—i.e., enable people to override wrong AI recommendations and adhere to correct ones. To empirically support our conceptual arguments, we conduct a randomized experiment with 600 participants for the task of occupation prediction from short bios, using a publicly available real-world dataset [24]. We address the following questions:

RQ1 How do explanations affect people's fairness perceptions towards an AI system?

RQ2 What is the relationship between people's fairness perceptions and their reliance on AI recommendations?

RQ3 What is explanations' role in enabling appropriate reliance in human-AI decision-making?

In our experiment, we assess differences in perceptions and reliance behavior when humans 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 in the context of occupation prediction, for which we train two AI models with access to different vocabularies. We randomly assign study 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. In both cases, study participants 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. Ultimately, we test for differences in perceptions and reliance behavior across conditions, and infer implications for the appropriate characterization of explanations' role in human-AI decision-making.

Findings and implications **First**, we do not observe any significant differences in overall task performance across conditions, i.e., study participants are not making more (or less) accurate decisions in the conditions with explanations compared to the baseline without explanations. We see, however, differences in reliance behavior: task performance in the *gendered* condition is composed of (i) less adherence to correct AI recommendations but (ii) more overriding of wrong AI recommendations, compared to the *task-relevant* condition and the baseline. In other words, people who see explanations highlighting the importance of gendered words override more AI recommendations, but the increase in overriding is independent of whether the AI is correct or wrong. In real-world settings, this type of reliance behavior would lead to fairer decisions only if it were desirable to override the AI based on its use of sensitive features (here: gendered words). However, prior research has shown that "fairness through unawareness" is neither a necessary nor sufficient condition for algorithmic fairness [7, 63, 31, 92, 23, 85]. Our findings, thus, challenge the common claim that existing explainability techniques are an enabler for algorithmic fairness. **Second**, we empirically show that while there is a strong negative relationship between fairness perceptions and reliance on AI recommendations, this affects both correct and wrong AI recommendations. This means that people who perceive an AI model as unfair override its recommendations more often, irrespective of their correctness. Moreover, we confirm prior works' findings by observing that study participants' fairness perceptions are significantly lower when explanations highlight gendered words compared to task-relevant words, but show that this is not a meaningful proxy to anticipate appropriate reliance. Overall, these findings suggest that fairness perceptions mediate the relationship between explanations and people's reliance on AI recommendations, but that perceptions solely influence the quantity of overrides and do *not* correlate with appropriate reliance. Hence, our study makes a specific and important case for why perceptions must not be confused as a proxy for appropriate reliance.

2 Background

In this section, we provide background and review related literature. First, we address the role of explanations in human-AI decision-making; then, we summarize prior work on the relationships between explanations and appropriate reliance as well as fairness.

2.1 Explanations of AI

Goals of explanations AI systems are getting increasingly complex and opaque, and researchers and policymakers have been calling for explanations to make AI systems more understandable to human stakeholders [82, 70, 36]. The EU GDPR, for instance, states that decision-subjects have the right to "meaningful information about the logic involved" [36] when consequential automated profiling takes place. Goodman and Flaxman [43], more concretely, argue that "any adequate explanation would, at a minimum, provide an account of how input features relate to predictions," referring to the GDPR requirements. 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 can be attributed to one or more different types of stakeholders [70, 97, 33]. For instance, system designers might be interested in facilitating trust in and acceptance of their systems through explanations, whereas a regulator likely wants to assess a system's compliance with moral and ethical standards [70]. These goals may, however, sometimes be impossible to accomplish simultaneously [117].

Types of explanations The scientific literature distinguishes explanations that aim at explaining individual predictions (*local* explanations) from explanations regarding the whole AI model (*global* explanations) [51]. However, it has been argued that combining local explanations can also lead to an understanding of global model behavior [77]. So-called *local model-agnostic* explanations, such as LIME [101] or SHAP [76], have gained popularity in the literature [1]. LIME, for instance, measures how much an AI model's prediction for a given data point changes when "wiggling" the feature values of this data point. The assumption is that if the prediction changes significantly, then the perturbed feature is locally important. These methods can, among others, generate saliency maps for computer vision tasks or a highlighting of important words for text classification. Other feature importance-type explanation techniques have been proposed, for instance, by Arras et al. [8] specifically for text classification, or by Altmann et al. [4] based on permutation importance [15]. For a comprehensive review of explainability methods we refer to [9, 1].

Criticism of explanations While explanations have been framed as the "sine qua non [i.e., necessary condition] for AI to continue making steady progress without disruption" [1], their realized impact has yet to be proven [27]. Out of all the desiderata of explanations that have been claimed or proposed in the literature, most of them are insufficiently studied or met with inconclusive or (seemingly) contradictory empirical findings [70]. A major line of criticism stems from the fact that explanations can mislead people: drawing on the concept of *dark UX design patterns* [44], Chromik et al. [21] discuss situations where system designers may create interfaces or misleading explanations to purposefully deceive more vulnerable stakeholders like auditors or decision-subjects. This could be accomplished through *adversarial attacks* on explanation methods—approaches to generate explanations that can nudge their receivers into trusting AI models by sacrificing some degree of faithfulness of the explanation to the underlying AI model [116, 69, 98, 29]. Pruthi et al. [98], for instance, construct attention-based explanations that can mask the use of problematic information (e.g., gender or race) by diminishing the weights assigned to such features. Similarly, Lakkaraju and Bastani [69] construct misleading explanations by leveraging correlations between problematic and legitimate features. In the extreme case of placebic explanations (i.e., explanations that convey no information about the underlying AI), Eiband et al. [35] 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 [34] highlight several challenges arising from unanticipated negative downstream effects of explanations, such as misplaced trust in AI, over- or underestimating the AI's capabilities, or over-reliance on certain explanation types.

2.2 Explanations and (appropriate) reliance

Effects on overall performance It has been argued that explanations are an enabler for better human-AI decision-making (e.g., [9, 30, 62, 39, 99]). A recent meta-study [106] on the effectiveness of explanations, however, implies that explanations in most empirical studies did not yield any significant benefits with respect to human-AI decision-making performance. Alufaisan et al. [5], for instance, find no conclusive evidence of explanations' influence on decision accuracy in income and recidivism prediction, and observe that explanations do not enable humans to detect when the AI was correct or incorrect. Green and Chen [46] and Narayanan et al. [84] similarly observe that explanations do not improve performance, and Liu et al. [75] find that interactive explanations do not remedy this. Lai and Tan [66], on the other hand, find that explanations greatly enhance decision-making performance for the case of deception detection. It is noteworthy, however, that a performance increase through explanations may solely be due to (i) an overall increase in adherence to a high-performing AI, or (ii) an overall decrease in adherence to a low-performing AI.

