

Social impacts of algorithmic decision-making: A research agenda for the social sciences

Big Data & Society
 January–June: 1–13
 © The Author(s) 2022
 Article reuse guidelines:
sagepub.com/journals-permissions
 DOI: 10.1177/20539517221089305
journals.sagepub.com/home/bds



Frederic Gerdon^{1,2} , Ruben L Bach³ , Christoph Kern³ 
 and Frauke Kreuter^{2,4} 

Abstract

Academic and public debates are increasingly concerned with the question whether and how algorithmic decision-making (ADM) may reinforce social inequality. Most previous research on this topic originates from computer science. The social sciences, however, have huge potentials to contribute to research on social consequences of ADM. Based on a process model of ADM systems, we demonstrate how social sciences may advance the literature on the impacts of ADM on social inequality by uncovering and mitigating biases in training data, by understanding data processing and analysis, as well as by studying social contexts of algorithms in practice. Furthermore, we show that fairness notions need to be evaluated with respect to specific outcomes of ADM systems and with respect to concrete social contexts. Social sciences may evaluate how individuals handle algorithmic decisions in practice and how single decisions aggregate to macro social outcomes. In this overview, we highlight how social sciences can apply their knowledge on social stratification and on substantive domains of ADM applications to advance the understanding of social impacts of ADM.

Keywords

Algorithms, social inequality, fair machine learning, social impacts of AI, algorithmic decision-making, artificial intelligence

Introduction

As the increasing use of algorithmic decision-making (ADM) has raised concerns about its social impacts and particularly about new or reinforced social inequalities, research quantifying consequences of ADM for social inequality remains in demand. Understanding the sources and effects of social inequality is one of the core competencies—and responsibilities—of the social sciences. To facilitate a cross-disciplinary discussion and additional research in this area, we use a process model of automated decision-making to highlight when and where social inequality may arise from ADM systems. Focusing on the data generation, data analysis, and implementation of ADM systems, we suggest a roadmap and research avenues for social scientists interested in answering the increased calls for the study of social impacts of ADM.

ADM is used as an umbrella term for a variety of systems that are used to assist or replace human deciders (see AlgorithmWatch, 2019). For instance, judges may use recidivism risk scores predicted by algorithms trained on decades of criminal records to determine bail decisions (Stevenson, 2018), mortgage lenders can base interest

rates on default risks predicted by algorithms (Bartlett et al., 2019), and public social services may draw on algorithmic support to make decisions on financial aids (Lind and Wallentin, 2020).

ADM systems are based on predictions from models that process historical data, which contain both inputs (“predictors,” “features,” “independent variables,” “x”) and one or more outputs (“label,” “outcome,” “dependent variable,” “y”). The goal of data processing is to “learn” associations between inputs and output from the past to make predictions where the output is still unknown. Predictions are

¹Mannheim Centre for European Social Research, University of Mannheim, Mannheim, Germany

²Department of Statistics, Ludwig-Maximilians-Universität (LMU) München, München, Germany

³School of Social Sciences, University of Mannheim, Mannheim, Germany

⁴University of Maryland, College Park, MD, USA

Corresponding author:

Frederic Gerdon, Mannheim Centre for European Social Research, University of Mannheim, A 5,6, Mannheim, 68131, Germany.

Email: fgerdon@mail.uni-mannheim.de



then used to decide whether some action should be taken or not. While our focus is on ADM systems that draw on some automated learning, these systems can generally vary in the complexity of how inputs determine outputs—including simple threshold rules for single input variables—, as well as in the extent to which humans are involved in the final decisions (see related definitions and surrounding discussions in AlgorithmWatch (2019) and European Parliament, Directorate General for Parliamentary Research Services et al. (2019)).

ADM seems promising as an alternative to (pure) human decision-making, as human decisions may be just as or even more biased than ADM, with ADM potentially having higher efficacy (Miller, 2018), transparency, and accountability (Mayson, 2019). However, concerns have been raised about algorithms exacerbating social inequality and discriminating against certain societal groups, for example, due to learning biases from historical training data (e.g. Zou and Schiebinger, 2018).

