

Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature

Big Data & Society
July–December: 1–16
© The Author(s) 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/20539517221115189
journals.sagepub.com/home/bds
 SAGE

Christopher Starke¹ , Janine Baleis², Birte Keller²
and Frank Marcinkowski²

Abstract

Algorithmic decision-making increasingly shapes people's daily lives. Given that such autonomous systems can cause severe harm to individuals and social groups, fairness concerns have arisen. A human-centric approach demanded by scholars and policymakers requires considering people's fairness perceptions when designing and implementing algorithmic decision-making. We provide a comprehensive, systematic literature review synthesizing the existing empirical insights on perceptions of algorithmic fairness from 58 empirical studies spanning multiple domains and scientific disciplines. Through thorough coding, we systemize the current empirical literature along four dimensions: (1) algorithmic predictors, (2) human predictors, (3) comparative effects (human decision-making vs. algorithmic decision-making), and (4) consequences of algorithmic decision-making. While we identify much heterogeneity around the theoretical concepts and empirical measurements of algorithmic fairness, the insights come almost exclusively from Western-democratic contexts. By advocating for more interdisciplinary research adopting a society-in-the-loop framework, we hope our work will contribute to fairer and more responsible algorithmic decision-making.

Keywords

Algorithmic decision-making, fairness perceptions, algorithmic fairness, artificial intelligence ethics, human-centric approach, systematic literature review

Introduction

Algorithms increasingly shape people's daily lives by making important decisions, for example, in public administration (AlgorithmWatch, 2019), the legal system (Chouldechova, 2017), and hiring (Acikgoz et al., 2020). Algorithmic decision-making (ADM) can lead to faster and better decision outcomes (Lepri et al., 2018). For instance, in South Korea, algorithms were used to relocate ambulance units so that more people could receive help within 5 min of making an emergency call (Nam, 2020). Furthermore, ADM improved social integration outcomes by successfully assigning refugees to resettlement locations (Bansak et al., 2018). However, ADM often includes a downside: unfair ADM systems can systematically reinforce racial or gender stereotypes, marginalize minorities, or flat-out denigrate certain members of society (Veale and Binns, 2017; Žliobaitė, 2017). The famous example of the COMPAS algorithm, which disproportionately assigned a higher risk score of recidivism to black than to white defendants, is evidence of existing algorithmic

discrimination (Chouldechova, 2017). The reasons for unfair ADM include biased input data, faulty algorithms, poor implementation, or transferring decision authority for sensitive issues from humans to algorithms in the first place.

Fairness has become a key element in developing algorithmic systems to counter such detrimental results (Hutchinson and Mitchell, 2019). Algorithmic fairness is endorsed as one of the four main principles for trustworthy Artificial Intelligence (AI) by the OECD (2019) and the European Commission (2019), and it has been featured in more than 80% of guidelines for AI ethics (Jobin et al.,

¹Amsterdam School of Communication Research, University of Amsterdam, Amsterdam, the Netherlands

²Department of Social Sciences, Heinrich Heine University Düsseldorf, Düsseldorf, Germany

Corresponding author:

Christopher Starke, University of Amsterdam, Amsterdam School of Communication Research, Nieuwe Achtergracht 166, 1018 WV Amsterdam, the Netherlands.

Email: christopher.starke@uva.nl



2019). However, addressing the societal implications of (un)fair ADM requires more than mere technological solutions (Barabas et al., 2020; Sloane and Moss, 2019): designing and implementing fair algorithms calls for a thorough empirical understanding of when and why citizens *perceive* ADM to be (un)fair. Insights into citizens' fairness perceptions are essential to facilitate human-centric AI by informing developers entrusted with designing ethical ADM and decision-makers tasked with implementing such systems in social contexts (Kieslich et al., 2022). Thereby, fairness perceptions can contribute to answering the call for a society-in-the-loop approach that emphasizes embedding societal values in the design of ADM systems (Rahwan, 2018). Here, we present the first authoritative systematic literature review mapping the existing empirical insights on *fairness perceptions of algorithmic decision-making*.

This paper synthesizes the interdisciplinary results of 58 empirical studies, incorporating over 33,000 unique observations of citizens' fairness perceptions of ADM. We apply a state-of-the-art approach to capturing academic and gray literature by combining a systematic online search with sequential citation searches. Screening more than 3000 entries, two coders identified the relevant literature in two subsequent coding steps. We systemize the existing literature along four main dimensions of perceived algorithmic fairness: (1) algorithmic predictors, (2) human predictors, (3) comparative effects (human decision-making (HDM) vs. ADM), and (4) consequences of ADM.

Bias in ADM

Due to unprecedented data availability and computing power, institutions of the private and public sectors increasingly implement algorithms to make important decisions such as screening suitable applicants, allocating treatment to patients, or predicting crime (AlgorithmWatch, 2019). Algorithms can reduce human biases in decision-making processes because they do not grow tired, have no agency, and are not distorted by emotional factors (Lee, 2018). However, ADM can also decrease fairness (Barocas and Selbst, 2016). For instance, ADM systems have arbitrarily excluded citizens from food support programs, mistakenly reduced their disability benefits, or falsely accused them of fraud (Richardson et al., 2019). Biases in ADM are often unintended and can have different reasons. They can occur in collecting and processing input data but also in selecting and specifying the algorithm (Veale and Binns, 2017). First, in terms of training data, ADM systems that learn from historical input data are likely to reproduce or even exacerbate existing societal biases, often with harmful outcomes for minority groups (Eubanks, 2018; Lepri et al., 2018). Data about individual or group features may be incomplete or unreliable, leading to a misrepresentation of certain groups (Köchling and Wehner, 2020). Second, algorithms may discriminate

if they are carelessly selected, designed, and specified, as some ADM systems may perform fairly on some specific tasks but unfairly on others (Veale and Binns, 2017).

However, reducing such bias is not merely a technical challenge (Lepri et al., 2018; Wong, 2020); problematic impacts can also occur in the implementation of an ADM system (Köbis et al., 2021)—for example, when the system violates privacy rights or is used for sensitive decisions that should not be made by an algorithm in the first place. On a more general note, “one can dispute whether an algorithm is fair by questioning the idea of fairness underlying the ‘fair’ algorithm in question” (Wong, 2020: 227). Thus, the following section focuses on the different existing notions of algorithmic fairness.

Concepts of algorithmic fairness

The concept of fairness has regained prominence as a core objective in designing AI (Jobin et al., 2019). The term “algorithmic fairness” generally means that decisions made by an algorithm should not produce unjust, discriminatory, or disparate consequences (Shin and Park, 2019). Two broad approaches to algorithmic fairness can be identified: first, literature that formalizes algorithmic fairness and derives mathematical definitions (Gajane and Pechenizkiy, 2017; Verma and Rubin, 2018; Žliobaitė, 2017); second, literature that draws on philosophical and social-science concepts of human fairness and applies them to algorithms (Binns, 2018a, 2018b; Marcinkowski and Starke, 2019).

By reviewing more than 20 different formal definitions, Verma and Rubin (2018) found that formal concepts can largely be clustered into three categories (for an elaborate historical discussion of fairness definitions, see Hutchinson and Mitchell, 2019). First, *statistical measures* (e.g. statistical parity, Dwork et al., 2012) are based on different calibrations of predicted probabilities, predicted outcomes, and actual outcomes. Second, *similarity-based measures* (e.g. fairness through awareness, Dwork et al., 2012) assume that similar individuals should be treated similarly, regardless of their classification in various specific groups. Third, *causal reasoning* (e.g. counterfactual fairness, Kusner et al., 2017) argues that structural equations can be used to estimate the effects of sensitive attributes and then design algorithms that ensure tolerable discrimination levels due to these attributes. Adding to the list of formal definitions, Zafar et al. (2017) introduce preference-based fairness (e.g. preferred treatment, preferred impact), conceived as a predictor that increases benefits for a group compared to another predictor.

The plethora of existing formal fairness definitions indicates that they refer to different notions of fairness, but these conceptions are also often incompatible (Kleinberg et al., 2017). Thus, as different fairness trade-offs emerge, several authors have highlighted the importance of the social context when assessing appropriate understandings of algorithmic fairness (Lepri et al., 2018; Wong, 2020).

In two seminal papers, Binns (2018a, 2018b) drew on moral and political philosophy to outline a concept of algorithmic fairness. He discussed how egalitarianism—the belief in equal treatment of people and the equal distribution of fundamental rights and goods—can inform theoretical notions of algorithmic fairness. For instance, algorithmic (un)fairness cannot only be assessed on the grounds of unequal distribution; instead, it should also consider how inequality is produced (Binns, 2018a).

Despite using a different theoretical approach, other authors have come to similar conclusions (Grgić-Hlača et al., 2018b; Marcinkowski and Starke, 2019). Drawing on organizational justice literature (Greenberg, 1990), other studies have assessed algorithmic fairness according to four dimensions. First, *distributive fairness* refers to the non-discriminatory allocation of resources based on equality, equity, or need (Deutsch, 1975). Second, *procedural fairness* indicates that decision-making is based on fair criteria, such as revocability or consistency (Leventhal, 1980). Third, *informational fairness* involves the transparency of ADM systems (Greenberg, 1993). Fourth, *interpersonal fairness* is achieved if an ADM refrains from using protected data and respects privacy rights (Greenberg, 1993).