Effects on appropriate reliance In the context of human-AI decision-making, *appropriate reliance* is generally understood as the behavior of humans of overriding wrong AI recommendations and adhering to correct ones [89, 107]. To promote appropriate reliance, explanations must enable humans to distinguish correct from wrong AI recommendations. It has been claimed, for instance, that "transparency mechanisms also function to help users to learn about how the system works, so they can evaluate the *correctness* of the outputs they experience and identify outputs that are incorrect" [99]. The empirical state of evidence, however, is less clear. Lai et al. [68] conduct a survey of empirical findings with respect to human-AI decision-making, and note that prior works have assessed reliance through a variety of inconsistent metrics and tasks. Poursabzi-Sangdeh et al. [96] analyze human-AI decision-making for the case of house price estimation and find that explanations are detrimental to appropriate reliance—likely due to information overload. Bansal et al. [10] find that providing explanations increases humans' adherence to AI recommendations, regardless of their correctness—a phenomenon often referred to as *over-reliance* [37] or *automation bias* [40]. Over-reliance through explanations has also been observed in the contexts of clinical decision support [17], fake review detection [107], diabetes self-management [123], and others. On the other end of the reliance spectrum is *under-reliance* (or *algorithm aversion* [28]); the tendency to ignore or override the AI even when it is correct. Kim et al.

[60] find that humans tend to exhibit algorithm aversion when they see the AI err shortly after they start using it. Jussupow et al. [59] mention three additional characteristics that influence aversion: high algorithm agency (e.g., fully autonomous systems), perceived lack of algorithm capabilities (e.g., lack of empathy), and low human involvement (e.g., no human quality assurance). Parasuraman and Riley [88] argue that distrust is another major driver of algorithm aversion.

Conflation of reliance and trust Many studies have treated reliance and trust interchangeably [68], sometimes calling reliance a "behavioral trust measure" [87]. Similar to the effect on reliance, Dzindolet et al. [32] find that explanations can both increase and decrease trust. However, definitions of trust—especially in HCI—are inconsistent [87, 72, 58]. Some sources have defined *trust* as the extent to which the trustee believes that an automated system will behave as expected [41, 87]. Others have stressed the importance of vulnerability of the trustee [79], as well the anticipation that the AI will adhere to some "implicit or explicit contract" [58]. Such terminological inconsistencies make empirical findings challenging to compare. More importantly, however, trust and reliance are different constructs [68]: reliance is the *behavior* of adhering to or overriding an AI recommendation; whereas trust is a subjective *attitude* regarding the whole system, which builds up and develops over time [100, 126, 88]. It has been argued that trust may impact adoption and reliance [32, 72, 113], but trust in an AI system is not a sufficient requirement for reliance when other factors, such as time constraints, perceived risk, or self-confidence, impact decision-making [72, 102, 25]. The discrepancy between trust and reliance has been shown empirically by Papenmeier et al. [87], who only find a weak positive correlation in their study on offensive Tweet detection. Hence, keeping trust and reliance separate might be beneficial for explainability research [104].

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 [26]. Examples of such behavior include race and gender stereotyping in job ad delivery [57], unfair treatment of people with disabilities in the hiring process [18], or discrimination of Latinx and African-American borrowers in algorithmic mortgage loan pricing [12]. When AI systems are used to inform such consequential decisions, it is important that a human can override problematic recommendations. To that end, the literature has been framing explanations as an important pathway towards improving fairness in human-AI decision-making. According to Langer et al. [70], approximately 10% of all claims around desiderata for explanations have been made with respect to assessing and increasing a system's fairness as well as its compliance with moral and ethical standards—only trailing claims on trust (15%). One representative claim is that "explainability should be considered as a bridge to avoid the unfair or unethical use of algorithms' outputs" [9]. However, there is no conclusive evidence showing that explanations lead to fairer decisions, and it remains unclear *how* explanations are to enable this [70].

Fairness perceptions Instead, prior work has mostly focused on assessing how people *perceive* the fairness of AI systems [118, 68]. Binns et al. [13] compare fairness perceptions across different explanation styles and scenarios—with inconclusive findings. In the study of Angerschmid et al. [6], explanations increase fairness perceptions, but the authors note that perceptions are use-case dependent. Dodge et al. [30] find that people perceive global and local explanations differently, but conclude that the effect of explanations depends on "the kinds of fairness issues and user profiles." Similarly, Shulner-Tal et al. [114] suggest that some explanations are more beneficial than others, but perceptions mainly depend on the outcome of the system. Surprisingly few works have examined downstream effects of fairness perceptions on actual behavior. Green and Chen [46, 45] study the interaction of lay people with risk assessment tools, and find that (i) people's judgment of the tools' accuracy and fairness are unrelated to their *actual* accuracy and fairness, and (ii) people make disparate use of the tools to the detriment of Black people. However, their focus is not on explanations.

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 [48, 50, 49, 122, 95, 85]. This type of information is, for instance, conveyed by feature importance-based explanations like LIME. In a series of studies, Grgić-Hlača et al. [49, 50, 48] examine which features people deem unfair in criminal justice and law enforcement cases, and find that these are mostly in line with what is typically regarded as *sensitive* or *protected* information, such as gender or race [23]. This has been confirmed in other cases like targeted advertising [95] or lending [112]. Nyarko et al. [85] conduct several empirical studies with lay people, computer scientists, and lawyers, and also observe that people are generally averse to the use of gender and race in AI-informed decision-making. Interestingly, people's perceptions towards these features change after they learn that "blinding" the AI to these features can lead to *worse* outcomes for marginalized groups. Along similar lines, Grgić-Hlača et al. [48] and Harrison et al. [53] find that people's perceptions towards the inclusion of sensitive features switch when they are told that this inclusion makes an AI model more accurate [48] or equalizes error rates across demographic groups [53]. In fact, prior research has shown that prohibiting an AI from using sensitive information is neither a necessary nor sufficient requirement for fair decision-making [7, 63, 31, 92, 23, 85], and provided real-world examples where the inclusion of sensitive features can make historically disadvantaged groups like Black people or women better off [94, 23, 115, 80].

3 Explanations, fairness, and appropriate reliance

As noted in Section 2, explanations are often touted for their importance towards improving algorithmic fairness in human-AI decision-making. There are many dimensions of fairness. Grounded on the organizational justice literature [22, 47], researchers have distinguished three main notions of algorithmic fairness: (i) *distributive fairness*, which refers to the fairness of decision outcomes [127], (ii) *procedural fairness*, which refers to the fairness of decision-making procedures [73], and (iii) *informational fairness*, which refers to the adequacy of information conveyed during these procedures [112].

Either implicitly or explicitly, a significant amount of research and deployment seeking to use explanations to advance fairness is concerned with *distributive fairness*, which is judged according to how fair decisions are in their effect on the distribution of benefits and resources [83]. Here, "fair" is typically defined in terms of statistical metrics such as disparities in error rates across groups [52, 20]; which is also referred to as *group fairness* [11]. When considering the effect of human-AI decision-making on distributive fairness, humans' ability to augment fairness depends on their ability to discern correct and wrong AI recommendations and to *appropriately rely* on them—i.e., to adhere to the AI when its recommendations are correct and to override whenever wrong, and doing so in a way that reduces disparities across groups.