A recent example of an ADM system that raised such concerns is a system that has been tested by the Public Employment Service Austria (AMS). This system classifies job seekers into three groups, depending on their predicted chances to find a new employment (Lopez, 2019). The system builds groups of feature combinations based on, for example, gender, age, nationality, education, and previous contact with AMS, and predicts short- as well as long-term chances of integration into the labor market (Gamper et al., 2020). The assignment to a group can influence which kind of assistance is given to an individual: for instance, Kopf (2019) argues that while all job seekers are supported by the employment agency, individuals with low chances for re-employment would usually profit more from intensified assistance than from, for example, qualification measures. However, concerns arise if, for example, women, with all other characteristics held equal, had lower scores than men. Such concerns sparked discourse regarding the discriminatory potentials of this system (see Kopf, 2019; Lopez, 2019).

While similar decisions have been made without algorithmic assistance in the past, novel ADM systems have specific features that create new and amplify old challenges. First, these systems make use of new technical devices and facilities, unprecedented amounts of data, and novel techniques of data analysis that allow deciders to employ new decision-making strategies to approach old problems. Second, ADM systems constitute socio-technical systems that entail machines and humans (Selbst et al., 2019): they are pervaded by human decisions and cultural notions that we need to scrutinize (Seaver, 2019) to understand potential detrimental effects for society.

Scholars from various disciplines have called for examining algorithmic outcomes to avoid or mitigate undesired consequences of ADM (Kusner and Loftus, 2020; Zou and Schiebinger, 2018). Previous research from computer

science (Mehrabi et al., 2019), legal studies (Wachter, 2020), and philosophical (Mittelstadt et al., 2016) perspectives discussed algorithmic, structural, and ethical problems with ADM. Joyce et al. (2021) and Liu (2021) provide a general overview of sociological perspectives on related artificial intelligence.

Drawing on previous literature, our own work on ADM systems, and a previously developed big data processing model (Weyer et al., 2018), we here highlight areas in which social scientists can (and should) use their expertise to contribute to the debate of equitable ADM. We show how a social science perspective on data generating processes, analytical challenges, and implementations can help anticipate (undesired) social impacts of ADM.

A process model of ADM

To understand how social inequality, here defined as “the unequal distribution of valued resources, opportunities, and positions among the members of a population in a given space and time” (Otte et al., 2021: 362), can arise or be amplified through ADM systems, attention needs to be paid to the distribution of opportunities and restrictions leading into and out of the ADM system. While inequality not always necessarily constitutes injustice, it is oftentimes considered an undesired property of ADM systems (see Kuppler et al. (2021) for a detailed discussion on distributive justice in ADM).

A major path via which algorithms—just like human-made decisions—may affect such distributions is discriminatory behavior. By *discrimination* we mean “an action or practice that excludes, disadvantages, or merely differentiates between individuals or groups of individuals on the basis of some ascribed or perceived trait” (Kohler-Hausmann, 2011), such as gender and race. Computer science research on *Fair Machine Learning* (Fair ML) aims to tackle discrimination by investigating how algorithms can be designed to make predictions *fair*, that is, without “prejudice or favoritism toward an individual or a group based on their inherent or acquired characteristics” (Mehrabi et al., 2019: 1).

Social implications of ADM systems do not only arise through biased predictions, but from implementations of decisions within a social environment. Even if fairness on the prediction level is present, disparate impact can occur (Feldman et al., 2015), for example, because impacts of ADM decisions are hard to factor into the preceding data analysis (Kusner et al., 2019), or because human deciders rely disproportionately on the ADM-based recommendations. Recent research extended the notion of fairness to include actual inequality effects resulting from algorithmic discrimination in the social context in which it is placed (see section “Data preparation and analysis—from fairness in algorithmic output to fairness in social impact”) and to frame such effects in terms of causal impact (Kasy and

Abebe, 2021). To investigate these social impacts of ADM, a social science perspective becomes particularly valuable.

To discuss how ADM systems may impact social inequality, we adapt a “big data process model” (Weyer et al., 2018: 74), by breaking down the ADM process into three steps (based on Weyer et al., 2018). We discuss how social inequalities may be shaped in each step:

- **Data generation:** Data bases may be biased, for example, due to historical discrimination against social groups or incomplete data availability.
- **Data preparation and analysis:** An algorithm may adapt or even reinforce biases that are already present in the data. This includes the choice and construction of variables that serve as the input for the algorithm, the choice of fairness metrics for identifying biases, and the choice of bias mitigation measures.
- **Implementation:** The way ADM systems ultimately affect inequality depends on their implementation within contexts (for contexts, see Weyer et al., 2018). Human decision-makers, if present, might handle algorithmic recommendations differently, and those affected by ADM-based decision might differ in their reactions. This step also includes how single decisions aggregate to social outcomes and how human behavior feeds back into the data.