Fair predictions by algorithms cannot be made without considering social questions. ADM systems do not operate in a vacuum but instead, need to be calibrated to the specific social context. As Shin and Park (2019: 279) stated: “What establishes an algorithm system as a socio-technical system is that it is generated by or related to a system adopted and used by social users in societies.” As humans are ultimately affected by the decisions made by ADM, several authors advocate for a more human-centric approach to researching algorithmic fairness to ensure that ADM systems are legitimized (Grgić-Hlača et al., 2018a; Kieslich et al., 2022). Many empirical studies have addressed this call by empirically investigating human perceptions of algorithmic fairness. It is that growing literature that we review in this paper.

Method

We systematically reviewed the empirical literature on people’s perceptions of algorithmic fairness using the approach recommended by Petticrew and Roberts (2006). It outlines seven steps to ensure a thorough literature review based on predefined and transparent criteria: (1) *research question or hypothesis*, (2) *inclusion criteria*, (3) *comprehensive literature search*, (4) *screening of the results*, (5) *critical evaluation*, (6) *summary*, and (7) *dissemination*.

Establishing the research question

A precise research question is vital for a systematic literature review (Booth et al., 2016). We applied the “PICOC” method (Petticrew and Roberts, 2006) to

break down the research question into five components: **P**opulation, **I**ntervention, **C**omparison, **O**utcome, and **C**ontext. This approach helps identify possible search terms for subsequent searches in databases (Booth et al., 2016). First, the *population* includes all individuals, irrespective of sociodemographic characteristics. This includes the general public, but also AI experts and decision-makers such as public officials or managers. Second, the *intervention* is defined as fairness in and through algorithms. Third, we did not specify a *comparison*, as it was not considered beneficial to include additional interventions. Fourth, the *outcome* involves individual fairness perceptions about ADM. Fifth, the *context*—conceived here as a country-specific and domain-specific setting—was not narrowed down, thereby making the literature review global in scope. Using these criteria, we derived the following research question: *How do individuals perceive the fairness of algorithmic decision-making?*

Inclusion criteria

The keywords in the research question (“*fairness*,” “*algorithmic*”) provide the basis for the search terms. We adopted the “pearl-growing” method for additional terms for the subsequent search (Booth et al., 2016). It draws on relevant articles (“pearls”) to identify further relevant search terms or keywords. For this study, we initially identified three articles as “pearls” because they are widely cited in the literature: Binns et al. (2018), Grgić-Hlača et al. (2018a), and Lee (2018). As all these articles correspond with our research interest, we added relevant keywords to the search terms we derived from our research question: “*justice*,” “*discrimination*,” “*machine learning*.” Subsequently, we clustered the search terms into two components: terms referring to ADM, and terms referring to the theoretical concept of fairness.

We inputted these search terms into the following Boolean operators (Lefebvre et al., 2008): (“big data” OR “artificial intelligence” OR “machine learning” OR “algorithm*”) AND (“fair*” OR “unfair*” OR “just*” OR “discrimina*” OR “bias” OR “disparate”).

If the predefined combination of search strings appeared in the title of a publication, we included that publication in the preliminary sample, which we then used for the first screening process. Due to the recent surge of literature on algorithmic fairness, we adopted the reasoning suggested by Favaretto et al. (2019) and only included studies written in English and published since January 2010.

Comprehensive literature search

We selected the final sample of empirical studies through a stepwise process (see PRISMA chart, Figure 1). As fairness of ADM has been investigated in different research disciplines,

the selection of the databases was based on thematic classification. We applied the Boolean logic to the following electronic databases: Web of Science, PsycINFO, IEEE Xplore, and Scopus. Our study also includes “gray literature”—working papers, pre-prints, or reports—as they account for recent research efforts. We used Google Scholar to manually search for relevant publications by searching for publications by key institutions (e.g. AI NOW, AlgorithmWatch).

Applying the search terms to the selected databases, we identified a total of 5117 contributions (24 February 2022), of which 3097 remained after filtering for duplicates. In a subsequent research step, we identified 101 potentially relevant articles by manually searching gray literature. We extracted the title, authors, keywords, journal information, and abstract for all publications.

Screening & critical evaluation

Two expert raters examined the 3198 publications based on their titles and abstracts for the initial screening. The screening was twofold. In the first step, for a publication to be considered for further analysis, both raters had to agree that it generally addressed the research interest. In this step, 389 publications were selected. In the second step, all publications were rated in terms of their applicability to the research question of the review and the use of an empirical method. All publications that were rated 1 (*applies*) on both criteria were selected for the literature review. This filtering process resulted in a sample consisting of 46 publications. Finally, we used Google Scholar to scan all publications that cited at least one of the “pearls” (Binns et al., 2018; Grgić-Hlača et al., 2018a; Lee, 2018), identifying twelve other relevant publications that we included in the final sample.

Two raters assessed the reliability of the selection process with a sample of 72 abstracts in the first test (Cohen’s $\kappa = .55$, “fair agreement”) and 39 abstracts in the second test (Cohen’s $\kappa = .74$, “good agreement”) (Higgins and Deeks, 2008).

Results

To start, we outline the descriptive results of the 58 studies included in the literature review. Then, we shed light on the underlying theoretical concepts of fairness and the specific measurements used in the empirical studies. We then reveal the main insights by clustering the empirical results from the existing literature on individuals’ perceptions of algorithmic fairness into four main categories: *algorithmic predictors*, *human predictors*, *comparative effects* (HDM vs. ADM), and *consequences of ADM* (see Figure 2).

Descriptive results

The descriptive results (shown in Table 1 in the Supplemental Material) indicate high homogeneity in

terms of the national context. Data on citizens’ perceptions of algorithms’ fairness was almost exclusively collected in Western democracies: 32 studies were conducted in the United States (US), three in China, three in Germany, two in the United Kingdom (UK), two in the Netherlands, one in South Korea, and one in Cyprus, with four studies collecting data in multiple countries.¹

Looking more closely at the empirical methods used to investigate citizens’ perceptions of algorithmic fairness, we find great diversity. Eight studies used a qualitative design, 37 studies used quantitative methods, and 13 studies combined qualitative and quantitative approaches in mixed-method designs.

With regard to the different domains within which the studies were located, the descriptive results reveal a focus on work-related decisions (16 studies), especially in hiring. Moreover, emphasis is also given to the criminal justice system as an area of application (13 studies)—most prominently pretrial risk assessment. Other domains include news recommendations, allocation of donations, university admissions, loan decisions, and targeted advertisements.

Concepts of (algorithmic) fairness

Despite the growing interest in algorithmic fairness “there is no consensus on a precise definition of (un)fairness” (Srivastava et al., 2019: 1). We find much heterogeneity in relation to algorithmic fairness notions (see Table 1 in the Supplemental Material). Twenty-three studies focus on the perceived fairness of decision outcomes—that is, distributive fairness, although many computer science studies do not use this term explicitly. Thirteen of them distinguished between different formal definitions of algorithmic fairness (e.g. demographic parity, equalized odds) (Saxena et al., 2020; Srivastava et al., 2019) or among equality, equity, and need (Schlicker et al., 2021). Thus, they conceived of algorithmic fairness in distribution norms. In this reasoning, fairness can be achieved by decreasing “discriminatory consequences for certain groups of individuals” (Dodge et al., 2019: 275).

Twenty-two studies went beyond distributive fairness and investigated the fairness of algorithmic processes. However, conceptual ambiguity exists. While some studies defined procedural fairness more narrowly as the inclusion of selected (sensitive) input features, others conceptualized it in broader terms, also addressing such criteria as the consistency or the revocability of decisions (Hsu et al., 2021; Schlicker et al., 2021). Thus, procedural fairness can be increased even if the decision outcome is biased by using appropriate input features and ensuring people can appeal an ADM decision. We further find that five studies also included interactional fairness, and six studies investigated informational fairness (Acikgoz et al., 2020; Binns et al., 2018; Schoeffer et al., 2021). These