However, studies considering the relationship between explanations and fairness have centered on perceptions. This approach has been criticized for its flaws with regard to *procedural fairness*. Specifically, it has been shown that explanations can mislead people's perceptions. By manipulating the faithfulness of explanations to the underlying AI models, Lakkaraju and Bastani [69], for instance, construct explanations to deceive people into trusting models that make decisions based on sensitive information (e.g., race or gender). This type of procedural "fairwashing" [2] of AI models can be achieved by exploiting correlations between sensitive and (seemingly) innocuous features. Ehsan and Riedl [34] discuss cases where uncalibrated perceptions can emerge even in the absence of any malicious intent.

In this work, we argue that perceptions are also an unreliable measure in regards to *distributive fairness*. Prior research has shown that fairness perceptions are influenced by knowing whether an AI system considers sensitive information like race and gender in its decision-making process [122, 85, 49, 50, 48, 95]. The inclusion or exclusion of sensitive information, however, has no bearing on distributive fairness, because sensitive information can be used both to the benefit or the detriment of demographic groups. Thus, in order for explanations to be a reliable pathway towards improving fairness, they must enable humans to better differentiate between correct and incorrect predictions.

Worryingly, there is a lack of research studying the relationship between perceptions and reliance [111], as illustrated in Figure 1. We conjecture that people's fairness perceptions influence their reliance behavior, but not their *appropriate* reliance. This would render perceptions an invalid measure when assessing the effect of explanations on distributive fairness. Furthermore, it would render explanations that influence perceptions but do not facilitate reasoning over algorithmic errors an unreliable mechanism for improving distributive fairness. We empirically test this hypothesis through a human subject study. Our study design is such that we can confirm prior findings on the impact of explanations on fairness perceptions. We then analyze people's reliance behavior given their perceptions and show that, indeed, fairness perceptions are not an indicator of appropriate reliance; i.e., they do not enable humans to discern correct and wrong AI recommendations.

4 Study design

In this section, we first describe in Section 4.1 the task and dataset that we use in our study. Then, we outline our experimental setup in Section 4.2, and we summarize how we recruited study participants in Section 4.3. After that, in Section 4.4, we explain the construction process of the AI models that we use for generating recommendations, and in Section 4.5 we describe how we selected the bios that we use in our questionnaires.

4.1 Task and dataset

Task Automating parts of the hiring funnel has become common practice of many companies; especially the sourcing of candidates online [14, 103]. 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

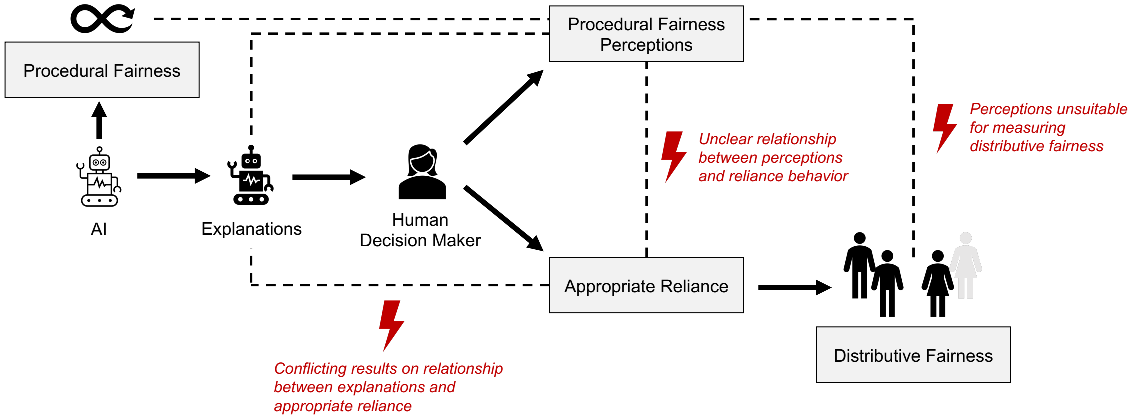

Procedural Fairness
Procedural Fairness Perceptions
AI
Explanations
Human Decision Maker
Unclear relationship between perceptions and reliance behavior
Perceptions unsuitable for measuring distributive fairness
Appropriate Reliance
Distributive Fairness
Conflicting results on relationship between explanations and appropriate reliance

Figure 1: **Conceptual summary of explanations, fairness, and appropriate reliance.** The considerations of the relationships between explanations, fairness, and appropriate reliance raise multiple challenges. Dotted lines with the lightning symbol indicate unclear or problematic relationships that we tackle in this work.

readily available in structured format and would, instead, have to be inferred from unstructured information found online. While this process lends itself to the use AI-based systems, it is susceptible to gender bias and discrimination [24, 14, 103]. De-Arteaga et al. [24] 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, woman rappers are significantly more often misclassified than man rappers because the occupation *rapper* is heavily man-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) [24]. Such misclassifications in hiring have tremendous repercussions for affected people because they may be systematically excluded from exposure to relevant jobs or other professional opportunities.

In our study, we instantiate a human-AI decision-making setup where study participants see short textual bios and are asked—with the help of an AI—to predict whether a given bio belongs to a professor or a teacher. Professors are historically a man-dominated occupation, whereas teachers have been mostly associated with women [82].[1]

Dataset We use the publicly available BIOS dataset, which contains approximately 400,000 online bios from the Common Crawl corpus, initially created by De-Arteaga et al. [24].[2] This data set has been used in other human-AI decision-making studies as well, such as [75, 93]. For each bio in the dataset we know the gender of the corresponding person as well as the true occupation. We note that the 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. As discussed, we only consider the subset of 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 [130, 129, 82], we have more man (55%) than woman (45%) bios of professors and more woman (60%) than man (40%) bios of teachers. We explain how we train AI models on the dataset in Section 4.4 and how we select the bios for our questionnaires in Section 4.5.