The different forms of biases may propagate through the ADM process and can be reinforced or mitigated along the way.

We briefly illustrate this three-step model with an example. The data basis of the Austrian AMS model that classifies individuals according to their labor market integration chances likely reflects historical unequal labor market participation rates, for example for women (*data generation*). The data analysis itself potentially manifests this bias, if, for example, the model resulting from training assigned women—all else being equal—lower employability scores than men (*data analysis*). Then, we also need to ask about the actual consequences of the system (*implementation*). For example, Kopf (2019) argues that women were ultimately under-represented in the lowest employability group. However, based on data reported by Gamper et al. (2020) on group assignment at the beginning of unemployment, Allhutter et al. (2020) note that the share of women was roughly double the share of men in the lowest employability group. Allhutter et al. (2020) suspect that varying conclusions may result from considering different models, time frames, or (sub-)populations. We therefore need to carefully scrutinize whether or to which extent women would be effectively disadvantaged by the system in practice. Furthermore, we need to know under which conditions human deciders adopt or disagree with the predicted scores and how job seekers subsequently change their job search behavior (also see Allhutter et al., 2020). In a feedback

loop, such behavior may flow into the data basis for future model building. Finally, we need to understand how potentially discriminatory decisions on the individual level will manifest on the macro-level in the long term.

Sources of bias and social impacts along the ADM process

In the following subsections, we explore each of the potential sources of inequalities that we described in the previous section. We follow the ADM process model step by step and show how the social sciences have already contributed to researching problems related to potentially discriminatory ADM. We also identify promising research questions related to social inequality impacts of ADM that the social sciences could investigate.

Data generation—historical bias and selective participation

Algorithms can be trained on a variety of data, ranging from governmental records, such as individual labor market histories, and survey data to digital data created through individuals’ online activities and interactions with digital devices. If bias is present in the data sets used, unfair or discriminatory outputs may result. While using data to make predictions is not exclusive to algorithms, specific aspects in data generation require heightened attention in ADM. In this section, we focus on exemplary problems in biased data sets (Rodolfa et al., 2021; Sen et al., 2019) and refer to previous literature for more general introductions (e.g. Groves, 2004).

Two major sources for bias can arise in the data generation step (for a detailed overview, e.g. Mehrabi et al., 2019). The first source covers all those cases where data used to develop an ADM system contains historical discrimination. That is, an outcome is unequally distributed between individuals with different characteristics such as gender and race, after controlling for other characteristics of the individuals that cause variation in the outcome. The mechanisms creating such discrimination are manifold and depend on the concrete context, for which social sciences can provide domain-specific knowledge. In the labor market example above, historical labor market records may show that, after controlling for other individual characteristics, women had worse re-employment chances than men after losing their job in the past. Similarly, historical criminal records may insinuate that, all other characteristics of an individual held constant, black offenders had higher risks of recidivism once released from jail than white offenders.

The second source comprises biases due to selective participation and representation of social groups in data generation and collection (see Mehrabi et al., 2019). Selective

participation can introduce a mismatch between the data that is used for training a prediction model and the ultimate target population that is affected in its application. If important subgroups are misrepresented in the training step, high error rates (and ultimately incorrect decisions) may result once the model is confronted with the target data in the deployment phase (Daumé III, 2017). Unequal participation in the generation and collection of digital data constitutes a particular challenge for those ADM systems that rely on them. Previous research has shown that the use of information and communication technology is often selective, for example, with respect to digital skills, age, and socio-economic status (Hargittai and Hsieh, 2013; Lutz, 2019). Models trained on such data may thus find relationships that hold only for the group of individuals using such technology. That is, individuals who are already disadvantaged because they do not use specific digital technologies could also be disadvantaged by an ADM system if the system cannot consider their behaviors and preferences (Lerman, 2013).