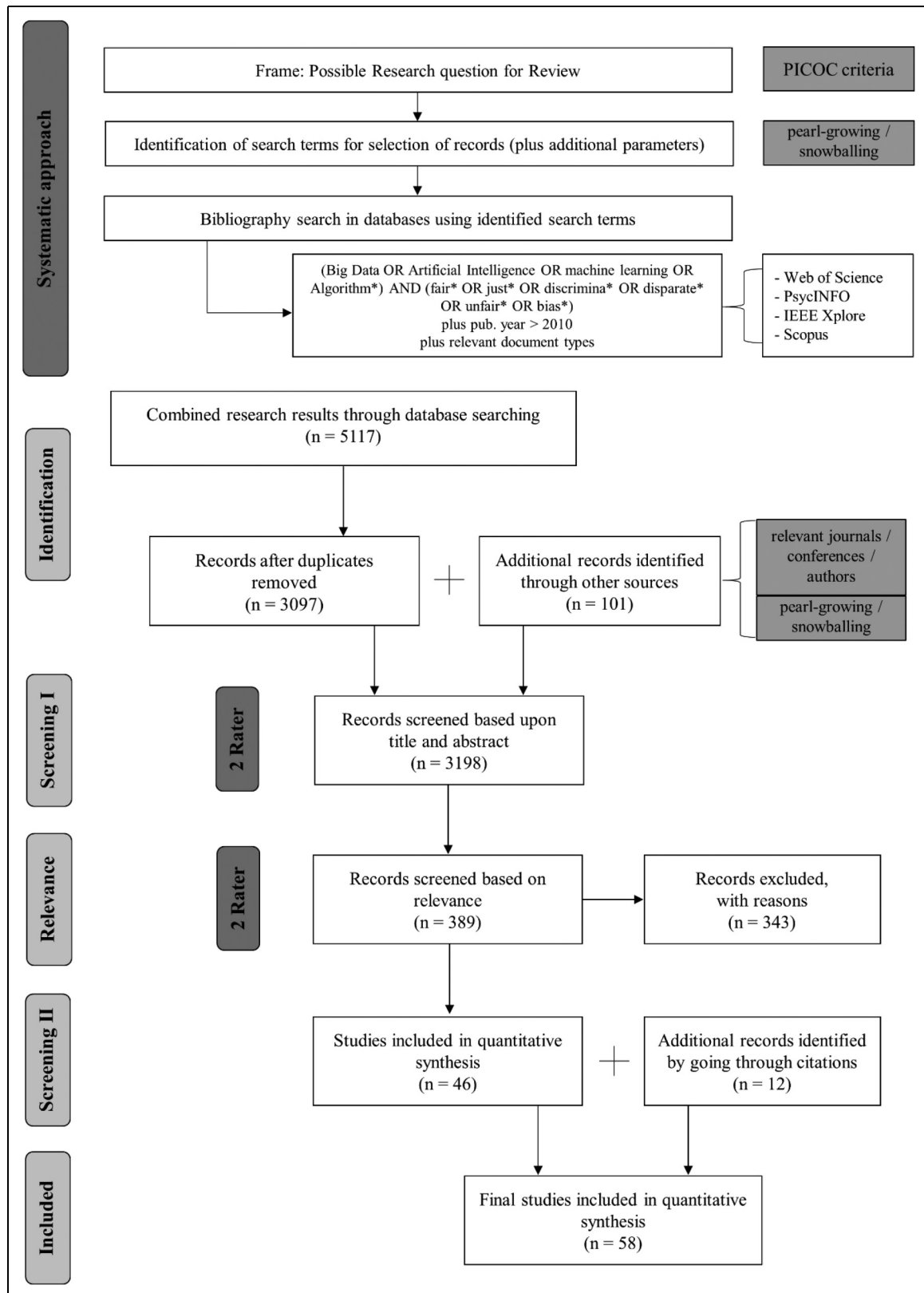


Figure 1. Flowchart documenting the selection process adopted by Moher et al. (2009).

studies zoom into the decision process and emphasize its social aspects, such as treating individuals with respect and receiving an explanation for an ADM decision (Schlicker et al., 2021). In this reading, fairness can be fostered by endowing ADM systems with fundamental aspects of social interactions indicating that algorithms should adhere to the same social norms as humans.

Measurements of algorithmic fairness

Most quantitative studies used a two-step process to measure fairness perceptions of ADM: respondents were confronted with an ADM process or outcome, and then they were asked about their perceived fairness. However, the results show much diversity in the measurement of algorithmic fairness. Fourteen studies simply used single items measured on 5-point or 7-point Likert scales, along the lines of “how fair did you perceive the decision?”. A total of 17 studies drew from existing fairness scales (e.g. Colquitt and Rodell, 2015) and adapted them for ADM. Five studies indirectly gauged fairness via stated preferences for one ADM over another. Ahnert et al. (2021) summarized the different measurement steps in their FairCeptron framework.

However, other measurements also exist. Pierson (2017) predefined a fair distribution of resources and asked participants’ approval on a 7-point Likert scale. Other studies used a conjoint design to investigate how algorithmic fairness hinges on *what*, *who*, and *how* resources are allocated (Hannan et al., 2021) and the relative importance of fairness in relation to other attributes such as transparency or accountability (Kieslich et al., 2022).

Algorithmic predictors of perceived algorithmic fairness

A significant strand of literature investigates how an ADM system’s technical design affects people’s perceptions of fairness. Overall, fairness is a crucial factor when evaluating algorithms (Bankins et al., 2022). Studies that tapped into people’s general notions of algorithmic fairness (Dodge et al., 2019; Shin and Park, 2019) yielded mixed results: some respondents perceived “the *very idea* of an algorithmic system making an important decision on the basis of past data [...] unfair” (Binns et al., 2018: 9), while other participants argued that algorithms are by definition impartial (Lee and Rich, 2021).

On a general note, Zhou et al. (2021) showed that a decrease in the de facto fairness of ADM led to a decline in respondents’ perceived fairness. However, fairness perceptions are highly context-dependent. Some studies revealed that algorithmic fairness is perceived as more problematic in some domains than in others (Hannan et al., 2021). For instance, discrimination by ADM in

housing or job recommendations is viewed as more harmful than in music or movie recommendations (e.g. Spotify, Netflix) (Smith et al., 2020). Also, less complex algorithmic tasks elicited higher fairness perceptions than more complex ones (Hsu et al., 2021).

A large group of studies goes beyond people’s basic understanding of algorithmic fairness and investigates how people’s perceptions of fairness are related to different outcome distributions. This is particularly intriguing as some formal fairness definitions cannot coexist (Kleinberg et al., 2017). Lee and colleagues (Lee et al., 2017; Lee et al., 2019b) tested people’s fairness perceptions regarding the allocation of resources based on equality, equity, or efficiency. They found much variation in the preferences for the three fairness concepts, both within and across different social groups differently impacted by the decision. Most respondents considered an outcome fair when it mirrored their input (equity). Yet, some respondents also believed that an equal allocation of tasks or resources was fair, emphasizing moral norms such as self-sacrifice (equality) (Lee and Baykal, 2017).

While these qualitative studies focused on basic fairness concepts, other quantitative studies tested more nuanced notions of fairness. Srivastava et al. (2019) used criminal risk and skin cancer as examples and matched people’s fairness choices with six different notions of group fairness. Their results showed that demographic parity best matched the fairness choices most respondents made in both scenarios. Thus, people favored algorithms aiming to equalize the positive rate across different groups. For instance, if ten percent of all applicants to a university get admitted, this rate should be equal for all gender groups. While the authors further found that in high-stakes situations, respondents weighed accuracy higher and inequality lower (Srivastava et al., 2019), a qualitative study by Koene et al. (2017) revealed that participants deemed ethical considerations more important than higher accuracy. Three other studies shed light on the relationship between fairness and accuracy: Kieslich et al. (2022) showed that fairness and accuracy were similarly important to respondents, Cheng et al. (2021) found that participants rather accepted disparities in accuracy across groups than give up overall accuracy, and Hsu et al. (2021) highlighted that accurate ADM is also perceived as fairer than inaccurate ADM.

Adding to this, two studies (Kasinidou et al., 2021a; Saxena et al., 2020) investigated a different set of formal fairness definitions in the context of loan decisions: equal distribution (“money is split equally among candidates”), meritocratic distribution (“all the money is distributed to the candidate with the highest payback rate”), and calibrated/proportional distribution (“money is split proportionally to candidates payback rates”). The results indicated that people perceived the calibrated model to be the fairest. Then again, Cheng et al. (2021) compared

three group fairness approaches in a child maltreatment predictive system and found that respondents most supported *equalized odds*, followed by *statistical parity* and *unawareness*.

Another study shed light on the trade-offs between different incompatible fairness definitions in the criminal justice context. The results indicated that respondents favored an algorithm that equalizes the false positive rate between groups over one that equalizes accuracy (Harrison et al., 2020).

However, the technical design of an algorithm refers not only to the decision outcome but also to the decision process. Six studies investigated the perceived fairness of input features. Grgić-Hlača et al. (2018b) used predictive policing as a case study. They showed that respondents perceived feature-accuracy fairness (“a feature is perceived as fair if it increases the accuracy of an algorithm”) to be the fairest process, followed by feature-a priori fairness (“a feature is perceived as fair, independent of its effect on the outcome”) and then feature-disparity fairness (“a feature is perceived as fair even if it increases disparity in the outcomes of an algorithm”).

Two other studies tested eight feature properties that determine if people perceive the use of said feature in an ADM system to be fair: while Grgić-Hlača et al. (2018a) indicated that *relevance*, *causes outcome*, and *reliability* are most important for respondents, Albach and Wright (2021) found that *relevance* and *increases accuracy* are the essential features when deciding whether it is fair to use a feature in an ADM system.