4.2 Experimental setup

General setup Study 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 study participants to conditions where they see recommendations and explanations either from

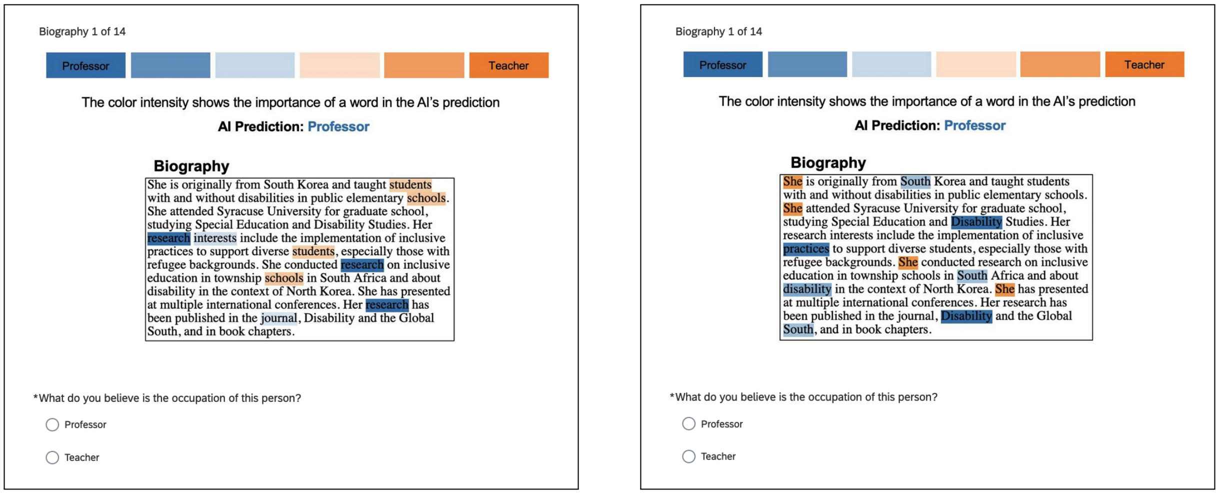
- an AI that uses *task-relevant* features, or

[1] See also [130, 129] on current demographic statistics for professors and teachers in the US.

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

- an AI that uses *gendered* (i.e., sensitive) features.

An exemplary bio including explanations is depicted in Figure 2. Note that the AI predictions and explanations stem from actual AI models that agree in their predictions in the 14 instances shown to participants; we outline the construction of these models in Section 4.4. Study participants in each condition first complete the task of predicting occupations (professor vs. teacher) for 14 such bios, and then—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, study participants provide some demographic information. A summary of our general setup in illustrated in Figure 3. Note that we ask about fairness perceptions *after* the task is completed, so as to prevent these questions from moderating respondents' reliance behavior [19].



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

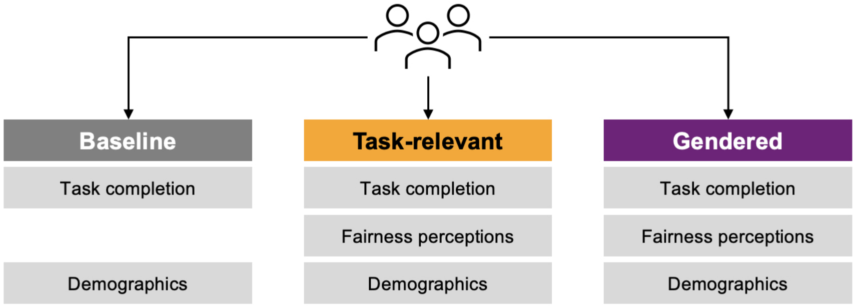


Figure 3: **Illustration of our experimental setup.** Study participants are randomly assigned to one of three conditions. In each condition, they first complete the task of predicting occupations from 14 short bios, and complete a demographic survey at the end. In the conditions with explanations (i.e., *task-relevant* and *gendered*), study participants are also asked about their fairness perceptions after completing the task.

Task completion Figure 2 shows the interface that study 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 human-AI text classification [67, 105, 75]. Study participants in the *task-relevant* and the *gendered* condition are confronted with 14 bios similar to the one in Figure 2, whereas study 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, study 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 also Section 4.3).

In order to be able to assess differences in appropriate reliance across conditions, study 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 study 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 woman teachers (WTT) and man professors (MPP), as well as 3 cases of wrongly recommended woman professors (WPT) and man 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. [60], we include the WPP and MTT scenarios early on in our questionnaires. Precisely, we randomize the order in which study 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 woman teachers are classified as professors, or where man professors are classified as teachers.

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

In our assessment of appropriate reliance, we distinguish four cases of reliance behavior, 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 *wrongful adherence*, to cases where humans override correct recommendations as *detrimental overriding*, and to cases where humans override wrong recommendations as *corrective overriding*. This taxonomy is similar to the one proposed by Liu et al. [75] for trust; however, we want to stress the difference between trust and reliance.

Table 2: **Different types of reliance on AI recommendations.** We distinguish four types of reliance in human-AI 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	Wrongful adherence	Corrective overriding

Fairness perceptions To measure fairness perceptions, we provide a brief introduction and then ask study 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 [22] 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 [22] and slightly rephrased to fit our case of human-AI decision-making. These items have been frequently used in other human-AI studies (e.g., [13, 110, 109, 78]). Colquitt and Rodell [22] propose up to eight measurement items for procedural justice in the organizational psychology context; however, several of these items (e.g., *Procedures offer opportunities for appeals of outcomes*) 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 of study participants, we ask (3) last and without possibility to modify responses for (1) and (2) retroactively.

4.3 Data collection

Our study has received clearance from an institutional ethics committee. Study participants were recruited via `Prolific`—a crowdworking platform for online research [86]. We required study participants to be at least 18 years of age, and to be fluent in English. We also sampled approximately equal amounts of men and women; no other pre-screeners were applied. After consenting to the terms of our study, 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 study 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. With respect to 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 study 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, study participants were paid on average £10.58 (approx. $12.70 at the time the study was conducted) per hour, excluding individual bonus payments of £0.05 per correctly predicted occupation. Study participants took on average 10:12min (baseline), 12:51min (*task-relevant*), and 12:27min (*gendered*) to complete the survey.

4.4 Task-relevant and gendered classifiers

In this subsection, we explain 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` [90]. In trial runs, we found that increasing n beyond 5000 does not significantly change the classifiers' behavior. 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 man and woman 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 man and woman 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 [90].
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 4.5). 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 [101].

4.5 Selection of bios

Pre-selection As outlined in Section 4.2, study 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 4.2.

5 Results and analysis

We now present results from our empirical study and relate them to our research questions. First, in Section 5.1, we present results on the effects of explanations on task performance as well as overriding behavior. In Section 5.2, we then present the effects on fairness perceptions. Finally, in Section 5.3, we analyze how fairness perceptions relate to overriding behavior—especially appropriate reliance.

5.1 Effects of explanations on task performance and overriding behavior

Effects on task performance First, we examine how task performance may be different between the baseline and the conditions with explanations, *task-relevant* and *gendered*. Mean task performances per condition are $M_{base} = 59.49\%$

($SD_{base} = 13.11$),[3] $M_{rel} = 56.94\%$ ($SD_{rel} = 13.86$), and $M_{gen} = 57.96$ ($SD_{gen} = 14.30$), as shown in Figure 4. We conduct a one-way ANOVA as omnibus test for differences across conditions. At $p = 0.179$, we cannot reject the null hypothesis that task performance means are equal across conditions. Hence, we conclude that there are no significant differences in task performance, regardless of whether people were shown the *task-relevant* explanations, the *gendered* explanations, or no explanations at all. This suggests that explanations did not aid human-AI decision-making when measured in terms of overall accuracy.