Social scientists are needed to identify coverage issues due to differences in social characteristics, digital skills, trust, and privacy concerns in the data collection process. Social scientists, and particularly survey researchers, can contribute to tackling representation issues of training data by applying weighting methods or improving data collections. Designing, conducting, and evaluating various forms of data collection processes such that the acquired sample resembles the target population of interest is a core task of survey research. Recent work in survey research investigates coverage and representation issues in the context of digital data and data collected via smartphones and sensors and introduces methods for adjusting non-random samples (Baker et al., 2013; Japac et al., 2015; Keusch et al., 2020). This includes, for example, pseudo-weighting approaches that allow to correct for biases due to selective participation by leveraging information from an auxiliary reference sample (Elliott and Valliant, 2017). Note that such techniques are closely related to adaptation approaches that have been proposed in computer science to account for covariate shift between training and test data (Daumé III, 2017). Weighting techniques from survey research could similarly be utilized to adjust (survey- and non-survey-based) training data if a suitable reference sample that resembles the target population can be found and both datasets include structural information about the entities of interest (e.g. socio-demographic attributes of individuals or make and type of digital devices). While applying pre-processing techniques such as re-weighting may not be feasible in all ADM contexts, recent work on post-processing predictions exemplifies how ideas from survey research (mass imputation; Yang and Kim, 2020) and computer science (multi-calibration; Hebert-Johnson et al., 2018) can be combined to tackle misrepresentation in training data (Kim et al., 2022).

In addition to historical bias and representation bias, ADM can be adversely affected by using mismeasured variables. Using proxy variables such as healthcare costs as a proxy for health needs can obscure differences in the true outcome of interest when, for example, black individuals generated lower healthcare costs than white individuals once the true health status is held constant (Obermeyer et al., 2019). Such measurement bias can be directly connected to social science work on measurement errors and thus represents one example of how social science already contributes to researching social impacts of ADM (Boeschoten et al., 2020; Jacobs and Wallach, 2021). Moreover, the contextual nature of some individual characteristics and behaviors may not be amenable to quantification and therefore, ADM system cannot cover these characteristics appropriately, such as context-sensitive combinations of protected attributes relating to intersectional discrimination (Mann and Matzner, 2019). These may be only subtly present in social interactions, lead to discrimination, and be insufficiently captured in automated analysis (see section “Data preparation and analysis—from fairness in algorithmic output to fairness in social impact”).

To conclude, biases in datasets gain renewed momentum in the context of ADM for three reasons. First, it is likely that the increased quantity of predictions and decisions that a model can make compared to a human decider will intensify inequalities that are already present in the data. Second, relying on ADM systems increases the importance of patterns in the data in comparison to the importance of heuristics of human decision-makers (but see section “Implementation—micro-interaction with ADM and macro-social outcomes”). Third, the amount of data produced and used in ADM systems has considerably increased with the advance of digital technologies. Social sciences can apply methodological and domain knowledge (a) to better understand how situation-specific biases may be present already in the data collection stage of ADM processes and (b) to explore how advances in survey methods can be used to correct such biases.

Research avenues:

- How can we utilize methodological advances in survey research to correct biases in data due to selective participation and improve the data input for ADM?
- How can we extend research on digital divides and technical competencies to study inequality in being covered by ADM systems (Lutz, 2019)?

Data preparation and analysis—from fairness in algorithmic output to fairness in social impact

Data preparation and analysis is the step in which developers work with data and construct and refine algorithms. This process entails manifold interpretations and decisions,

including, ideally, considerations on how to produce fair outputs. In this section, we outline how algorithms may produce biased predictions due to biased data or decisions during the modeling process. We give a brief survey on the mainly computer scientific research field of Fair ML. Then, we show how a social science perspective can contribute to the identification of meaningful fairness criteria in social contexts, particularly when considering social impacts of ADM systems on macro-level social outcomes and public perceptions of fairness.

Approaches in fair machine learning. Fair ML is a research field that investigates the fairness of machine learning algorithms. This research branch produces important contributions by proposing fairness metrics and improving algorithm design such that individuals are less likely to be assessed by characteristics that should not matter for taking a decision (“protected attributes”). Such steps are necessary as otherwise, algorithms might reproduce existing biases or exacerbate inequalities even when the data sources are unbiased (Aghaei et al., 2019). For example, this is the case when prediction error rates differ between groups (Rodolfa et al., 2021). Various steps in the construction of variables (“feature engineering”), such as how race is coded, may also introduce biases (Rodolfa et al., 2021).

Fairness definitions oftentimes are formal measures based on rates of correct and incorrect predictions for individuals of different social groups for which non-discrimination should be ensured (Corbett-Davies and Goel, 2018). For instance, an algorithm might be tasked with assigning job seekers into two classes: those with high or low chances of finding a new job. The algorithm could be trained with data that show past job market outcomes of job seekers. The algorithm tries to combine the characteristics of these individuals to build a model that predicts chances of labor market integration as accurately as possible. The prediction outputs can be evaluated by comparing the predicted with the observed outcomes in the data.