Other studies shed further light on procedural fairness showing that respondents perceived features that directly relate to the issue at hand to be fairest and perceived unrelated features to be the most unfair (Grgić-Hlača et al., 2018b; Plane et al., 2017; van Berkel et al., 2019). Nyarko et al. (2021) qualified this finding by discovering that respondents were generally opposed to including sensitive features, like race and gender, in an ADM system. However, after respondents were told that including these features in the model can lead to better outcomes for minority groups, support for such “non-blind algorithms” increased substantially. In one of the few studies conducted in a non-Western context, Sambasivan et al. (2021) further found that missing data and misrepresentation of subgroups in the existing data were critical reasons for algorithmic unfairness perceived. Moreover, respondents’ fairness perceptions hinged upon different design factors of the algorithmic process, such as the possibility to appeal a decision made by an algorithmic system (Hsu et al., 2021) and people’s control to avoid algorithmic discrimination (Sun and Tang, 2021).

Another set of studies looked at explanations for a decision as a critical aspect of perceived fairness. Explanations for an algorithmic decision significantly increased

respondents’ perceptions of fairness in several studies (Binns et al., 2018; Dodge et al., 2019; Shin, 2021; Shulner-Tal et al., 2022). However, the results are very nuanced. Schlicker et al. (2021) found that while explanations for ADM had no impact on interpersonal and distributive fairness, they influenced perceptions of informational fairness. Lee et al. (2019a) added that the direction of the effects largely depended on the context. When explanations helped respondents understand biased distributions, perceived fairness decreased, yet, when explanations helped respondents understand equality in utility distribution, perceived fairness increased. Moreover, several studies showed that different explanation styles (e.g. *case-based*, *sensitivity-based*, *demographic-based*, *input influence-based*) affected respondents’ perceived fairness differently (Binns et al., 2018; Dodge et al., 2019; Schoeffer et al., 2021; Shulner-Tal et al., 2022).

Looking at transparency more generally, Perez Vallejos et al. (2017) found that young people demanded more information about an algorithm to perceive it as fair. However, empirical, experimental evidence offers inconclusive results. While Wang (2018) found that algorithmic transparency increased perceptions of fairness, Wang et al. (2020) observed that different degrees of transparency had no significant effect on algorithmic fairness. Van Berkel et al. (2021) focused on visualizations as a specific form of transparency and showed that scatterplots led to lower perceived fairness levels than text.

Human predictors of perceived algorithmic fairness

Another strand of empirical studies investigates human predictors of perceived algorithmic fairness. Only a few studies found an impact of sociodemographic variables. For instance, three studies suggested a significant influence of gender: female respondents opposed gender as an input feature more strongly than male respondents (Grgić-Hlača et al., 2020; Pierson, 2017) and overall showed a lower level of perceived algorithmic fairness (van Berkel et al., 2021). Helberger et al. (2020) found that education and age affected both perceptions of algorithmic fairness. In terms of age, young people perceived the inclusion of a sensitive feature in ADM systems as unfair (Grgić-Hlača et al., 2020) and demanded global approaches to regulating fairness for algorithms (Perez Vallejos et al., 2017). Moreover, higher educated respondents showed higher levels of perceived algorithmic fairness (van Berkel et al., 2021) and deemed fairness an important feature when evaluating algorithms (Kieslich et al., 2022). Other studies found that algorithmic fairness perceptions can vary between ethnic groups (Albach and Wright, 2021; Lee and Rich, 2021).

Five studies found that perceived algorithmic fairness hinged on self-interest, indicating that people tend to perceive algorithms as fairer when the ADM yields a positive

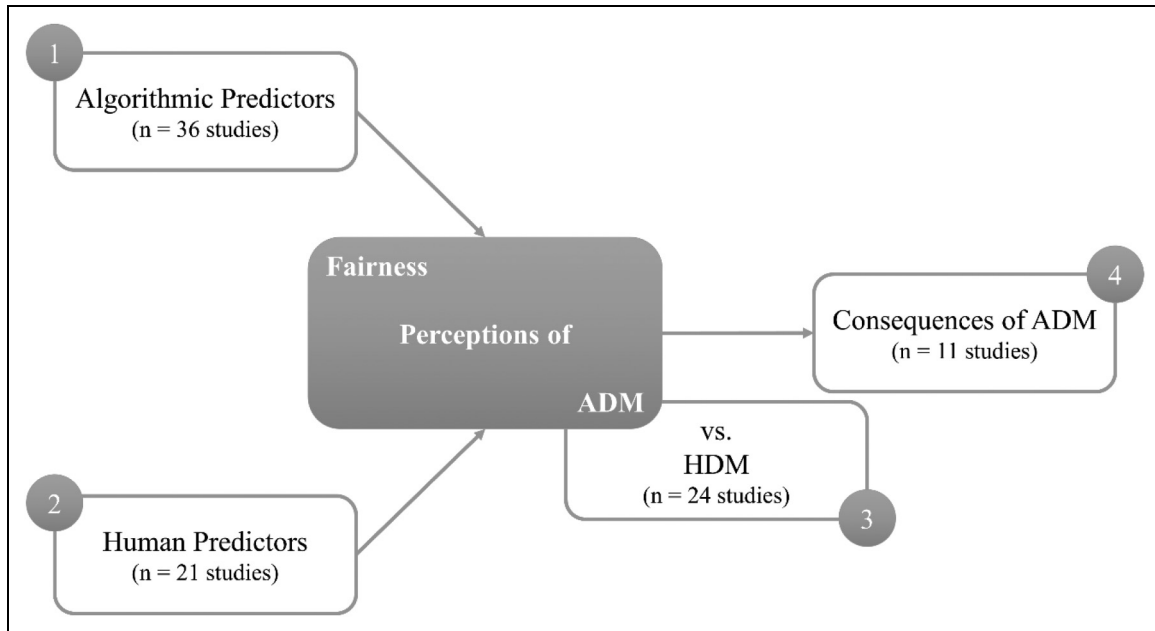


Figure 2. Summary of the results in a process model.

outcome for them (Bankins et al., 2022; Grgić-Hlača et al., 2020; Noble et al., 2021; Shulner-Tal et al., 2022; Wang et al., 2020). Along similar lines, respondents perceived decision outcomes as fairer when their social group benefited from them, suggesting in-group favoritism (Hannan et al., 2021). In addition, Lee et al. (2019a) presented evidence that the more significant the gap between prediction and the actual outcome, the more unfair an algorithm was perceived.

Another aspect that received considerable attention is people's familiarity with data and algorithms. While participating in a workshop about ethical AI raised awareness about algorithmic fairness (Kasinidou et al., 2021b), understanding the mathematical definition of the fairness concept led respondents to reject the fairness concept (Saha et al., 2020). While some other studies indicated that higher levels of AI literacy were associated with lower levels of perceived algorithmic fairness (Lee and Baykal, 2017; Schoeffer et al., 2021), Araujo et al. (2020) found that the effects ran in the opposite direction.

Two other studies examined political ideology as a predictor of algorithmic fairness perceptions. Grgić-Hlača and colleagues (Grgić-Hlača et al., 2020; Grgić-Hlača et al., 2018a) found that conservative users perceived the inclusion of sensitive features such as gender and race in an ADM system as fairer than liberal users did.

Comparative effects (HDM vs. ADM)

Our literature review found 24 studies that examined whether decisions made by humans or algorithms are

perceived to be fairer (see Figure 2), with ambiguous results. After being asked, "Who would, according to you, make a fairer decision: a human or artificial intelligence/computer?," 54% of respondents answered that they believed AI makes fairer decisions (compared to 33% for humans) (Helberger et al., 2020). Schoeffer et al. (2021) investigated open-ended questions and also found that respondents perceived ADM as less biased than HDM. Other studies support this general finding in university admission decisions (Marcinkowski et al., 2020) and algorithmic work assignments (Bai et al., 2020).

However, several studies provided different evidence that suggests that humans are viewed as fairer compared to algorithms, in criminal justice (Harrison et al., 2020; Wang, 2018), healthcare (Lee and Rich, 2021), work-related decisions (Acikgoz et al., 2020; Bankins et al., 2022; Newman et al., 2020; Noble et al., 2021), sport referee decisions (Wonseok et al., 2021), and social division tasks (Lee and Baykal, 2017). The main reason for this finding is that respondents believed that algorithms do not consider qualitative information or context. However, Wang (2018) found that respondents were more willing to accept discriminatory outcomes when the outcomes were attributed to algorithms rather than humans. Other studies found evidence for both positions, as HDM and ADM had different effects on different proxies of procedural fairness: while algorithms were rated higher in terms of consistency, human agents were rated higher in terms of personableness (Kaibel et al., 2019; Schlicker et al., 2021). Then again, other studies found no significant differences in perceived fairness between HDM and ADM (Plane et al., 2017; Suen et al., 2019).

