What is known about the impact of algorithm bias on the perceived fairness: A Scoping Review

Amirhossein Hajigholam Saryazdi

Supervisor: Dr. Mahdi Mirhoseini

Abstract

Background: Artificial intelligence (AI) based algorithm increasingly shapes our daily life. These algorithms are not seamless and can be biased in a way that compounding injustice in society. This affects people's perceived fairness of the algorithm, but little is known about its mechanism and scope.

Methods: A comprehensive scoping review of published academic articles was performed to answer the **question**: how algorithmic bias impacts perceived fairness of users. The review was accomplished in **three steps**: (1) identifying the question and relevant literature; (2) selecting the literature; (3) charting, collating, and summarizing the information. Overall themes and concepts were identified from this process.

Results: The query of search terms in all searchable fields resulted in 41 papers from all the searched sources. After reviewing the abstracts and applying inclusion criteria, 12 works are fully reviewed. A further 10 articles were included for review after hand searching reference lists (backward/forward search). The result discussed around seven themes identified from the English-language academic literature namely: (1) algorithm bias; (2) algorithm fairness; (3) perceived fairness; (4) individual characteristics; (5) social characteristics; (6) task characteristics; (7) technology characteristics.

Conclusions: Shedding light on the relationship between algorithm bias and perceived fairness was the initial goal of this study. Although previous works indicated how deviating from fairness metrics impacts perceived metrics, we are among the first who proposes a theoretical framework on how algorithm (model) biases influence algorithm fairness and what are the contextual factors.

1. Introduction

Artificial Intelligence (AI) algorithms are extensively employed by businesses, governments, and organizations to make decisions regarding individuals and society. Despite the enormous benefits of utilizing AI in the decision-making process, their black-boxed nature may entail hard-to-detect ethical risks at different levels. A main ethical concern is that AI algorithms are likely to reproduce and strengthen the biases that exist in society (O'Neil, 2019). These biases can lead to discrimination toward certain population groups and have been already observed in a variety of cases. For example, the COMPAS system for forecasting the re-offending risk was found to calculate higher risk for black defendants compared to their white counterparts, everything else was equal (Angwin et al., 2016)(racial bias). A further example, a machine learning (ML) advertisement tool on the internet was observed to serve significantly fewer ads for high-paid job positions to women than to men (Datta et al., 2015) (gender bias). This violation of the norms of

Electronic copy available at: https://ssrn.com/abstract=4437243

justice and equality is referred to as **algorithm bias** and happens when the result of an ML-based model advantage or disadvantage certain people more than others without a reasonable reason for such unfair impacts (Kordzadeh & Ghasemaghaei, 2022). Although computer scientists have been studying mathematical techniques to cure biases in machine learning algorithms and fairness is a pilar of algorithm development in trustworthy AI area (Chatila et al., 2021) (Ntoutsi et al., 2020), how people understand this phenomenon and what are the contextual factors influencing this understanding are still not fully understood. Shedding light on this subject and discovering fairness perception of algorithms, provide a valuable insight about its social implication such as trust and acceptance of these models.

Perception of algorithms is increasingly receiving scholarly attention in various fields such as human-computer interaction (HCI) and information systems (IS). It has been discussed that people build mental models about algorithmic systems regardless of how they actually work, and so, a mathematically-proven fair algorithm may not be perceived as fair for a group of people because it may not match the concepts of fairness of them (M. K. Lee, 2018). An essential dimension of algorithm perception is fairness perception that influence people's attitude (e.g. trust) and behaviors (e.g. acceptance) (Acikgoz et al., 2020). **Perceived fairness** refers to the extent to which the process and output of an algorithmic decision making are perceived to be fair (Saxena et al., 2019). Although mentioning philosophical conceptions of fairness is beyond the purpose of this work, one dominant philosophical theory of fairness is developed by the 20th-century philosopher John Rawls, who equated fairness and justice (Rawls, 1999), stating generally that fairness is "a demand for impartiality". Since then, researchers have often used "justice" and "fairness" interchangeably (Colquitt & Rodell, 2015). In the field of algorithmic decision-making, scholars mainly use the term "fairness" over "justice". This work does so to be in harmony with the literature and to differentiate algorithmic bias, as a computationally measured construct, and perceived fairness, as a subjective one (Kordzadeh & Ghasemaghaei, 2022; Robert et al., 2020).

In the most similar work to this study, Kordzadeh & Ghasemaghaei (2022) well identified algorithm bias as an antecedent of fairness perception. However, how algorithm bias impacts perceived fairness and what are the specific factors moderating this relationship is crucial are not fully understood. Thus, the goal of this paper is to shed light on this mechanism by reviewing related literature.

2. Method

The review poses a broad question: what is known about the impact of algorithm bias on perceived fairness? This work follows the scoping review protocol suggested by Arksey and O'Malley (2005). In the rest of this section the steps undertaken to accomplish this review is presented.

2.1. Identifying the Question and Relevant Literature

The first stage is to develop the scoping question, which was done through research team meetings to discover a potentially rewarding and beneficial topic to focus on within the area of human computer interaction. Next, a search strategy defined for the identification of relevant literature. To achieve a full coverage, Web of science is used to search through any kind of published work by all the publishers without applying any filters or time constraint. The list of publishers includes but not limited to Elsevier, Springer, IEEE, ACM, Wiley, Sage, Taylor and Francis, Emerald, etc. Papers published in the proceedings of the International Conference on Information Systems (ICIS) are also included since this IS conference has mini tracks on AI and fairness. Plus, The

Electronic copy available at: https://ssrn.com/abstract=4437243

Social Science Research Network (SSRN) website was explored to check for the papers that are not generally peer-reviewed but have the potential to have meaningful theoretical and practical contributions. Finally, after applying inclusion criteria and during the full text review stage, the references and citations of the articles were manually reviewed to expand the search and ensure no related work has been missed (backward/forward).

The two main keywords in the research question ("algorithm bias" and "perceived fairness") provide the basis for the search terms. "Pearl-growing" method is adopted for additional terms for the subsequent search (Booth et al., 2016). It is based on relevant papers ("pearls") to detect additional relevant search terms or keywords. For this study, two review articles were initially identified as "pearls" because of their relevancy to the research question and being widely cited in the literature: (Kordzadeh & Ghasemaghaei, 2022; Starke et al, 2022). Relevant keywords are added to the search terms we derived from our research question, and they were searched across all searchable field in the database including title, abstract, keywords, and text. The query used in this stage is provided in *Appendix 1*.