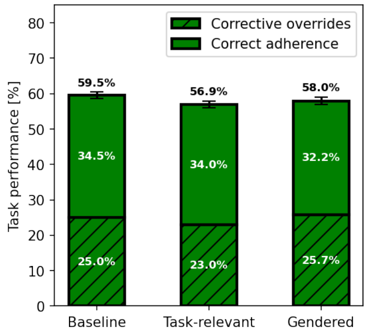


Figure 4: **Task performance by condition.** Differences are not significant across condition. Error bars represent standard errors.

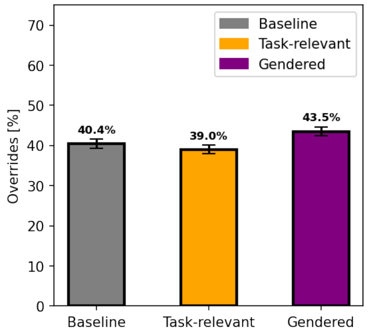


Figure 5: **Overrides by condition.** Overrides are significantly higher in the *gendered* condition vs. *task-relevant* and the baseline.

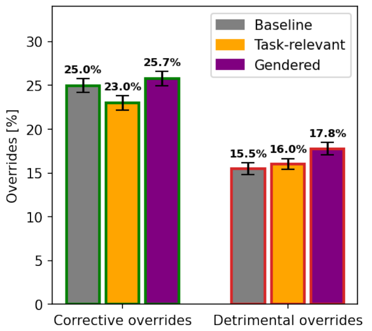


Figure 6: **Overriding behavior.** Both corrective (green border) and detrimental (red border) overrides are higher in the *gendered* condition.

Effects on overriding behavior However, we observe that the task performance in the *gendered* condition is composed of comparably less correct adherence to and more corrective overriding of AI recommendations. We examine the differences in overriding behavior between conditions further in Figures 5 and 6, where we see that study participants in the *gendered* condition override more AI recommendations than in the *task-relevant* condition ($p = 0.003$) and the baseline ($p = 0.041$). From Figure 6 we further conclude that *both* corrective *and* detrimental overrides are highest in the *gendered* condition, with detrimental overrides being significantly higher than in the *task-relevant* condition ($p = 0.031$) and the baseline ($p = 0.008$). In the *task-relevant* condition, we see that overall overrides are lowest across conditions (see Figure 5), with corrective overrides being significantly lower than the baseline ($p = 0.042$); the difference in detrimental overrides between *task-relevant* condition and baseline is not significant. Hence, even though the difference in total adherence to AI recommendations is not significantly higher in the *task-relevant* condition compared to the baseline ($p = 0.153$), we observe a tendency towards automation bias—especially because people in the *task-relevant* condition show an increased adherence to the AI in cases where it would be beneficial to override. Overall, we conclude that people's reliance behavior is affected by how the AI explains its recommendations; specifically, people override the AI significantly more often when sensitive features are highlighted. However, both corrective *and* detrimental overrides increase and neutralize each other at the performance level, indicating that gendered explanations fostered algorithm aversion.

5.2 Effects of explanations on fairness perceptions

Study participants in the *task-relevant* and *gendered* conditions have significantly different ($p = 1.37 \times 10^{-23}$) perceptions of fairness towards the AI. 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. Recall that we measure all items on 5-point Likert scales, ranging from 1 (unfair) to 5 (fair). This means that people who are shown a highlighting of task-relevant words perceive the underlying AI as fairer than people who are shown gendered words as being important for given AI recommendations. Figure 7 shows the distribution of perceptions in both conditions. We also display fitted third-order polynomials to highlight the difference between conditions. Interestingly, we also see several responses suggesting that the use of gendered words for predicting occupations was fair. Understanding such perceptions better will be an interesting aspect for future research. Overall, we confirm prior works' findings and conclude the following:

[3] 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 our conditions.

Regarding **RQ1**, our findings suggest that an AI is perceived as significantly less fair when explanations point at the use of sensitive features compared to cases where explanations point at task-relevant features.

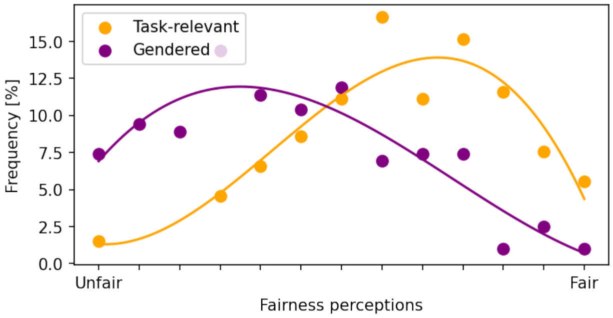


Figure 7: **Distribution of fairness perceptions.** Fairness perceptions are significantly higher in the *task-relevant* condition compared to the *gendered* condition. We also fitted third-order polynomials to highlight the differences.

5.3 Effects of fairness perceptions on overriding behavior

When we look at people's overriding behavior as a function of their fairness perceptions, we find an overall strong negative relationship ($p = 1.10 \times 10^{-11}$) between fairness perceptions and overriding of AI recommendations, i.e., study participants override the AI more often when their fairness perceptions are lower. Concretely, we see that people override on average 52% of AI recommendations when their fairness perceptions are lowest, and only 31% when their fairness perceptions are highest. This negative relationship is consistent in both the *task-relevant* ($p = 8.19 \times 10^{-4}$) and the *gendered* ($p = 1.97 \times 10^{-7}$) condition, and it also persists when we disentangle corrective ($p = 1.45 \times 10^{-10}$) and detrimental ($p = 7.40 \times 10^{-4}$) overrides at the aggregate level. Figure 8 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 8 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 report all results from OLS regressions of overrides on perceptions in Table 3. Interestingly, we observe that as study participants override 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. This is visualized in Figure 9; and we note that the slope of the fitted regression line there is not significantly different from zero ($p = 0.935$). Overall, we conclude the following:

Regarding **RQ2**, our findings suggest that people's fairness perceptions influence their reliance behavior in a way that low perceptions lead to more overrides than high perceptions. However, both corrective *and* detrimental overrides increase as fairness perceptions decrease.

Aggregating the results from Section 5.1 and Section 5.3, we further conclude:

Regarding **RQ3**, our findings suggest that explanations that point at the use of sensitive features lead to low fairness perceptions, which, in turn, translate to more overriding of AI recommendations, regardless of their correctness—i.e., algorithm aversion.

6 Discussion and conclusion

In this section, we discuss broader implications of our work. First, we briefly summarize our main findings. After that, we discuss implications for the assessment and design of explanations. Finally, we address limitations and conclude our manuscript.