The inconsistency of the empirical evidence suggests that fairness perceptions of HDM versus ADM are highly context-dependent. Consequently, several studies looked at conditional effects or distinguished between different kinds of decisions. An experiment on ADM versus HDM in policing suggested that procedural fairness perceptions hinge on social group representation (Miller and Keiser, 2021). Black respondents rated ADM as fairer—but only when they felt their social group lacked representation in the HDM condition.

Lee (2018) compared different decisions and found that ADM and HDM were perceived as equally fair for tasks requiring mechanical skills. However, respondents perceived HDM as fairer than ADM for tasks requiring human skills. Araujo et al. (2020) investigated high- and low-impact decisions. Overall, the findings suggested no significant differences between ADM and HDM in perceived fairness; however, ADM was perceived as fairer than HDM in high-impact decisions in the health and justice sectors. In contrast to this finding, Nagtegaal (2021) distinguished between high- and low-complexity decisions and found that HDM was perceived as fairer than ADM in high-complexity tasks. However, for low-complexity tasks, respondents viewed algorithms as fairer.

Another strand of literature went beyond the binary distinction between HDM versus ADM and included hybrid decision-making forms. For instance, Nagtegaal (2021) found that decision-making systems involving algorithms and humans are perceived as the fairest in high-complexity situations. However, Newman et al. (2020) suggested that while algorithmic decisions with human oversight increased fairness perceptions, it was still outranked by pure HDM. Lee et al. (2019a) showed that allowing respondents to adjust the algorithmic allocation of resources and thereby overrule decisions made by the ADM system increased people's perceptions of algorithmic fairness. Other studies varied the degree of human involvement in the decision-making process: While Wang et al. (2020) found no significant differences across the experimental conditions, de Cremer and McGuire (2022) suggested that respondents even incurred financial costs to avoid the algorithm led to the decision because they perceived it to be unfair.

Consequences of perceived fairness

Only eleven studies investigated the implications of the perceived fairness of algorithms. Shin and Park (2019) found that perceived fairness has a significant positive impact on satisfaction with algorithms. Several studies focused on the relationship between perceived algorithmic fairness and trust. While Kasinidou et al. (2021a) found no correlation between the two variables, other studies (Shin, 2020, 2021; Shin et al., 2020; Sun and Tang, 2021) suggested that fairness perception had a positive effect on trust in an algorithm. Woodruff et al. (2018) asked respondents belonging to traditionally marginalized groups about

possible consequences of perceived algorithmic unfairness. The results indicated that “algorithmic fairness (or lack thereof) could substantially affect their trust in a company or product” (p. 1). The authors concluded that perceptions of fairness play an essential role in adopting algorithms.

Concerning ADM in human resources decisions, empirical evidence suggests that perceptions of procedural and interactional algorithmic unfairness were associated with lower organizational attraction or commitment and lower job pursuit intention (Acikgoz et al., 2020; Newman et al., 2020). Furthermore, low levels of perceived interactional fairness increased the likelihood of pursuing litigation against a company using ADM systems (Acikgoz et al., 2020). In a field experiment, Bai et al. (2020) found higher perceived fairness was associated with higher productivity among workers in a warehouse.

Marcinkowski et al. (2020) tested how fairness perceptions influence students' intentions to protest, students' willingness to exit, and the institution's reputation if ADM was used for university admissions. The results yield three main insights. First, distributive and procedural fairness perceptions negatively influenced students' intention to protest against an ADM system. Second, perceptions of procedural fairness negatively affected students' likelihood of exiting the university. Third, perceptions of distributive fairness had a positive effect on the university's reputation.

We summarize four main insights: First, preferences for different distribution norms are highly context-dependent and can vary substantially across domains, tasks, and algorithmic designs. However, respondents favored more straightforward fairness definitions over more complex ones. The literature further yielded tentative evidence that explanations can increase perceived fairness. Second, while studies on human predictors delivered inconclusive results regarding sociodemographic variables, they also indicated that political ideology and self-interest influence citizens' fairness perceptions of ADM. Third, studies comparing fairness perceptions of HDM versus ADM revealed ambiguous results: fairness perceptions are highly context-sensitive, making generalizations about the perceived fairness of HDM versus ADM infeasible. Fourth, while little empirical research examined the consequences of perceived algorithmic (un)fairness, initial empirical insights suggested adverse effects for institutions using an ADM system if that system is perceived to be unfair, especially to its reputation.

Discussion

Computer science studies have dominated the literature on algorithmic fairness to mitigate bias and discrimination from machine learning models. However, as “social scientists have long argued, the fairness of a data science project

extends far beyond the technical properties of a given model” (Barabas et al., 2020: 174). Thus, there has been a recent push for social science research that takes a human-centric approach and investigates fairness perceptions of those most affected by ADM. Earlier review papers have addressed human trust in AI (Glikson and Woolley, 2020), algorithmic discrimination in human resources (Köchling and Wehner, 2020), big data discrimination (Favaretto et al., 2019), and formal definitions of algorithmic fairness (Verma and Rubin, 2018). This systematic literature review is the first to shed light on perceptions of algorithmic fairness.

Before discussing the main insights this literature review yielded, we need to point out four limitations: First, the selection of search strings and databases only allows us to make assertions about studies published in English. Thus, we cannot claim that this review is exhaustive or that our findings are representative of all empirical studies on algorithmic fairness published in other languages. Second, we only included published research and did not publically call for unpublished studies. To mitigate this limitation, we included pre-prints and other non-peer-reviewed work. Third, we used our search strings for the title and subtitle of a publication. Hence, our approach was not sensitive to studies that only used the relevant keywords in the main text. Fourth, even though this paper reviews the interdisciplinary literature on algorithmic fairness, the authors are social scientists. Thus, we succumb to disciplinary biases and inadvertently read the studies through a social science lens.

Theoretical groundwork

A key takeaway of this review is that the perceived fairness of ADM systems is highly context-dependent. Fairness perceptions are determined not only by the technical design of the algorithm but also by the area of application (e.g. pre-trial risk assessment, hiring) and the specific task at hand (e.g. high-stakes vs. low-stakes). However, we attribute some of the inconclusiveness of the empirical results to the lack of coherent theoretical frameworks for perceived algorithmic fairness.

Three aspects stand out. First, to avoid conceptual confusion, researchers should clearly state the fairness dimension (e.g. distributive, procedural) they investigate and discuss their results through the respective lens. Second, existing theoretical fairness approaches are not used consistently in the literature. For instance, while some authors define procedural fairness as input features (e.g. Grgić-Hlača et al., 2018b), others conceive of it in broader terms that also include consistency and revocability (Marcinkowski et al., 2020). Third, most studies rely on fairness concepts that have been developed for decisions made by humans (Greenberg, 1990). However, empirical evidence suggests that people base their evaluations of

ADM on different factors than those they use to evaluate HDM (Dietvorst et al., 2015). It is likely that people also include other criteria for assessing the fairness of ADM and HDM. Following these arguments, we echo the call voiced by other authors (Lepri et al., 2018; Wong, 2020) for more theoretical groundwork on human perceptions of algorithmic fairness. We argue that a fruitful avenue for future research lies in developing a coherent, multi-dimensional theoretical concept of perceived algorithmic fairness. For instance, such a framework could encompass the four main fairness dimensions *distributive*, *procedural*, *interactional*, and *informational* fairness, and scrutinize the specific differences between human–human and human–machine interactions.

Diversification versus harmonization of research

We argue that a need exists for diversification and harmonization of empirical research on algorithmic fairness perceptions. First, while the descriptive results show a variety of research methods, they also reveal that the studies included in this review were almost exclusively conducted in Western democracies, predominantly the US. The problematic aspects of generalizing so-called *WEIRD* (White, Educated, Industrialized, Rich, Democratic) samples are widely acknowledged (Henrich et al., 2010). The seminal work of Henrich et al. (2010) showed that fairness in decision-making is considerably dependent on the sociocultural context. The same likely applies to fairness perceptions of algorithms as cross-national variance exists globally in citizens’ perceptions of AI (Kelley et al., 2021). Diversification of countries under investigation—especially those in which ADM systems are already widely implemented, such as China and South Korea—would greatly enrich the existing literature.

In addition, we also call for more diversification in terms of the investigated domains and tasks. Thus far, the literature has been dominated by fairness perceptions around ADM in the criminal justice system and human resources. However, ADM systems have recently surged in many other areas of society, such as distributing social benefits (Noriega-Campero et al., 2020). Thus, more empirical research is needed to compare fairness perceptions across various domains and tasks systematically.

Second, we argue in favor of more harmonization in measurements. Ideally, reliable measures of perceived algorithmic fairness should be developed and validated following the theoretical groundwork outlined above. These would make new findings more comparable and allow for more nuanced interpretations of the results. A multi-dimensional measurement would provide information about the specific deficiencies of an algorithmic process instead of merely indicating that a process is perceived as (un)fair. For example, knowing whether consistency or accuracy concerns drive a perception of unfairness could

enable developers to fine-tune ADM processes according to citizens' desires. Recent studies have tackled this challenge for ADM-adjacent concepts, such as "threats of AI" (Kieslich et al., 2021), introducing validated scales easily adaptable for use in different domains.