Although this study falls within the realm of IS, no field discrimination was applied and all papers regardless of their theme have been collected. This process resulted in a total of 40 results in the preliminary dataset.

2.2 Selecting the literature: post-hoc Inclusion/Exclusion

In this stage, the articles' titles, keywords, and abstracts were screened to exclude conceptually or contextually irrelevant studies. Then, 'Post hoc' inclusion criteria were developed and applied. Such a criterion is central to the scoping review procedure as it is unlikely to recognize grounds for exclusion or inclusion at the beginning, and this is one of the main differences between the scoping and systematic review methods (Arksey and O'Malley, 2005).

A publication would be considered relevant if 1) it mostly focused on both notion of algorithmic bias and perceived fairness and 2) the conceptualizations of these two notions were consistent with the sociotechnical description of them that emphases on social biases integrated into algorithms that may result in discriminatory outcomes for individual or groups (Domanski, 2019). This explanation has been extensively applied in the IS, computer science, data science, and law literature. For instance, a paper that only focuses on either "algorithmic bias" or "perceived fairness" or uses these terms or their similar constructs only in its list of keywords, did not meet the first requirement. Similarly, a paper that studied the role of bias in the algorithmic calculation of food consumption did not meet the second criterion. As part of the exclusion procedure, abstracts, panel reports, posters, and proposals were also removed from the papers list, as they did not provide major insights. Moreover, highly technical papers were excluded from the analysis due to their primary purpose of providing mathematical solutions or theorizations. They provided computational proofs or tested their proposed techniques using hypothetical or real-world datasets to demonstrate the effectiveness of their techniques in developing less biased and more fair algorithms. These topics, however, were beyond the scope of this study.

After applying the above-mentioned criteria and filtering based on abstract review, the remained 12 sources were reviewed in full. Backward and forward search of the references and citations are also applied manually with the help of Google Scholar during the full-text review to make sure no relevant article has been missed. This process resulted in adding more 14 papers. The summary of all the work is shown in *Table 1*.

Electronic copy available at: https://ssrn.com/abstract=4437243

Table 1: Summary of the reviewed papers

Paper	Theory	Algorithm Bias	Algorithm Fairness	Perceived Fairness	Method	Case	Context factors
(Gupta et al., 2022)	NA	Racial bias Gender bias	NA	Single-item measure (5-point Likert scale)	Survey N=387 MTurk Control (Age, Gender, Daily Internet usage)	Financial service Recruitment Facial recognition Hotel or flight booking Blocking online content Healthcare space	Enviro. Ch.: Cultural values
(Bonezzi & Ostinelli, 2021)	NA	Racial and gender disparities	NA	Single-item measure (7-point Likert scale)	N=150	Uni admission Hiring Parole	Task Ch.: AI vs Human
(Angerschmid et al., 2022)	NA	Gender bias	Gender fairness	Single-item measure (4-point Likert scale)	N=25	Health insurance decision Medical treatment	Tech Ch.: explainability
(Hannan et al., 2021)	NA	NA	NA	Conjoint analysis with likert-scale questions	N=747	Service allocation	Task Ch.: high/low impact
(van Berkel et al., 2021)	NA	NA	NA	Scale 0-100%	N=80 Mixed model COMPAS Lendingclub	Recidivism Lending	Tech Ch: Visualizati Ind Ch.: Gender, Education

Electronic copy available at: https://ssrn.com/abstract=4437243

(Shin, 2021)	NA	NA	NA	NA	N=350 Survey		Tech. Ch: Explainability
(R. Wang et al., 2020)	NA	Biased outcome: high error rates in protected groups. Unbiased outcome: similar error rates across different groups.	NA	Six questions about fairness with 7-point likert	Randomized between subject experiment N=590	Awarding a promotion as a worker	Ind. Ch.: outcome favorability, Education Tech Ch.
(M. K. Lee, 2018)	Folk theory	NA	NA	7 point likert	N= 321	Mech: Work assignment/ Scheduling Human: hiring/ Work evalu	Task Ch.: Mechanic vs human skills required
(Saxena et al., 2019)	NA	NA	Calibrated: Selecting individuals in proportion to their merit	NA	Between subject experiment	Loan decision	Ind Ch.
(Grgic-Hlaca et al., 2018)	NA	NA	NA	Assessing the fairness of using different features	N=576	COMPAS	Task ch
(Kordzadeh & Ghasemaghaei, 2022)	Stimulus organism theory	NA	NA	NA	Systematic literature review	Algorithmic bias, perceived fairness,	NA

Electronic copy available at: https://ssrn.com/abstract=4437243

	Organizational Justice					behavioral response	
(Mehrabi et al., 2022)		NA	NA	NA	Survey	NA	Algorithm bias, Algorithm fairness
(M. K. Lee et al., 2017)	NA	NA	Efficient: min distance Equality: equal chance Equity: higher chance for greater needs	NA	Semi-structured interview. N=31	Food allocation	
(G. Wang et al., 2023)	NA	NA	Procedural, informational, distributive	7 point likert	Survey experiment N=1360	Public affair	Task Ch.: Rule driven vs data driven
(Grgić-Hlača et al., 2022)	NA	NA	Procedural		N=	Bail decision based on COMPAS	Ind. Ch.: Political view, experience
(Bankins et al., 2022)	NA	NA	Interactional justice: belief of being treated with dignity and respect	Colquitt's four item scale	Experimental survey N=638 Control (demographic and work experience)	HRM	Task Ch.: AI vs Human
(C. Wang et al., 2022)	NA	Gender bias	NA	NA	Between subject experiment	College major recommendation	Ind. Ch.: Human bias

Electronic copy available at: https://ssrn.com/abstract=4437243

					N=200 college students		
(M. K. Lee & Rich, 2021)	NA	NA	NA	NA	Between subject experiment N=280	Healthcare	Ind Ch.: Mistrust in human systems
(Marcinkowski et al., 2020)	NA	NA	Procedural, distributive	NA	340 Germany	Higher education admission	Human vs algorithm
(Acikgoz et al., 2020)	Organizational justice, Signaling, Fairness Heuristic	NA	Procedural Interactional	Selection Procedural Justice Scale (SPJS) (Bauer et al., 2001)	N=320	Job interview	Task Ch.: AI vs Human
(M. K. Lee et al., 2019)	NA	NA	Procedural	Likert	Within subject lab study	Food division	Tech Ch.: Explanation, Outcome control