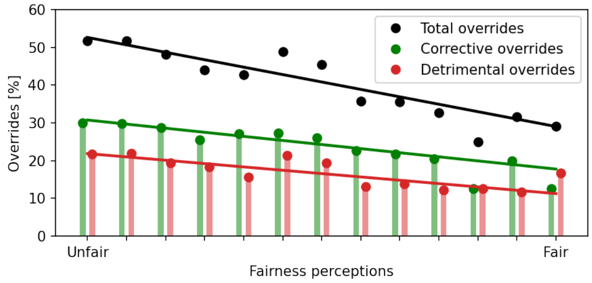


Figure 8: **Overrides over perceptions (*gendered*).** Significant negative relationship between fairness perceptions and overrides, both corrective and detrimental, as well as overall.

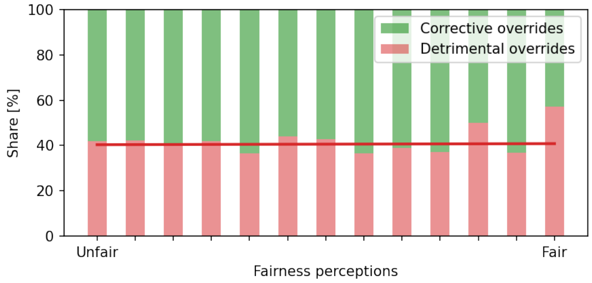


Figure 9: **Type of overrides over perceptions (*gendered*).** Ratio of corrective to detrimental overrides is approximately 60:40, independent of fairness perceptions.

Table 3: **OLS regression table.** Results for regressions of overrides (in %) on fairness perceptions, overall and disaggregated by *gendered* and *task-relevant* condition. Standard errors are provided in parentheses.

	Overall$_1$	**Overall$_2$**	**Overall$_3$**	**Gendered$_1$**	**Gendered$_2$**	**Gendered$_3$**	**Task-rel.$_1$**	**Task-rel.$_2$**	**Task-rel.$_3$**
Dep. variable	Tot. overr.	Corr. overr.	Detr. overr.	Tot. overr.	Corr. overr.	Detr. overr.	Tot. overr.	Corr. overr.	Detr. overr.
Constant	56.56***	35.08***	21.48***	58.52***	34.00***	24.52***	53.82***	38.50***	15.31***
	(2.31)	(1.72)	(1.42)	(2.98)	(2.22)	(1.87)	(4.48)	(3.35)	(2.64)
Perceptions	-5.04***	-3.53***	-1.51***	-5.90***	-3.25***	-2.65***	-4.20***	-4.40***	0.21
	(0.72)	(0.54)	(0.44)	(1.09)	(0.81)	(0.69)	(1.24)	(0.92)	(0.73)
p-value (F-stat.)	0.000	0.000	0.001	0.000	0.000	0.000	0.001	0.000	0.079
R^2 (adj.)	0.107	0.096	0.026	0.123	0.069	0.065	0.051	0.100	-0.005
N	400	400	400	202	202	202	198	198	198

*$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Summary of our findings and implications In this work, we argue that explanations can only foster distributive fairness if they enable people to discern correct and wrong AI recommendations—i.e., foster appropriate reliance. Our empirical findings, however, show that while explanations influence reliance, they need not lead to appropriate reliance, casting doubt on their reliability as a mechanism towards distributive fairness. Our study design lets us also draw conclusions about the role of fairness perceptions. We first confirm prior works' findings showing that people's fairness perceptions towards the AI are lower when explanations point at the importance of sensitive features in the decision-making process, compared to explanations that highlight task-relevant words. Moreover, we show that perceptions influence people's reliance on AI recommendations such that low perceptions lead to more overrides. Crucially, however, we find that when explanations highlight sensitive words, this leads to both corrective and detrimental overrides—a phenomenon sometimes called *algorithm aversion*. Hence, perceptions mediate the effect of explanations on reliance behavior, but not on appropriate reliance. The main implications of these results are twofold: first, they challenge the common claim that explanations are an enabler for better and fairer human-AI decision-making; second, they highlight that fairness perceptions are not a valid proxy for appropriate reliance and, hence, do not imply an ability to promote distributive fairness.

Implications for assessing explanations Our work has several implications for assessing the effectiveness of explanations. Our main argument is that claims around explanations fostering distributive fairness must be assessed against explanations' ability to enable appropriate reliance first, as opposed to their effects on perceptions. Importantly, while research on fairness perceptions has been thriving [118], our findings show that fairness perceptions have no bearing on people's ability to appropriately rely on AI recommendations and, hence, foster distributive fairness. We thus suggest that fairness perceptions must not be conflated with distributive fairness.

While we do not argue against evaluating fairness perceptions (this may be of interest in many cases), we urge all stakeholders of AI systems to keep in mind the following: first, our results—in line with previous work [48, 50, 49, 85, 95, 112, 122]—suggest that fairness perceptions are strongly influenced by whether people think that an AI system uses sensitive features or not. While this is not worrisome per se, it has been shown that explanations can be exploited to "mask" the use of such sensitive information [29, 69, 98]. This means that system designers with malicious intents can use explanations to deceive people into perceiving an AI system as fair when in reality it may be procedurally unfair.

Second, the notion of "fairness through unawareness", which deems an AI model to be fair if it does not make use of information that is evidently indicative of a person's demographics, has been shown to be deeply flawed. For explanations to constitute a meaningful pathway to improving fairness, they must go beyond a human-in-the-loop operationalization of "fairness through unawareness". Nyarko et al. [85] note that justifications for such a notion of fairness are often along one of two ethical principles: first, people may oppose the use of sensitive information like gender and race based on *deontological* principles. Under such an account, it may be argued that using demographic information like gender or race to inform consequential decisions (e.g., the allocation of certain goods or benefits) is itself fundamentally unethical [119]. On the other hand, it may also be the case that people oppose the use of sensitive information under a *consequentialist* account. A justification for excluding sensitive information under this principle assumes that "unawareness" will lead to fairer outcomes and likely benefit historically disadvantaged groups. This, however, is not generally true [31, 63, 23, 92]. While there are several examples—both inside and outside of AI—where "scrubbing" sensitive information can reduce biases (e.g., "unaware" orchestra auditions to increase gender diversity [42]), other examples exist where including such information can have the same effect (e.g., affirmative action [56]). Hence, fairness perceptions may be based on false assumptions.

Third, and perhaps most importantly, perceptions are strongly related to reliance behavior but do *not* imply the ability to discern correct and wrong AI recommendations. Hence, fairness perceptions are not a reliable indicator of distributive fairness and should not be treated as such. Moreover, since we know that perceptions can be misled through explanations, this also means that explanations can manipulate *reliance* behavior, either towards algorithm aversion (low perceptions) or—the other extreme—automation bias (high perceptions).