Understanding consequences of ADM systems

This review shows that the literature dedicates more attention to investigating the drivers of algorithmic (un)fairness than to its social implications. This emphasis is understandable since insights into the predictors of fairness perception are vital to designing fair ADM systems. However, only eleven studies investigated the consequences of algorithmic fairness perceptions. This is surprising because attitude changes or actions resulting from perceived (un)fairness may have profound societal ramifications. Not only may reputation losses offset potential gains in reducing costs or increasing impartiality. More importantly, understanding the social implications of algorithmic (un)fairness is vital to shedding light on what is at stake. Suppose public and private institutions fail to design and implement ADM systems perceived as fair and deny citizens a voice in this process. In that case, citizens could become alienated or lose trust in public institutions and thereby become more vulnerable to populist rhetoric.

Conceptualizing perceptions of algorithmic fairness as a mediator variable in structural equation models can be a valuable way to study both the drivers and the consequences of algorithmic fairness (Shin, 2021; Shin and Park, 2019). We encourage more research that builds on such approaches and examines the potential implications of perceived algorithmic unfairness for other attitudinal and behavioral variables.

From human-in-the-loop to society-in-the-loop

The results of the perceived fairness of HDM versus ADM are inconclusive; they provide evidence for and against the assumption that algorithms are seen as fairer than HDM. This variation of results can be found across domains (e.g. recidivism risk prediction vs. hiring) and different tasks within the same domain (e.g. two different hiring algorithms). This shows that fairness perceptions are highly context-specific and that every algorithm requires thorough investigation before being widely used. However, simply differentiating HDM and ADM falls short of capturing the natural world's complexity (Binns, 2022). In most real-life tasks, ADM systems do not decide entirely independently; instead, humans are also involved in the decision-making process. For instance, COMPAS makes a prediction, which serves as a recommendation for the final decision by a human judge. Such forms of hybrid decision-making (Starke and Lünich, 2020) can be fine-tuned and deserve more nuanced attention from the empirical literature. For instance, the European

Commission (2019) outlines three approaches to human oversight. First, humans are involved in every step of the decision cycle of an ADM system and can intervene at any point (human-in-the-loop). Second, humans can intervene during the design cycle of an ADM system and monitor the system's operation (human-on-the-loop). Third, humans oversee an ADM system's overall economic, societal, legal, and ethical impacts and have authority over its use in any situation (human-in-command).

This leads to a broader point of institutional implementation. Introducing an ADM system into an institutional context raises critical questions that extend far beyond the algorithmic design (e.g. input data, code) and the specifics of human involvement in the decision-making process. This review revealed a blind spot in the existing literature, as few studies considered the broader institutional context and the emerging questions in implementing an ADM system. For instance, how are humans who might use ADM systems within an institution trained? How are decisions made by an algorithm communicated to those most affected by the decision (e.g. job applicants, defendants)—are they even made aware that an algorithm made the final decision? Do affected citizens have an opportunity to appeal, and who should ultimately be liable for false and/or discriminating classifications?

These are just a few examples of the many emerging questions that are likely to affect citizens' fairness perceptions. Addressing these blind spots requires more interdisciplinary research involving computer scientists, social scientists, legal scholars, and ethicists (Lepri et al., 2018). For instance, computer scientists add technical expertise in how ADM systems work and allow using real algorithms instead of hypothetical scenarios. Social scientists add the methodological expertise necessary to research human perceptions, attitudes, and behaviors. Such research collaborations can help put "society in the loop," which Rahwan (2018) defined as human-in-the-loop plus social contract. This approach argues that the more ADM takes over decisions that profoundly affect citizens' livelihoods, the more essential it becomes to embed social values and norms in designing and implementing such systems.

Conclusion

As ADM increasingly penetrates all sectors of society, concerns about the fairness of such systems have arisen. As scholars and policymakers have demanded a human-centric approach to designing and implementing ADM, the empirical literature on perceived algorithmic fairness is surging. This systematic literature review crystallizes the insights of 58 empirical studies along four dimensions: algorithmic predictors, human predictors, comparative effects (HDM vs. ADM), and consequences of ADM. In conclusion, we call for more research from non-Western contexts, along with more theoretical and methodological groundwork to

harmonize concepts and measurements of algorithmic fairness perceptions. Finally, we advocate for more interdisciplinary research that provides empirical evidence for developers to design and decision-makers to implement ADM adhering to a society-in-the-loop framework.

Acknowledgements

We express our gratitude to Kimon Kieslich, Nils Köbis, and Marco Lünich for their valuable feedback on previous versions of the manuscript.



Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This literature review was conducted as part of the project Fair Artificial Intelligence Reasoning (FAIR). From May 2019 to April 2020 the project was funded by the VolkswagenStiftung (Volkswagen Foundation), Germany [grant number 95998].

ORCID iDs

Christopher Starke  <https://orcid.org/0000-0001-7899-6029>
Birte Keller  <https://orcid.org/0000-0002-3145-5206>

Supplemental material

Supplemental material for this article is available online.

Note

1. Ten studies do not indicate the country in which the data was collected.

References

- Acikgoz Y, Davison KH, Compagnone M, et al. (2020) Justice perceptions of artificial intelligence in selection. *International Journal of Selection and Assessment* 28(4): 399–416. DOI: 10.1111/ijsa.12306
- Ahnert G, Smirnov I, Lemmerich F, et al. (2021) The FairCepton: A Framework for Measuring Human Perceptions of Algorithmic Fairness. In: Adjunct Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization, pp. 401–403. Utrecht, Netherlands. DOI: 10.1145/3450614.3463291.
- Albach M and Wright JR (2021) The Role of Accuracy in Algorithmic Process Fairness Across Multiple Domains. In: Proceedings of the ACM Conference on Economics and Computation, pp. 29–49. Budapest, Hungary. DOI: 10.1145/3465456.3467620.
- AlgorithmWatch (2019) *Automating Society: Taking Stock of Automated Decision-Making in the EU*. Berlin. Available at: https://algorithmwatch.org/wp-content/uploads/2019/01/Automating_Society_Report_2019.pdf (accessed 18 July 2022).
- Araujo T, Helberger N, Kruijkemeier S, et al. (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & Society* 35(3): 611–623. DOI: 10.1007/s00146-019-00931-w
- Bai B, Dai H, Zhang D, et al. (2020) The impacts of algorithmic work assignment on fairness perceptions and productivity: Evidence from field experiments. *SSRN Electronic Journal* 2020: 1–32. DOI: 10.2139/ssrn.3550887.
- Bankins S, Formosa P, Griep Y, et al. (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*: 2022 1–19. DOI: 10.1007/s10796-021-10223-8.
- Bansak K, Ferwerda J, Hainmueller J, et al. (2018) Improving refugee integration through data-driven algorithmic assignment. *Science* 359(6373): 325–329. DOI: 10.1126/science.aao4408
- Barabas C, Doyle C, Rubinovitz JB, et al. (2020) Studying Up: Reorienting the study of algorithmic fairness around issues of power. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 167–176. New York, USA. DOI: 10.1145/3351095.3372859.
- Barocas S and Selbst A (2016) Big data's disparate impact. *California Law Review* 104(1): 671–729. DOI: 10.15779/Z38BG31
- Binns R (2018a) Fairness in machine learning: Lessons from political philosophy. *Proceedings of Machine Learning Research* 81: 149–159. Available at: <http://arxiv.org/abs/1712.03586> (accessed 18 July 2022).
- Binns R (2018b) What can political philosophy teach us about algorithmic fairness? *IEEE Security & Privacy* 16(3): 73–80. DOI: 10.1109/MSP.2018.2701147
- Binns R (2022) Human judgment in algorithmic loops: Individual justice and automated decision-making. *Regulation & Governance* 16(1): 1–15. DOI: 10.1111/rego.12358.
- Binns R, Van Kleek M, Veale M, et al. (2018) 'It's Reducing a Human Being to a Percentage'; Perceptions of Justice in Algorithmic Decisions. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14. Montreal, Canada. DOI: 10.1145/3173574.3173951.
- Booth A, Sutton A and Papaioannou D (2016) *Systematic Approaches to a Successful Literature Review*. Los Angeles: Sage Publications.
- Cheng H-F, Stapleton L, Wang R, et al. (2021) Soliciting Stakeholders' Fairness Notions in Child Maltreatment Predictive Systems. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, pp. 1–17. Yokohama, Japan. DOI: 10.1145/3411764.3445308.
- Chouldechova A (2017) Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5(2): 153–163. DOI: 10.1089/big.2016.0047
- Colquitt JA and Rodell JB (2015) Measuring justice and fairness. In: Cropanzano RS and Ambrose ML (eds) *The Oxford Handbook of Justice in the Workplace*. Oxford: Oxford University Press, pp. 187–202.
- de Cremer D and McGuire J (2022) Human–algorithm collaboration works best if humans lead (because it is fair!). *Social Justice Research* 35(1): 33–55. DOI: 10.1007/s11211-021-00382-z
- Deutsch M (1975) Equity, equality, and need: What determines which value will be used as the basis of distributive justice?