3. Discussion: Algorithm Fairness VS Algorithm Bias

Although algorithm fairness was not among the search terms of this study, it appears remarkably in the AI decision-making field and made the team to pursue it in the forward/backward stage. Thus, it is worthy to clarify the differences between algorithmic bias and algorithm fairness. They are related concepts that have received considerable attention in the literature on algorithmic decision-making. Algorithm bias refers to the presence of unfair or discriminatory outcomes in algorithmic decision-making. This can occur when the model trained on biased data or the model itself is prone to generate biases due to its design (Barocas & Selbst, 2016). On the other hand, algorithm fairness, refers to the absence of unfair or discriminatory outcomes in algorithmic decision-making. It can be achieved through a variety of techniques, such as designing algorithms to be transparent, allowing for input from affected communities, and using techniques such as counterfactual analysis to test for potential biases (Dwork et al., 2012).

Additionally, fairness is a normative concept, meaning that it is based on a set of values and principles, while bias is more of a descriptive concept, describing what is actually happening in a particular context (Friedler et al., 2019). Another important difference is that bias is often seen as a technical problem that can be addressed through better data collection, model design, and testing,

Electronic copy available at: https://ssrn.com/abstract=4437243

while fairness is seen as a social and ethical problem that requires engagement with affected communities and consideration of social values (Whittaker et al., 2018).

Algorithm bias is mainly discussed in the literature through identifying the types of bias and then suggesting mitigation solutions (Barocas & Selbst, 2016). However, Algorithm fairness is trying to operationalize the notion of fairness in AI by providing equality metrics (Dwork et al., 2012). Algorithm fairness metrics are used to quantify the effect of algorithmic bias in the algorithmic decision making.

4. Results

To give an overview of the studies, a large group of researchers take *normative (prescriptive)* approach by surveying their respondents regarding the fairness of a scenario, assuming that there is a collective agreement around what creates fairness concept as an objective matter. Another group, pursue a descriptive (comparative) approach to explore human perception of fairness in technology-based decision making (Grgic-Hlaca et al., 2018).

This section presents the results including themes and theoretical concepts identified from the review. Based on the knowledge obtained from the literature, propositions are developed regarding the concepts' relationship. The theoretical model is presented in figure 1. To be consistent with the literature, the contextual factors of the framework is presented in a similar way to Kordzadeh & Ghasemaghaei (2022).

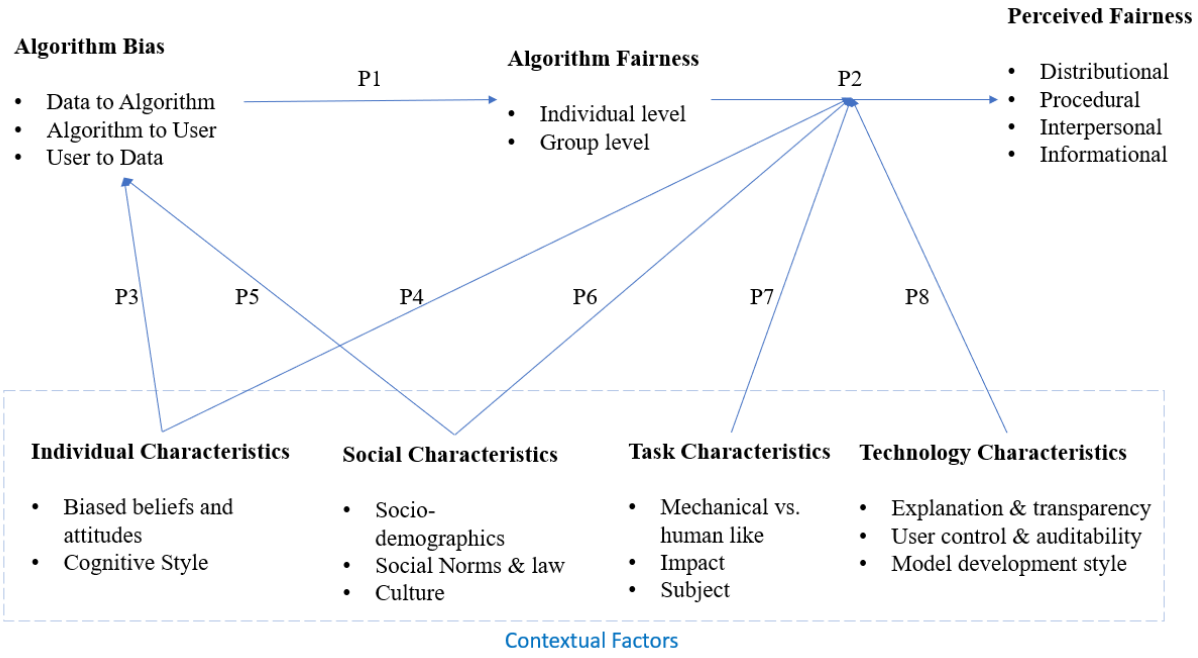

Figure 1: Theoretical model

4.1. Algorithmic bias

Despite different definitions for algorithm bias, there are two major aspects in common across all definitions: 1) the output of a biased algorithm reflects the deviation from an equality principle leading to discrimination, and 2) that deviation happens systematically and repeatably, not at

Electronic copy available at: https://ssrn.com/abstract=4437243

random. Thus, algorithmic bias is a systematic deviation from equality that appears in the outputs of an algorithm (Kordzadeh & Ghasemaghaei, 2022). Algorithmic bias can be rooted in different stages of developing an ML model. For example, it may be due to the unbalanced biased data which representing a well-established bias from society. Or it may happen due to incorporating a sensitive feature (e.g., gender, race, etc.) into the calculation. Computer scientists are trying to mitigate these biases at different stages of model development (Ntoutsi et al., 2020). Table 2 indicates the most important sources of algorithm bias under three categories.