Overall, given several concerns with measuring and interpreting perceptions, we urge that researchers clearly articulate when and why perceptions are a meaningful objective in assessing human-AI decision-making. In line with the argumentation by Lai et al. [68], our findings also stress the general importance of separating subjective versus objective (i.e., behavioral) measurements. While prior work has found that, for instance, trust and reliance behavior must not be conflated [87, 72], our work emphasizes the importance of explicitly distinguishing fairness perceptions from distributive fairness when measuring the effects of interventions—such as explanations—on fairness properties. The choice for measuring either subjective attitudes or objective behavior should be clearly stated and reflected in the constructs being measured.

Implications for designing explanations Our research has several implications for system designers creating future AI applications as well as policymakers who shape the surrounding legal frameworks, like the European Artificial Intelligence Act [64] or the American Artificial Intelligence Initiative Act [81].

Systems designers typically follow a narrative of implementing explanations to "open the black box" of AI systems [91]. However, what this means is not often clear. 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 [74]. Seen through this lens, our work emphasizes the importance of designing explanations with the explicit purpose of facilitating appropriate reliance by helping humans identify and reason about algorithmic errors, 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. Lebovitz et al. [71] 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. They point out that if AI is to add value, it will sometimes make recommendations that are at odds with the experts' knowledge, increasing their uncertainty. Thus, appropriate reliance requires processes and tools that help them reconcile both views. To this point, it is not clear that what is necessary are always explanations as they relate to AI's inner workings, instead of explanations of the broader socio-technical system. For instance, interventions that help humans reason over the information that is and is not available to the algorithm may help them reconcile disagreement and better integrate multiple sources of information [54, 55]. Furthermore, additional interventions such as cognitive forcing functions have been shown to foster more productive reliance [16].

In the context of trust, Schlicker and Langer [108] have conceptually examined discrepencies between *perceived* trust and *actual* trustworthiness of AI systems. They draw on different psychological theories [125, 38] to identify four dimensions of cues based on which people form their perceptions about the *actual* trustworthiness of a system: cue relevance, cue availability, cue detection, and cue utilization. It could be helpful for system designers to evaluate explanations along these dimensions. Concretely, they should ask *what* cues an explanation provides, and *whether* they are relevant towards achieving a certain desideratum (e.g., fairness). For the case of commonly-used explanations like LIME or SHAP, for instance, our work suggests that there is a fundamental mismatch in relevance between the cues these explanations provide and the goal of enabling people to enhance algorithmic fairness.

Policymakers currently urge for making AI applications more transparent [3, 124]. The political will for providing more transparency is important and welcome, but policymakers should be careful when demanding explanations without contextualizing what they are supposed to achieve. As laid out, implementing explanations to increase fairness perceptions can backfire and increase algorithm aversion without giving involved decision-makers the right instruments to actually make improved decisions. Future policies should aim at increasing the effective collaboration of human-AI teams [128] without the involved human parties drifting into either algorithm aversion or automation bias. Finally, our work stresses that the much-cited "right to explanation" is not a sufficient condition for ensuring the responsible use of AI in decision-making. Only if policies facilitate explanations which lead to better and fairer outcomes, a meaningful symbiosis between human and AI entity can thrive.

Limitations and outlook We acknowledge several limitations of our work and outline directions for follow-up research. First, our study setup assigns people to either the *gendered* or the *task-relevant* condition; i.e., study participants see either only explanations with a 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, these cases would likely be mixed. Concretely, an AI might sometimes highlight only sensitive features, sometimes only task-relevant features, and at other times a mix of both. While our study design does not explicitly account for such cases, even if perceptions vary at the instance level, our findings suggest that reliance would depend on the inclusion of sensitive features, which is flawed for reasons outlined above. However, future work that studies how instance-level perceptions relate to aggregate-level perceptions, and how these interdependencies shape reliance behavior, is important to inform the design of novel explanation frameworks and socio-technical systems that facilitate productive human-AI decision-making.

Our work shows a study in which explanations do not enable appropriate reliance. In particular, it shows that while we replicate prior results of explanations' impact on perceptions, this does not translate to the desirable reliance behavior that is usually implied when making claims about explainablity and fairness. This, however, should not be taken to mean that explanations never enable appropriate reliance. 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.

References

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

[2] Ulrich Aïvodji, Hiromi Arai, Olivier Fortineau, Sébastien Gambs, Satoshi Hara, and Alain Tapp. Fairwashing: The risk of rationalization. In *International Conference on Machine Learning*, pages 161–170. PMLR, 2019.

[3] Gabriele Spina Alì and Ronald Yu. Artificial intelligence between transparency and secrecy: From the EC whitepaper to the AIA and beyond. *European Journal of Law and Technology*, 12(3), 2021.

[4] André Altmann, Laura Toloşi, Oliver Sander, and Thomas Lengauer. Permutation importance: A corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

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

[6] Alessa Angerschmid, Jianlong Zhou, Kevin Theuermann, Fang Chen, and Andreas Holzinger. Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction*, 4(2):556–579, 2022.

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

[8] Leila Arras, Franziska Horn, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. "What is relevant in a text document?": An interpretable machine learning approach. *PLOS ONE*, 12(8):e0181142, 2017.

[9] 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. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58:82–115, 2020.

[10] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 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*, pages 1–16, 2021.

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

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

[13] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 'It's reducing a human being to a percentage'; Perceptions of justice in algorithmic decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

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

[15] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[16] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–21, 2021.

[17] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. The role of explanations on trust and reliance in clinical decision support systems. In *2015 International Conference on Healthcare Informatics*, pages 160–169. IEEE, 2015.

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

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

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

[21] Michael Chromik, Malin Eiband, Sarah Theres Völkel, and Daniel Buschek. Dark patterns of explainability, transparency, and user control for intelligent systems. In *IUI Workshops*, volume 2327, 2019.

[22] Jason A Colquitt and Jessica B Rodell. Measuring justice and fairness. 2015.

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

[24] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 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*, pages 120–128, 2019.

[25] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–12, 2020.

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

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

[28] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General*, 144(1):114, 2015.

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

[30] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. Explaining models: An empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 275–285, 2019.

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

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

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

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

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

[36] European Union. General Data Protection Regulation. 2016. URL `https://eur-lex.europa.eu/eli/reg/2016/679/oj`.

[37] A Michael Froomkin, Ian Kerr, and Joelle Pineau. When AIs outperform doctors: Confronting the challenges of a tort-induced over-reliance on machine learning. *Ariz. L. Rev.*, 61:33, 2019.

[38] David C Funder. On the accuracy of personality judgment: A realistic approach. *Psychological Review*, 102 (4):652, 1995.

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

[40] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association*, 19(1):121–127, 2012.