- Journal of Social Issues* 31(3): 137–149. DOI: 10.1111/j.1540-4560.1975.tb01000.x
- Dietvorst BJ, Simmons JP and Massey C (2015) Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144(1): 114–126. DOI: 10.1037/xge0000033
- Dodge J, Vera Liao Q, Zhang Y, et al. (2019) Explaining models: An empirical study of how explanations impact fairness judgment. In: Proceedings of the International Conference on Intelligent User Interfaces, pp. 275–285. Marina del Rey, USA. DOI: 10.1145/3301275.3302310.
- Dwork C, Hardt M, Pitassi T, et al. (2012) Fairness Through Awareness. In: Proceedings of the 3rd Innovations in Theoretical Computer Science, pp. 214–226. New York, USA. DOI: 10.1145/2090236.2090255.
- Eubanks V (2018) *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York: St Martin's Press.
- European Commission (2019) *Ethics guidelines for trustworthy AI*. Available at: <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai> (accessed 18 July 2022).
- Favaretto M, De Clercq E and Elger BS (2019) Big data and discrimination: Perils, promises and solutions. A systematic review. *Journal of Big Data* 6(12): 1–27. DOI: 10.1186/s40537-019-0177-4
- Gajane P and Pechenizkiy M (2017) *On Formalizing Fairness in Prediction with Machine Learning*. Available at: <http://arxiv.org/abs/1710.03184> (accessed 18 July 2022).
- Glikson E and Woolley AW (2020) Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14(2): 627–660. DOI: 10.5465/annals.2018.0057
- Greenberg J (1990) Organizational justice: Yesterday, today, and tomorrow. *Journal of Management* 16(2): 399–432. DOI: 10.1177/014920639001600208
- Greenberg J (1993) The social side of fairness: Interpersonal and informational classes of organizational justice. In: Cropanzano R (ed) *Justice in the Workplace: Approaching Fairness in Human Resource Management*. Hillsdale: Lawrence Erlbaum Associates, pp. 79–103.
- Grgić-Hlača N, Redmiles EM, Gummadi KP, et al. (2018a) Human Perceptions of Fairness in Algorithmic Decision Making: A Case Study of Criminal Risk Prediction. In: Proceedings of the World Wide Web Conference, pp. 903–912. New York, USA. DOI: 10.1145/3178876.3186138.
- Grgić-Hlača N, Zafar MB, Gummadi KP, et al. (2018b) Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In: Proceedings of the 32nd AAAI Conference on Artificial Intelligence, pp. 51–60. New Orleans, USA. DOI: 10.1609/aaai.v32i1.11296
- Grgić-Hlača N, Weller A and Redmiles EM (2020) *Dimensions of Diversity in Human Perceptions of Algorithmic Fairness*. Available at: <https://arxiv.org/abs/2005.00808v1>.
- Hannan J, Chen H-YW and Joseph K (2021) Who Gets What, According to Whom? An Analysis of Fairness Perceptions in Service Allocation. In: Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, pp. 555–565. Virtual Event, USA. DOI: 10.1145/3461702.3462568.
- Harrison G, Hanson J, Jacinto C, et al. (2020) An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 392–402. Barcelona, Spain. DOI: 10.1145/3351095.3372831.
- Helberger N, Araujo T and de Vreese CH (2020) Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Computer Law and Security Review* 39: 1–16. DOI: 10.1016/j.clsr.2020.105456
- Henrich J, Heine SJ and Norenzayan A (2010) The weirdest people in the world? *Behavioral and Brain Sciences* 33(2–3): 61–83. DOI: 10.1017/S0140525X0999152X
- Higgins JP and Deeks JJ (2008) Selecting studies and collecting data. In: Higgins JP and Green S (eds) *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, West Sussex: John Wiley & Sons Ltd, pp. 151–185.
- Hsu S, Li TW, Zhang Z, et al. (2021) Attitudes Surrounding an Imperfect AI Autograder. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–15. Yokohama, Japan. DOI: 10.1145/3411764.3445424.
- Hutchinson B and Mitchell M (2019) 50 Years of Test (Un)fairness: Lessons for Machine Learning. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 49–58. New York, USA. DOI: 10.1145/3287560.3287600.
- Jobin A, Ienca M and Vayena E (2019) The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1(9): 389–399. DOI: 10.1038/s42256-019-0088-2
- Kaibel C, Koch-Bayram I, Biemann T, et al. (2019) Applicant perceptions of hiring algorithms—uniqueness and discrimination experiences as moderators. *Academy of Management Proceedings* 2019(1): 1–6. DOI: 10.5465/AMBPP.2019.210.
- Kasinidou M, Kleanthous S, Barlas P, et al. (2021a) I agree with the decision, but they didn't deserve this. Future Developers' Perception of Fairness in Algorithmic Decisions. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 690–700. New York, USA. DOI: 10.1145/3442188.3445931.
- Kasinidou M, Kleanthous S, Orphanou K, et al. (2021b) Educating Computer Science Students about Algorithmic Fairness, Accountability, Transparency and Ethics. In: Adjunct Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization, pp. 484–490. Utrecht, Netherlands. DOI: 10.1145/3430665.3456311.
- Kelley PG, Yang Y, Heldreth C, et al. (2021) “Happy and assured that life will be easy 10 years from now.”: Perceptions of artificial intelligence in 8 countries. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 1–12. Virtual Event, USA. Available at: <https://arxiv.org/abs/2001.00081>.
- Kieslich K, Lünich M and Marcinkowski F (2021) The threats of artificial intelligence scale (TAI): Development, measurement and test over three application domains. *International Journal of Social Robotics* 13(1): 1–15. DOI: 10.1007/s12369-020-00734-w.
- Kieslich K, Keller B and Starke C (2022) Artificial intelligence ethics by design. Evaluating public perception on the importance of ethical design principles of artificial intelligence. *Big Data & Society* 9(1): 1–19. DOI: 10.1177/2F20539517221092956
- Kleinberg J, Mullainathan S and Raghavan M (2017) Inherent trade-offs in the fair determination of risk scores. *Leibniz International Proceedings in Informatics* 67(43): 1–23. DOI: 10.4230/LIPIcs.ITCS.2017.43

- Köbis N, Starke C and Rahwan I (2021) Artificial Intelligence as an Anti-Corruption Tool (AI-ACT)—Potentials and Pitfalls for top-down and bottom-up approaches. Available at: <http://arxiv.org/abs/2102.11567>.
- Köchling A and Wehner MC (2020) Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research* 13(3): 1–54. DOI: 10.1007/s40685-020-00134-w.
- Koene A, Perez E, Ceppi S, et al. (2017) Algorithmic Fairness in Online Information Mediating Systems. In: Proceedings of the ACM on Web Science Conference, pp. 391–392. New York, USA. DOI: 10.1145/3091478.3098864.
- Kusner M, Loftus J, Russell C, et al. (2017) Counterfactual Fairness. In: Proceedings of the International Conference on Neural Information Processing Systems, pp. 4066–4076. Red Hook, USA.
- Lee MK (2018) Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5(1): 1–16. DOI: 10.1177/2053951718756684
- Lee MK and Baykal S (2017) Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In: Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing, pp. 1035–1048. Portland, USA. DOI: 10.1145/2998181.2998230.
- Lee MK and Rich K (2021) Who Is Included in Human Perceptions of AI?: Trust and Perceived Fairness around Healthcare AI and Cultural Mistrust. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14. Yokohama, Japan. DOI: 10.1145/3411764.3445570.
- Lee MK, Kim JT and 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. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 3365–3376. Denver, USA. DOI: 10.1145/3025453.3025884.
- Lee MK, Jain A, Cha HJIN, et al. (2019a) 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. DOI: 10.1145/3359284
- Lee MK, Kusbit D, Kahng A, et al. (2019b) Webuildai: Participatory framework for algorithmic governance. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–35. DOI: 10.1145/3359283
- Lefebvre C, Manheimer E and Glanville J (2008) Searching for studies. In: Higgins JP and Green S (eds) *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: John Wiley & Sons Ltd, pp. 95–150.
- Lepri B, Oliver N, Letouze E, et al. (2018) Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges. *Philosophy and Technology* 31(4): 611–627. DOI: 10.1007/s13347-017-0279-x
- Leventhal GS (1980) What should be done with equity theory? In: Gergen KJ, Greenberg MS and Willis RH (eds) *Social Exchange*. Boston: Springer, US, pp. 27–55. DOI: 10.1007/978-1-4613-3087-5_2.
- Marcinkowski F, Starke C, et al. (2019) Wann ist künstliche Intelligenz (un)fair? Ein sozialwissenschaftliches Konzept von KI-fairness. In: Hofmann J, Kersting N and Ritz C (eds) *Politik in Der Digitalen Gesellschaft: Zentrale Problemfelder Und Forschungsperspektiven*. Bielefeld: transcript, pp. 269–288.
- Marcinkowski F, Kieslich K, Starke C, et al. (2020) Implications of AI (un-)fairness in higher education admissions: The Effects of Perceived AI (Un-)Fairness on Exit, Voice and Organizational Reputation. In: Proceedings of the ACM Conference on Fairness, Accountability and Transparency, pp. 122–130. Barcelona, Spain. DOI: 10.1145/3351095.3372867.
- Miller SM and Keiser LR (2021) Representative bureaucracy and attitudes toward automated decision making. *Journal of Public Administration Research and Theory* 31(1): 150–165. DOI: 10.1093/jopart/muaa019
- Moher D, Liberati A, Tetzlaff J, et al. (2009) Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. *BMJ (Online)* 339(7716): 332–336. DOI: 10.7326/0003-4819-151-4-200908180-00135
- Nagtegaal R (2021) The impact of using algorithms for managerial decisions on public employees’ procedural justice. *Government Information Quarterly* 38(1): 1–10. DOI: 10.1016/j.giq.2020.101536
- Nam T (2020) Do the right thing right! understanding the hopes and hypes of data-based policy. *Government Information Quarterly* 37(3): 1–10. DOI: 10.1016/j.giq.2020.101491
- Newman DT, Fast NJ and Harmon DJ (2020) When eliminating bias isn’t fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes* 160: 149–167. DOI: 10.1016/j.obhdp.2020.03.008
- Noble SM, Foster LL and Craig SB (2021) The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment* 29(2): 139–153. DOI: 10.1111/ijssa.12320
- Noriega-Campero A, Garcia-Bulle B, Cantu LF, et al. (2020) Algorithmic targeting of social policies: Fairness, accuracy, and distributed governance. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 241–251. Barcelona, Spain. DOI: 10.1145/3351095.3375784.
- Nyarko J, Goel S and Sommers R (2021) Breaking Taboos in Fair Machine Learning: An Experimental Study. In: Proceedings of Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 1–11. New York, USA. DOI: 10.1145/3465416.348329.1.
- OECD (2019) *Recommendation of the Council on OECD Legal Instruments Artificial Intelligence*. Paris. Available at: <https://www.oecd.ai/ai-principles> (accessed 18 July 2022).
- Perez Vallejos E, Koene A, Portillo V, et al. (2017) Young people’s policy recommendations on algorithm fairness. In: Proceedings of the ACM on Web Science Conference, pp. 247–251. New York, USA. DOI: 10.1145/3091478.3091512.
- Petticrew M and Roberts H (2006) *Systematic Reviews in the Social Sciences*. Oxford: Blackwell Publishing Ltd.
- Pierson E (2017) *Demographics and discussion influence views on algorithmic fairness*. Available at: <http://arxiv.org/abs/1712.09124>.
- Plane AC, Redmiles EM, Mazurek ML, et al. (2017) Exploring user perceptions of discrimination in online targeted