Table 2 Sources of algorithm bias from Mehrabi et. al (2022)

Category	Type of Bias	Definition
Data to Algorithm	Measurement Bias	Ex. Using irrelevant proxy to measure an attribute to be used in model
	Omitted Variable Bias	one or more important variables are left out of the model
	Representation/Sampling Bias	Non-representative or non-random samples fails to be diverse, inclusive, and generalizable.
	Aggregation Bias (Ecological Fallacy)	Drawing false conclusions about individuals from observing the entire population and ignoring individual or subgroup differences.
	Longitudinal Data Fallacy	The heterogeneous cohorts can bias cross-sectional analysis of temporal data, resulting in different conclusions than longitudinal analysis
	Linking Bias	When user behavior is misrepresented by network attributes derived from their connections, activities, or interactions.
Algorithm to User	Algorithm Bias	When the bias is not existed in the input data and is added solely by the algorithm.
	User Interaction Bias	A bias due to the user interface or user itself by imposing their self-selected biased behavior.
	Popularity Bias	Popular items tend to be exposed more, but popularity criteria can be subject to manipulation by, for example, fake reviews.
	Emergent Bias	When the bias occurs because of interacting with users. For example, after A change in population or values.
	Evaluation Bias	
User to Data	Historical Bias	When biases that are already existed in the world find its way in model through data even after a careful sampling.
	Population Bias	When the user population of the platform is different from the original target population

Electronic copy available at: https://ssrn.com/abstract=4437243

		statistically, demographically, and characteristically.
	Self-selection Bias	When individual can voluntarily to participate in a study
	Temporal Bias	Variation in populations and behaviors over time
	Content Production Bias	Structural, lexical, semantic, and syntactic differences in the content created by users due to different background, age, gender, etc.
	Behavioral Bias	Different user behavior across platforms, datasets, or contexts
	Social Bias	When we act under the influence of observing others' action

4.2. Algorithm fairness

Defining algorithm fairness is depending on the definition of equality and justice which varies across different philosophical paradigms, culture, religion or legal systems (Saxena et al., 2019). Researcher mostly categorize fairness notion in algorithmic context into individual-level, group-level, and sub-group level. Individual level is concerned with similar outcomes for individuals who have similar qualifications with respect to a task. Group-level, aka statistical notion of fairness, seeks to ensure that no group is being disproportionately affected by algorithm outcomes in a task. Individual-level and group-level notion of fairness can be conflicting. For example, if an employee using a technology-based model for recruiting and want the model to equalize the number of men and women for the interview. When there is not enough qualified candidate in one group, this model give interview to an unqualified candidate with the cost of ignoring qualified candidate in the other group (Binns, 2020). Thus, subgroup fairness is introduced that intends to address this conflict by using the best properties of the group and individual notions of fairness (Kearns et al., 2018). Table 3 presents different algorithmic fairness definitions and their relations to the above-mentioned categories. It is important to note that it is impossible to satisfy all fairness constraints for a task and the proper one should be selected based on the context (Kleinberg et al., 2016).

Table 3: Algorithm fairness notions, adapted from Mehrabi et. al (2021)

Category	**Fairness notion/ metric**	**Description**	**Constraint**
Group Fairness	Demographic parity	Selection probabilities should be equal between disadvantaged and advantaged groups.	$Pr(\hat{Y} = 1 \mid A = 0) = Pr(\hat{Y} = 1 \mid A = 1)$
	Conditional statistical parity	Selection probabilities should be equal between disadvantaged and	$Pr(\hat{Y} = 1 \mid L=1, A = 0) = Pr(\hat{Y} = 1 \mid L=1, A = 1)$

Electronic copy available at: https://ssrn.com/abstract=4437243

		advantaged groups given a set of legitimate factors L.	
	Predictive Parity	Positive predictive values should be equal between disadvantaged and advantaged groups	$Pr(Y = 1 \mid A = 0, \hat{Y} = 1) = Pr(Y = 1 \mid A = 1, \hat{Y} = 1)$
	Equalized odds	True positive rates and false positive rates should be equal between disadvantaged and advantaged groups	$Pr(\hat{Y} = 1 \mid A = 0, Y = 1) = Pr(\hat{Y} = 1 \mid A = 1, Y = 1)$, $Pr(\hat{Y} = 1 \mid A = 0, Y = 0) = Pr(\hat{Y} = 1 \mid A = 1, Y = 0)$
	Equal opportunity	True positive rates should be equal between disadvantaged and advantaged groups.	$Pr(\hat{Y} = 1 \mid A = 0, Y = 1) = Pr(\hat{Y} = 1 \mid A = 1, Y = 1)$
	Treatment equality (Error rate balance)	False positive rates and false negative rates should be equal between disadvantaged and advantaged groups.	$Pr(\hat{Y} = 1 \mid A = 0, Y = 0) = Pr(\hat{Y} = 1 \mid A = 1, Y = 0)$, $Pr(\hat{Y} = 0 \mid A = 0, Y = 1) = Pr(\hat{Y} = 0 \mid A = 1, Y = 1)$
	Test fairness (Within group Calibration)	A score S = S(x) is test fair (well-calibrated) if it reflects the same likelihood irrespective of the individual's group membership, R)	$P(Y =1\|S=s, R=b)=P(Y =1\|S=s, R=w)$
Individual Fairness	Fairness through unawareness	Protected attributes are not explicitly used in the decision-making process.	
	Fairness through awareness	Individuals who are similar based on a similarity metric defined for a particular task should receive a similar outcome.	
	Counterfactual fairness	An algorithm is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group	

Algorithm biases mentioned in the previous section impact fairness metrics. As presented in table 2, algorithm bias is either rooted in data or the model itself. Thus, a group of scholars argue that *"an algorithm is only as good as the data it works with"* (Barocas & Selbst, 2016). If a model trained on data biased toward a particular group, for example, due to sampling or historical bias or lack of balance, the outcome perpetuates this bias and lead to violating fairness-related metrics (Chouldechova, 2017). However, supplying a neutral data is not enough to guarantee algorithm fairness. ML models can produce biased outcome by using features not related to the task. This type of bias is more nuanced because those features are not necessarily sensitive attributes such as race or gender, in fact, they may look innocent at the first glance. T. Wang et al. (2019) discover that image processing and visual detection models for tasks such as gender detection are prone to associate objects presents in the image with the task and use them as an attribute to train on. For

Electronic copy available at: https://ssrn.com/abstract=4437243

example, if there are some photos of a woman in a kitchen, the model link kitchen utensils to the gender.