[41] Nazila Gol Mohammadi, Sachar Paulus, Mohamed Bishr, Andreas Metzger, Holger Könnecke, Sandro Hartenstein, Thorsten Weyer, and Klaus Pohl. Trustworthiness attributes and metrics for engineering trusted internet-based software systems. In *International Conference on Cloud Computing and Services Science*, pages 19–35. Springer, 2013.

[42] Claudia Goldin and Cecilia Rouse. Orchestrating impartiality: The impact of "blind" auditions on female musicians. *American Economic Review*, 90(4):715–741, 2000.

[43] Bryce Goodman and Seth Flaxman. European Union regulations on algorithmic decision-making and a "right to explanation". *AI Magazine*, 38(3):50–57, 2017.

[44] Colin M Gray, Yubo Kou, Bryan Battles, Joseph Hoggatt, and Austin L Toombs. The dark (patterns) side of UX design. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2018.

[45] Ben Green and Yiling Chen. Disparate interactions: An algorithm-in-the-loop analysis of fairness in risk assessments. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 90–99, 2019.

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

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

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

[49] Nina Grgić-Hlača, Elissa M Redmiles, Krishna P Gummadi, and Adrian Weller. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. In *Proceedings of the 2018 World Wide Web Conference*, pages 903–912, 2018.

[50] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[51] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR)*, 51(5):1–42, 2018.

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

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

[54] Patrick Hemmer, Max Schemmer, Niklas Kühl, Michael Vössing, and Gerhard Satzger. On the effect of information asymmetry in human-AI teams. *arXiv preprint arXiv:2205.01467*, 2022.

[55] Kenneth Holstein, Maria De-Arteaga, Lakshmi Tumati, and Yanghuidi Cheng. Toward supporting perceptual complementarity in human-AI collaboration via reflection on unobservables. *arXiv preprint arXiv:2207.13834*, 2022.

[56] Harry Holzer and David Neumark. Assessing affirmative action. *Journal of Economic Literature*, 38(3):483–568, 2000.

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

[58] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 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*, pages 624–635, 2021.

[59] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion. *Proceedings of the 28th European Conference on Information Systems (ECIS)*, 2020.

[60] Antino Kim, Mochen Yang, and Jingjing Zhang. When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms. *ACM Trans. Comput.-Hum. Interact.*, 2022. ISSN 1073-0516. doi: 10.1145/3557889.

[61] Jennifer Kite-Powell. Explainable AI is trending and here's why. *Forbes*, 2022.

[62] René F Kizilcec. How much information? Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 2390–2395, 2016.

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

[64] Mauritz Kop. EU Artificial Intelligence Act: The European approach to AI. Stanford-Vienna Transatlantic Technology Law Forum, 2021.

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

[66] Vivian Lai and Chenhao Tan. 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*, pages 29–38, 2019.

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

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

[69] Himabindu Lakkaraju and Osbert Bastani. "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*, pages 79–85, 2020.

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

[71] Sarah Lebovitz, Hila Lifshitz-Assaf, and Natalia Levina. 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):126–148, 2022.

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

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

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

[75] Han Liu, Vivian Lai, and Chenhao Tan. 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):1–45, 2021.

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

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

[78] Frank Marcinkowski, Kimon Kieslich, Christopher Starke, and Marco Lünich. 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*, pages 122–130, 2020.

[79] Roger C Mayer, James H Davis, and F David Schoorman. An integrative model of organizational trust. *Academy of Management Review*, 20(3):709–734, 1995.

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

[81] Jeffrey Mervis. US law sets stage for boost to artificial intelligence research, 2021.

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

[83] Lily Morse, Mike Horia M Teodorescu, Yazeed Awwad, and Gerald C Kane. Do the ends justify the means? Variation in the distributive and procedural fairness of machine learning algorithms. *Journal of Business Ethics*, pages 1–13, 2021.

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

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

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

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

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

[89] Samir Passi and Mihaela Vorvoreanu. Overreliance on AI: Literature review. Technical report, Microsoft Research, 2022.

[90] 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. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[91] Dino Pedreschi, Fosca Giannotti, Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, and Franco Turini. Meaningful explanations of black box AI decision systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9780–9784, 2019.

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

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

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

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

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

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

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

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

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

[101] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “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*, pages 1135–1144, 2016.

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

[103] Javier Sánchez-Monedero, Lina Dencik, and Lilian Edwards. 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*, pages 458–468, 2020.

[104] Nicolas Scharowski, Sebastian AC Perrig, Nick von Felten, and Florian Brühlmann. Trust and reliance in XAI – Distinguishing between attitudinal and behavioral measures. *arXiv preprint arXiv:2203.12318*, 2022.

[105] Max Schemmer, Patrick Hemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. Should I follow AI-based advice? Measuring appropriate reliance in human-AI decision-making. *arXiv preprint arXiv:2204.06916*, 2022.

[106] Max Schemmer, Patrick Hemmer, Maximilian Nitsche, Niklas Kühl, and Michael Vössing. A meta-analysis on the utility of explainable artificial intelligence in human-AI decision-making. *arXiv preprint arXiv:2205.05126*, 2022.

[107] Max Schemmer, Niklas Kühl, Carina Benz, and Gerhard Satzger. On the influence of explainable AI on automation bias. *arXiv preprint arXiv:2204.08859*, 2022.

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

[109] Nadine Schlicker, Markus Langer, Sonja K Ötting, Kevin Baum, Cornelius J König, and Dieter Wallach. What to expect from opening up ‘black boxes’? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior*, 122:106837, 2021.

[110] Jakob Schoeffer, Yvette Machowski, and Niklas Kuehl. 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*, 2021.

[111] Jakob Schoeffer, Maria De-Arteaga, and Niklas Kuehl. On the relationship between explanations, fairness perceptions, and decisions. *ACM CHI 2022 Workshop on Human-Centered Explainable AI (HCXAI)*, 2022.

[112] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. “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*, pages 1616–1628, New York, NY, USA, 2022. Association for Computing Machinery. doi: 10.1145/3531146.3533218.

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

[114] Avital Shulner-Tal, Tsvi Kuflik, and Doron Kliger. 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):1–13, 2022.

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

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

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

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

[119] Sonja B Starr. Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.*, 66: 803, 2014.

[120] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American Journal of Surgical Pathology*, 42(12):1636, 2018.

[121] Sian Townson. AI can make bank loans more fair. *Harvard Business Review*, 2020.

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

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

[124] Darrell M West and John R Allen. *Turning point: Policymaking in the era of artificial intelligence*. Brookings Institution Press, 2020.

[125] Bernhard Wolf. Brunswik's original lens model. *University of Landau, Germany*, 9:1–9, 2005.

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

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

[128] Qiaoning Zhang, Matthew L Lee, and Scott Carter. You complete me: Human-AI teams and complementary expertise. In *CHI Conference on Human Factors in Computing Systems*, pages 1–28, 2022.

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

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