- advertising. In: Proceedings of the USENIX Security Symposium, pp. 935–951. Vancouver, Canada.
- Rahwan I (2018) Society-in-the-loop: Programming the algorithmic social contract. *Ethics and Information Technology* 20(1): 5–14. DOI: 10.1007/s10676-017-9430-8
- Richardson R, Schultz JM and Southerland VM (2019) *Litigating Algorithms 2019 US Report: New Challenges to Government Use of Algorithmic Decision Systems*. New York. Available at: <https://ainowinstitute.org/litigatingalgorithms-2019-us.html> (accessed 18 July 2022).
- Saha D, Schumann C, McElfresh DC, et al. (2020) Human Comprehension of Fairness in Machine Learning. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp. 152–152. New York, USA. DOI: 10.1145/3375627.3375819.
- Sambasivan N, Arnesen E, Hutchinson B, et al. (2021) Re-imagining Algorithmic Fairness in India and Beyond. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, pp. 315–328. Virtual Event, Canada. DOI: 10.1145/3442188.3445896.
- Saxena NA, Huang K, DeFilippis E, et al. (2020) How do fairness definitions fare? Examining public attitudes towards algorithmic definitions of fairness. *Artificial Intelligence* 283: 1–15. DOI: 10.1016/j.artint.2020.103238
- Schlicker N, Langer M, Ötting SK, et al. (2021) What to expect from opening up ‘black boxes’? Comparing perceptions of justice between human and automated agents. *Computers in Human Behavior* 122: 1–16. DOI: 10.1016/j.chb.2021.106837.
- Schoeffer J, Machowski Y and Kuehl N (2021) A Study on Fairness and Trust Perceptions in Automated Decision Making. In: Joint Proceedings of the ACM IUI 2021 Workshops, pp. 1–12. College Station, USA. DOI: 10.48550/arXiv.2103.04757.
- Shin D (2020) User perceptions of algorithmic decisions in the personalized AI system: Perceptual evaluation of fairness, accountability, transparency, and explainability. *Journal of Broadcasting & Electronic Media* 64(4): 541–565. DOI: 10.1080/08838151.2020.1843357
- Shin D (2021) The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human Computer Studies* 146: 102551. DOI: 10.1016/j.ijhcs.2020.102551
- Shin D and Park YJ (2019) Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior* 98: 277–284. DOI: 10.1016/j.chb.2019.04.019
- Shin D, Zhong B and Biocca FA (2020) Beyond user experience: What constitutes algorithmic experiences? *International Journal of Information Management* 52: 1–11. DOI: 10.1016/j.ijinfomgt.2019.102061
- Shulner-Tal A, Kuflik T and Kliger D (2022) Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users’ perceptions of fairness toward an algorithmic system. *Ethics and Information Technology* 24(1): 1–13. DOI: 10.1007/s10676-022-09623-4
- Sloane M and Moss E (2019) AI’s social sciences deficit. *Nature Machine Intelligence* 1: 330–331.
- Smith J, Sonboli N, Fiesler C, et al. (2020) Exploring User Opinions of Fairness in Recommender Systems. In: CHI’20 Workshop on Human-Centered Approaches to Fair and Responsible AI, pp. 1–4. Honolulu, USA.
- Srivastava M, Heidari H and Krause A (2019) Mathematical notions vs. Human perception of fairness: A descriptive approach to fairness for machine learning. In: Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 2459–2468. Anchorage, USA. DOI: 10.1145/3292500.3330664.
- Starke C and Lünich M (2020) Artificial intelligence for political decision-making in the European Union: Effects on citizens’ perceptions of input, throughput, and output legitimacy. *Data & Policy* 2: e16. DOI: 10.1007/s10618-017-0506-1
- Suen HY, Chen MYC and Lu SH (2019) Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior* 98: 93–101. DOI: 10.1016/j.chb.2019.04.012
- Sun L and Tang Y (2021) Data-Driven discrimination, perceived fairness, and consumer trust—the perspective of consumer attribution. *Frontiers in Psychology* 12: 1–13. DOI: 10.3389/fpsyg.2021.748765.
- van Berkel N, Goncalves J, Hettiachchi D, et al. (2019) Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–21. DOI: 10.1145/3359130
- van Berkel N, Goncalves J, Russo D, et al. (2021) Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–13. Yokohama, Japan. DOI: 10.1145/3411764.3445365.
- Veale M and Binns R (2017) Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society* 4(2): 1–17. DOI: 10.1177/2053951717743530
- Verma S and Rubin J (2018) Fairness Definitions Explained. In: Proceedings of the International Workshop on Software Fairness, pp. 1–7. Gothenburg, Sweden. DOI: 10.1145/3194770.3194776.
- Wang AJ (2018) Procedural justice and risk-assessment algorithms. *SSRN Electronic Journal* 2018: 1–31. DOI: 10.2139/ssrn.3170136.
- Wang R, Harper FM and Zhu H (2020) Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14. Honolulu, USA. DOI: 10.1145/3313831.3376813.
- Wong PH (2020) Democratizing algorithmic fairness. *Philosophy and Technology* 33(2): 225–244. DOI: 10.1007/s13347-019-00355-w
- Wonseok J, Young Woo K and Yeonheung K (2021) Who made the decisions: Human or robot umpires? The effects of anthropomorphism on perceptions toward robot umpires. *Telematics and Informatics* 64: 1–10. DOI: 10.1016/j.tele.2021.101695
- Woodruff A, Fox SE, Rouso-Schindler S, et al. (2018) A qualitative exploration of perceptions of algorithmic fairness. In: Proceedings of the CHI Conference on Human Factors in Computing Systems, pp. 1–14. Montreal, Canada. DOI: 10.1145/3173574.3174230.
- Zafar MB, Valera I, Rodriguez MG, et al. (2017) From parity to preference-based notions of fairness in classification. In:

- Proceedings of the International Conference on Neural Information Processing Systems, pp. 1–11. Long Beach, USA.
- Zhou J, Verma S, Mittal M, et al. (2021) Understanding Relations Between Perception of Fairness and Trust in Algorithmic Decision Making. In: Proceedings of the International Conference on Behavioral and Social Computing (BESC 2021), pp. 1–8. Doha, Qatar. DOI: 2109.14345v1.
- Žliobaitė I (2017) Measuring discrimination in algorithmic decision making. *Data Mining and Knowledge Discovery* 31(4): 1060–1089.