Proposition 1: Algorithm bias negatively influences algorithm fairness.

4.3. Perceived fairness

When it comes to defining fairness, there are two schools of thoughts. One is taking prescriptive approach which is seemingly an objective definition of fairness, and the other understands it as a subjective matter which exists in the minds of individuals and being affected by who is being asked to judge, who are the involved groups, and what algorithm is doing (Hannan, Chen, Joseph, et al., 2021). Thus, perceived fairness can be seen as the extent to which an algorithm is perceived to be fair (Saxena et al., 2019; Woodruff et al., 2018). However, when it comes to operationalization, it is contextual dependent as (Srivastava et al., 2019) found that the simple notion of demographic parity is preferable in the contexts of recidivism and medical prediction and laypeople less prefer complex notions that come with cognitive burdens.

One broadly used theory in conceptualizing perceived fairness in algorithmic decisions is ***organizational justice theory*** (Binns et al., 2018; Robert et al., 2020). It considers employee's view and states that justice "reflects the degree to which one's company or top management is perceived to act consistently, equitably, respectfully, and truthfully in decision contexts" (Colquitt & Rodell, 2015). The existing literature on this theory has discovered three commonly used types of fairness: distributive, procedural, and interactional (Greenberg & Colquitt, 2005). Distributive justice is seeking the allocation of outcomes such as pay and other resources whereas procedural justice refers to the perceived fairness of the processes that lead to decision outcomes (Thibaut et al., 1973). Interactional justice is defined under the two elements of interpersonal and informational. Interpersonal refers to the degree that employees are treated with respect and dignity. Informational aspect indicates the degree that information is provided to help employees understand processes taken to achieve fairness. Depending on the context and task, the importance associated with each component of fairness can be changed. For instance, Lee et al. (2019) focus on the process and proposed a framework for procedural justice in algorithmic decisions, and Bankins et al. (2022) aim to study interpersonal aspects of AI decision making.

Since the development of organizational justice theory by Greenberg (1987), it has been customized and used to build models in various non-algorithmic settings such as employment selection (Gilliland, 1993). More recently, these models vae been found useful and are being used in the context of algorithmic decisions such as AI interview (Acikgoz et al., 2020).

Furthermore, Acikgoz et al. (2020) discussed the application of **signaling theory** from the economy (Connelly et al., 2011) and **fairness heuristic theory** (Bos, 2001; Lind et al., 1993) in algorithmic decision setting. From the perspective of signaling theory, using AI in decision-making may be interpreted as a lack of humanity, then lower perceived interactional justice (negative), or as a standard process treating equally, then higher perceived procedural justice (positive). Based on the fairness heuristic theory, people rely on cognitive shortcuts to make decisions about justice perception especially when they are in uncertain situations with limited knowledge. Therefore, signals people receive regarding the algorithms or the information that organizations present about their AI models, impact the formation of the fairness heuristic.

Electronic copy available at: https://ssrn.com/abstract=4437243

Algorithm fairness that mentioned in the previous section can impact perceived fairness. Lee et al. (2017) show that applying equality and equity-based measures (corresponding to group and individual fairness respectively) to algorithm decision making lead leads to different fairness perception in different groups in the context of food allocation. Saxena et al. (2019) also indicate that satisfying calibrated fairness measure in loan distribution enhancing fairness perception of people (distributive fairness).

Proposition 2: Algorithmic fairness positively influences perceived fairness.

4.4. Individual characteristics

Individual characteristics that are pertinent in the context of an algorithmic decision encompass beliefs, attitudes, and socio-demographic characteristics. In this context, individuals denote whoever involved are involved in AI decision-making system from developers and users to people being affected by the outcome. Attitudes and beliefs refer to biased predispositions and stereotypes that individuals may keep regarding particular social groups (N. T. Lee, 2018).

Individuals' beliefs and attitudes can impact bias in the data used to train algorithms. For example, if individuals have preconceived notions or biases about certain groups or individuals, they may unintentionally select data or program an algorithm that reinforces those biases (confirmation bias). Thus, the resulting algorithm may be biased (Mehrabi et al., 2022)

Proposition 3: Individual characteristics impact algorithm biases.

Human biases and stereotypical beliefs can also impact fairness perception of an algorithm. For instance, gender stereotype in career recommendations have a significant effect on accepting a career recommended by an algorithm. In general the results from a gender-aware algorithm had a higher acceptance rate than a gender-neutral model (C. Wang et al., 2022). The two following human biases are identified in the reviewed papers:

Outcome (favorability bias): Bankins et al., (2022) find that although people generally see human decision makers to be more fair than algorithm in the context of HRM, favorable AI decisions is perceived more just than unfavorable human decision. (R. Wang et al., 2020) discover that people perceive the algorithm's decision more just when it predicts in their favor, even if the algorithm being described biased against particular demographic groups. More interestingly, they observe that the impact of favorable outcome on the fairness perception of individuals is greater than maintaining fairness at group level, suggesting that satisfying a group fairness metric can be beaten by outcome (favorability) bias.

In-group bias: This cognitive bias is not limited to individuals themselves. People also shows *group justifying* (or *ego-justifying)* perspective making them to be inclined seeing similar individuals to themselves more deserving of a treatment than others who are different from them (Hannan et al., 2021).

Apart from human biases, on a more general sense, individuals have separate views toward the usage of technology in different settings. Some are skeptical and show more mistrust while some tend to trust more in automation and to ignore its potential pitfalls such as unfairness (M. K. Lee & Rich, 2021). The level of knowledge about the context in which the algorithm is applied can also shape the perception of fairness. G. Wang et al. (2023) discover that when using an algorithm in the public affairs, low familiarity with the context increases the perception of fairness.

Electronic copy available at: https://ssrn.com/abstract=4437243

Moreover, having the experience of attending a bail hearing negatively affects the perceived fairness of using juvenile criminal history in building AI based decision support system (Grgić-Hlača et al., 2022).

Proposition 4: Individual characteristics moderate the relationship between algorithmic fairness and perceived fairness.

4.5. Social characteristics

Like individual biases, culture, social norms, and group biases can find their ways to data and then algorithm bias. Social factors such as the cultural and historical context in which the data was collected, and the biases and assumptions of the individuals who labeled and curated the data influence the input of models. In the same way, when data is collected within a society, structural discrimination and systematic inequalities along with stereotypes from that society can affect the quality of data and the model trained on that (Mehrabi et al., 2021).

Proposition 5: Social characteristics impact algorithmic bias.

Social characteristics also play role in how people perceive the fairness of algorithms. Social psychology research suggests that demographic factors affect people's moral judgments (Graham et al., 2013; Thompson & Loewenstein, 1992). Gender and educational accomplishment significantly influence perceived fairness. Female or higher-educated respondents perceived a set of predictors as less fair than other respondents (van Berkel et al., 2021; R. Wang et al., 2020). In addition to socio-demographic factors, political views influence fairness perception. Grgić-Hlača et al. (2022) show that left-leaning people generally perceive using an algorithm for bail decisions as less fair than their right-leaning counterparts.

Cultural values also moderate the relationship between algorithm bias and perceived fairness. It is discovered that respondents with cultural values associated with collectivism, masculinity, and uncertainty avoidance are more likely to question racial and gender-biased recommendations by AI (Gupta et al., 2022).

Proposition 6: Social characteristics moderate the relationship between algorithmic fairness and perceived fairness.

4.6. Task characteristics

Studies show the characteristics of a task matters to people when they are developing a sense of justice toward the algorithmic decision making. When it comes to applying a decision attribute respondents consider what resource is being allocated and if what would be the impact (low vs. high) of the decision on people's lives. For example, it is understood to be fair to consider a person's background when allocating affordable housing, but not a COVID medication (Hannan et al., 2021).

There is also a rich line of literature that discusses the impact of AI versus human decision-maker on perceived fairness of the outcome, but there is still no consensus among the studies. A group of researchers believe that algorithmic decisions that yield disparities are less likely to be perceived as biased than human decisions. This is because, they argue, people believe algorithms, unlike humans, decontextualize decision-making by abandoning individual characteristics and equally applying rules and regulations regardless of whom they are judging. Bankins et al. (2022) found that having a human decision-maker instead of AI results in higher perception of interactional

Electronic copy available at: https://ssrn.com/abstract=4437243

fairness in the context of human resource management. In the same context, Acikgoz et al. (2020) reveals that AI-based interviewing is generally perceived as less procedurally and interactionally fair than human-based interviewing. However, when it comes to deciding about school admission in higher education, German students evaluated algorithmic decision making (ADM) higher than human decision making (HDM) in terms of both procedural and distributive fairness (Marcinkowski et al., 2020). Building upon this line of literature, decisions can be categorized in two groups based on the associated skills: a group that needs "human" skills (e.g. subjective judgment and emotional capability) and those that requires more ''mechanical'' skills (e.g. processing quantitative data for objective measures). For mechanical tasks, algorithmic and human decisions are equally fair, but for tasks that involve human skills, human decisions are perceived to be fairer (M. Lee, 2018).

Proposition 7: Task characteristics moderate the relationship between algorithmic fairness and perceived fairness.

4.7. Technology characteristics

In the context of algorithmic systems, technology characteristics refers to the interactive and socio-technical characteristics that are incorporated into user interfaces. These features can improve the usability of the systems and allow users to identify and respond to biases more effectively (Binns et al., 2018; Dodge et al., 2019). Explaining the behavior of model in the form of using a n instance of the data (Example-based explanation) or clarifying important features of it (feature importance-based explanation) increases the perception of fairness (Angerschmid et al., 2022; Shin, 2021). However, outcome explanation can have no effect on the perceived fairness when the outcome itself is seen as being unfair. Moreover, it can reduce the perception of fairness when the explanation violating the expectation (M. K. Lee et al., 2019).

In the literature of trustworthy AI, explainability is closely related to the transparency. As mentioned, explainable AI is when the model can explain its finding and reasoning, while transparent AI refers the concept of developing algorithms without hiding the computations and choices behind (Kaur et al., 2023; Li et al., 2023). Interestingly, it is found that the negative effect of algorithm bias on perceived fairness is aggravated by providing a higher level of transparency about the process of model development (R. Wang et al., 2020). Triggering a feeling that the procedure should be fairer, information overload, and increasing the expectations are possible justifications for this effect.

The adoption of outcome control has been found to have a positive impact on the perceived fairness of decision-making. This is largely attributed to the ability of outcome control to facilitate an understanding of the inherent limitations of decision-making and to redistribute resources in a manner that is more aligned with contextual factors. Furthermore, outcome control is recognized for its capacity to integrate human elements into the final decision-making process, thereby increasing the level of perceived fairness (M. K. Lee et al., 2019).

More nuancedly, the way algorithms are developed also influences the perceived fairness. Comparing rule-driven and data-driven approaches in developing decision making models, the former is generally perceived as fairer and more acceptable than the latter (G. Wang et al., 2023). In addition to that, the negative effect of algorithm bias on perceived fairness is exacerbated if the model development procedure is described as outsourced. This may be due to the feeling that internal developers, who are presumably aware of the context, should have has good reasons for

Electronic copy available at: https://ssrn.com/abstract=4437243

biases (e.g., maximizing accuracy) (R. Wang et al., 2020). The means of visualization can also impact the fairness perception of algorithms. A team of researcher shows a model with text-based visualization of the result is perceived fairer than a commonly used scatter plot (van Berkel et al., 2021).

Proposition 8: Technology characteristics moderate the relationship between algorithmic fairness and perceived fairness.

5.Conclusion

It is well studied that algorithms are prone to be biased. With the ever spreading of algorithmic decision making in human's life, understanding people's fairness perception of these models are necessary since it plays a main role in adopting or dropping this technology. In this review, how algorithm bias impacts perceived fairness of users and third persons is studied. This is an interdisciplinary area between the fields of computer science (CS) and information systems (IS) and so, required a careful consideration of approaches and lingos of both sides. Unlike previous works in IS, this study differentiates the constructs of algorithm bias and algorithm fairness based on SC scholars' approach. It is found that algorithm bias affects algorithm fairness and then algorithm fairness impacts perceived fairness of people. Contextual factors of these two relationships are also identified, and interestingly, some are moderating both impacts. Although theoretical model and propositions are developed based on the published research, there is a lack of empirical research in this area.

APPENDIX 1:

(ALL=("algorithm* bias*" OR "algorithm* fair*" OR "algorithm* discriminat*" OR "analy* bias*" OR "analy* fair*" OR "analy* discriminat*" OR "AI bias*" OR "AI fair*" OR "AI discriminat*" OR "ML bias*" OR "ML fair*" OR "ML discriminst*" OR "DL bias*" OR "DL fair*" OR "DL discriminat*" OR "data bias*" OR "data fair*" OR "data discriminat*" OR "auto* bias*" OR "auto* fair*" OR "auto* discriminat*") AND ALL=("perceived fairness" OR "perceived bias" OR "perceived parity" OR "perceived justice" OR "perceived equality" OR "fair* perception*" OR "bias* perception*" OR "parity perception*" OR "justice perception*" OR "equality perception*" OR "fair* understanding" OR "bias* understanding " OR "parity understanding " OR "justice understanding " OR "equality understanding " OR "percept* of bias*" OR "percept* of fair*" OR "percept* of justice*" OR "percept* of equality*" OR "percept* of parity*" OR " understanding of bias*" OR " understanding of fair*" OR " understanding of justice*" OR " understanding of equality*" OR " understanding of parity*" OR "sense of bias*" OR " sense of fair*" OR " sense of justice*" OR " sense of equality*" OR " sense of parity*"))

References:

Acikgoz, Y., Davison, K. H., Compagnone, M., & Laske, M. (2020). Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment*, *28*(4), 399–416. https://doi.org/10.1111/ijsa.12306

Electronic copy available at: https://ssrn.com/abstract=4437243

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *MACHINE LEARNING AND KNOWLEDGE EXTRACTION*, *4*(2), 556–579. https://doi.org/10.3390/make4020026

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias *. In *Ethics of Data and Analytics*. Auerbach Publications.

Arksey and O'Malley. (2005). *Scoping studies: Towards a methodological framework*. https://doi.org/10.1080/1364557032000119616

Bankins, S., Formosa, P., Griep, Y., & Richards, D. (2022). AI Decision Making with Dignity? Contrasting Workers' Justice Perceptions of Human and AI Decision Making in a Human Resource Management Context. *Information Systems Frontiers*, *24*(3), 857–875. https://doi.org/10.1007/s10796-021-10223-8

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *California Law Review*, *104*(3), 671–732.

Binns, R. (2020). On the apparent conflict between individual and group fairness. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 514–524. https://doi.org/10.1145/3351095.3372864

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). *"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions* [Preprint]. SocArXiv. https://osf.io/9wqxr

Bonezzi, A., & Ostinelli, M. (2021). Can Algorithms Legitimize Discrimination? *JOURNAL OF EXPERIMENTAL PSYCHOLOGY-APPLIED*, *27*(2), 447–459. https://doi.org/10.1037/xap0000294

Booth, A., Sutton, A., & Papaioannou, D. (2016). *Systematic approaches to a successful literature review* (Second edition). Sage.

Bos, K. van den. (2001). *Theoretical and Cultural Perspectives on Organizational Justice*. IAP.

Electronic copy available at: https://ssrn.com/abstract=4437243

Buolamwini, J., & Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, 77–91. https://proceedings.mlr.press/v81/buolamwini18a.html

Chatila, R., Dignum, V., Fisher, M., Giannotti, F., Morik, K., Russell, S., & Yeung, K. (2021). Trustworthy AI. In B. Braunschweig & M. Ghallab (Eds.), *Reflections on Artificial Intelligence for Humanity* (Vol. 12600, pp. 13–39). Springer International Publishing. https://doi.org/10.1007/978-3-030-69128-8_2

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

Colquitt, J. A., & Rodell, J. B. (2015). ## Measuring Justice and Fairness. In *The Oxford Handbook of Justice in the Workplace*. Oxford University Press. http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199981410.001.0001/oxfordhb-9780199981410-e-8

Datta, A., Tschantz, M. C., & Datta, A. (2015). *Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination* (arXiv:1408.6491). arXiv. http://arxiv.org/abs/1408.6491

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K. E., & Dugan, C. (2019). Explaining models: An empirical study of how explanations impact fairness judgment. *Proceedings of the 24th International Conference on Intelligent User Interfaces*, 275–285. https://doi.org/10.1145/3301275.3302310

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226. https://doi.org/10.1145/2090236.2090255

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., & Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 329–338. https://doi.org/10.1145/3287560.3287589

Electronic copy available at: https://ssrn.com/abstract=4437243

Gilliland, S. W. (1993). The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. *The Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.2307/258595

Graham, J., Haidt, J., Koleva, S., Motyl, M., Iyer, R., Wojcik, S. P., & Ditto, P. H. (2013). Moral Foundations Theory. In *Advances in Experimental Social Psychology* (Vol. 47, pp. 55–130). Elsevier. https://doi.org/10.1016/B978-0-12-407236-7.00002-4

Greenberg, J. (1987). A Taxonomy of Organizational Justice Theories. *The Academy of Management Review*, *12*(1), 9–22. https://doi.org/10.2307/257990

Greenberg, J., & Colquitt, J. A. (Eds.). (2005). *Handbook of Organizational Justice*. Psychology Press. https://doi.org/10.4324/9780203774847

Grgić-Hlača, N., Lima, G., Weller, A., & Redmiles, E. M. (2022). *Dimensions of Diversity in Human Perceptions of Algorithmic Fairness* (arXiv:2005.00808). arXiv. http://arxiv.org/abs/2005.00808

Grgic-Hlaca, N., Redmiles, E., Gummadi, K., Weller, A., & Assoc Comp Machinery. (2018). Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. *Saarland University*, 903–912. https://doi.org/10.1145/3178876.3186138

Gupta, M., Parra, C., & Dennehy, D. (2022). Questioning Racial and Gender Bias in AI-based Recommendations: Do Espoused National Cultural Values Matter? *INFORMATION SYSTEMS FRONTIERS*, *24*(5), 1465–1481. https://doi.org/10.1007/s10796-021-10156-2

Hannan, J., Chen, H., Joseph, K., & ASSOC COMP MACHINERY. (2021). Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. *State University of New York (SUNY) System*, 555–565. https://doi.org/10.1145/3461702.3462568

Hannan, J., Chen, H.-Y. W., & Joseph, K. (2021). Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, 555–565. https://doi.org/10.1145/3461702.3462568

Hardt, M., Price, E., Price, E., & Srebro, N. (2016). Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems*, *29*.

Electronic copy available at: https://ssrn.com/abstract=4437243

https://proceedings.neurips.cc/paper/2016/hash/9d2682367c3935defcb1f9e247a97c0d-Abstract.html

Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2023). + Trustworthy Artificial Intelligence: A Review. *ACM Computing Surveys*, *55*(2), 1–38. https://doi.org/10.1145/3491209

Kearns, M., Neel, S., Roth, A., & Wu, Z. S. (2018). Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness. *Proceedings of the 35th International Conference on Machine Learning*, 2564–2572. https://proceedings.mlr.press/v80/kearns18a.html

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent Trade-Offs in the Fair Determination of Risk Scores* (arXiv:1609.05807). arXiv. http://arxiv.org/abs/1609.05807

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Lee, M. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *BIG DATA & SOCIETY*, *5*(1). https://doi.org/10.1177/2053951718756684

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

Lee, M. K., Jain, A., Cha, H. J., Ojha, S., & Kusbit, D. (2019). Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1–26. https://doi.org/10.1145/3359284

Lee, M. K., Kim, J. T., & Lizarondo, L. (2017). A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management that Allocates Donations to Non-Profit Organizations. *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, 3365–3376. https://doi.org/10.1145/3025453.3025884

Electronic copy available at: https://ssrn.com/abstract=4437243

Lee, M. K., & Rich, K. (2021). Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3411764.3445570

Lee, N. T. (2018). Detecting racial bias in algorithms and machine learning. *Journal of Information, Communication and Ethics in Society*, *16*(3), 252–260.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). + Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, *55*(9), 1–46. https://doi.org/10.1145/3555803

Lind, E. A., Kulik, C. T., Ambrose, M., & Park, M. V. de V. (1993). Individual and Corporate Dispute Resolution: Using Procedural Fairness as a Decision Heuristic. *Administrative Science Quarterly*, *38*(2), 224–251. https://doi.org/10.2307/2393412

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–130. https://doi.org/10.1145/3351095.3372867

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2022). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 1–35. https://doi.org/10.1145/3457607

Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.-E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., Kompatsiaris, I., Kinder-Kurlanda, K., Wagner, C., Karimi, F., Fernandez, M., Alani, H., Berendt, B., Kruegel, T., Heinze, C., … Staab, S. (2020). ++Bias in data-driven artificial intelligence systems—An introductory survey. *WIREs Data Mining and Knowledge Discovery*, *10*(3), e1356. https://doi.org/10.1002/widm.1356

Electronic copy available at: https://ssrn.com/abstract=4437243

O'Neil, C. (2019). Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy. *Vikalpa: The Journal for Decision Makers*, *44*(2), 97–98. https://doi.org/10.1177/0256090919853933

Rawls, J. (1999). *A theory of justice* (Rev. ed). Belknap Press of Harvard University Press.

Robert, L. P., Pierce, C., Marquis, L., Kim, S., & Alahmad, R. (2020). ## Designing fair AI for managing employees in organizations: A review, critique, and design agenda. *Human–Computer Interaction*, *35*(5–6), 545–575. https://doi.org/10.1080/07370024.2020.1735391

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., & Liu, Y. (2019). How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 99–106. https://doi.org/10.1145/3306618.3314248

Shin. (2021). *The Effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI | Elsevier Enhanced Reader*. https://reader.elsevier.com/reader/sd/pii/S1071581920301531?token=1D3980A189CF24BF0E6FCCBA12140CC013375DC7E3D330F8B5824EB11491D8F894CA7BE229DD3FEBBFE0E497861769E0&originRegion=us-east-1&originCreation=20230218174331

Srivastava, M., Heidari, H., Krause, A., & Assoc Comp Machinery. (2019). Mathematical Notions vs. Human Perception of Fairness: A Descriptive Approach to Fairness for Machine Learning. *Stanford University*, 2459–2468. https://doi.org/10.1145/3292500.3330664

Starke et al. (2022, December). *Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature*. https://journals.sagepub.com/doi/epub/10.1177/20539517221115189

Thibaut, J., Walker, L., LaTour, S., & Houldent, P. (1973). Procedural Justice as Fairness. *STANFORD LAW REVIEW*.

Thompson, L., & Loewenstein, G. (1992). Egocentric interpretations of fairness and interpersonal conflict. *Organizational Behavior and Human Decision Processes*, *51*(2), 176–197.

Electronic copy available at: https://ssrn.com/abstract=4437243

van Berkel, N., Goncalves, J., Russo, D., Hosio, S., & Skov, M. B. (2021). Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3411764.3445365

Wang, C., Wang, K., Bian, A., Islam, R., Keya, K. N., Foulds, J., & Pan, S. (2022). Do Humans Prefer Debiased AI Algorithms? A Case Study in Career Recommendation. *27th International Conference on Intelligent User Interfaces*, 134–147. https://doi.org/10.1145/3490099.3511108

Wang, G., Guo, Y., Zhang, W., Xie, S., & Chen, Q. (2023). What type of algorithm is perceived as fairer and more acceptable? A comparative analysis of rule-driven versus data-driven algorithmic decision-making in public affairs. *Government Information Quarterly*, 101803. https://doi.org/10.1016/j.giq.2023.101803

Wang, R., Harper, F., Zhu, H., & ACM. (2020). Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. *Carnegie Mellon University*. PROCEEDINGS OF THE 2020 CHI CONFERENCE ON HUMAN FACTORS IN COMPUTING SYSTEMS (CHI'20). https://doi.org/10.1145/313831.3376813

Wang, T., Zhao, J., Yatskar, M., Chang, K.-W., & Ordonez, V. (2019). Balanced Datasets Are Not Enough: Estimating and Mitigating Gender Bias in Deep Image Representations. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 5309–5318. https://doi.org/10.1109/ICCV.2019.00541

Whittaker, M., Crawford, K., Dobbe, R., Fried, G., Kaziunas, E., Mathur, V., West, S. M., Richardson, R., Schultz, J., & Schwartz, O. (2018). *AI now report 2018*. AI Now Institute at New York University New York.

Woodruff, A., Fox, S. E., Rousso-Schindler, S., & Warshaw, J. (2018). A Qualitative Exploration of Perceptions of Algorithmic Fairness. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3174230

Electronic copy available at: https://ssrn.com/abstract=4437243

Electronic copy available at: https://ssrn.com/abstract=4437243