Several fairness definitions specify how error rates should be balanced across different groups of individuals. As an example, an algorithm may be considered fair if it results in equal false negative rates (*equal opportunity*; Hardt et al., 2016) or equal false positive rates (*predictive equality*; Rodolfa et al., 2021) between members of different groups (e.g. men and women). A related definition is *equalized odds*, which means that members of different groups experience both false negatives and false positives at the same rate (Hardt et al., 2016). This principle can be applied to various error metrics and their combinations (e.g. false discovery rates, false omission rates, accuracy), resulting in a variety of group-based fairness notions. Furthermore, subgroup fairness (Hebert-Johnson et al., 2018) and individual fairness (Dwork et al., 2012) notions have been proposed that expand beyond comparisons of

error rates on the group level (e.g. by considering intersections of gender and race).

Research in Fair ML resulted in various methods and tools that may mitigate biases at different stages of the modeling pipeline (Berk et al., 2018; Mehrabi et al., 2019). *Pre-processing* techniques can be used to eliminate sources of unfairness in the data prior to model training, for example, by removing dependencies between legitimate factors and protected attributes (Johndrow and Lum, 2017). *In-processing* techniques aim at modifying the model building process itself, for example, by introducing fairness constraints in the objective function (Berk et al., 2017). *Post-processing* methods may be used to alter the output of a prediction algorithm after model training, for example, by “nudging” predictions towards the true outcome for subgroups where high errors are observed (Kim et al., 2019). These procedures have been shown to mitigate different notions of unfairness at the prediction stage of the ADM process in several applications (Friedler et al., 2018).

Competing fairness definitions and the importance of social context. Many fairness definitions and correction methods have been proposed, and it may prove difficult to choose the definition and technique that is the most appropriate for the given prediction task (see Makhoul et al., 2020). Moreover, some fairness definitions were found incompatible with each other and in conflict with overall accuracy (Berk et al., 2018), while Selbst et al. (2019) discuss as “formalism trap” whether an appropriate mathematical definition of the complex concept of fairness was even possible.

One major concern in handling fairness boils down to the question: is it better to ignore specific individual characteristics such as gender or race altogether, or should we try to balance, for example, error rates between groups based on these features (Corbett-Davies and Goel, 2018)? Neglecting group membership may, for instance, lead to aggregation bias, meaning that one model is used for all groups although the model works worse for some of the groups (Suresh and Guttag, 2020). In the case of race, Benthall and Haynes (2019) discuss that ignoring race would still possibly lead to racial discrimination as effects of correlates relevant to race and inequality persisted. However, explicitly including race reified this category. Instead, they propose a third alternative of algorithmically finding latent categories that mirror racial segregation. From a social science perspective, this discussion extends to the question which features are considered protected attributes in the first place. While gender and race represent attributes that are commonly considered sensitive and are protected by legislation, sociological research on the intergenerational transmission of resources and education (e.g. van Doorn et al., 2011) raises questions on which concepts purely measure individual merit and which attributes may constitute “hybrid” characteristics that are (at least partly)

socially inherited. Relatedly, using traditional concepts of gender and race for defining protected groups will fail to account for individuals who do not find themselves represented by those categories. Particular attention needs to be paid to intersectional discrimination that may disadvantage individuals based on multiple protected attributes at the same time, for example, gender *and* race: automated analysis of large data bases may contain a plethora of potential protected attributes, suggest new associations between these attributes, and thereby statistically form new groups of people that may then be discriminated against (Mann and Matzner, 2019). Social scientists can scrutinize data, analytical decisions, and outputs with respect to intersectionality in different contexts of ADM applications, and suggest groupings of protected attributes that are contextually relevant.

Eventually, fairness is context-specific. Among others, the choice of a fairness metric may depend on the outcome and which resources will be distributed (Kuppler et al., 2021). The idea of social context is not new in the realm of computer science (Selbst et al., 2019) and is part of pursuing “algorithmic realism” (Green and Viljoen, 2020). For instance, this entails the question whether a system aims at helping or punishing individuals, which implies an emphasis on disparate distributions towards either false negatives (incorrectly excluded from a positive intervention) or false positives (incorrectly included in receiving a negative intervention) (Saleiro et al., 2019). This discussion extends to the broader question on the just or desired allocation principle in a specific ADM application context. Sociological discourse on distributive justice can enlarge computer science’s decision space when it comes to designing allocation systems and selecting bias correction techniques by highlighting which design choices may serve which principle (Kuppler et al., 2021).

Empirical findings on fairness perceptions. Fairness perceptions matter to ADM development for two reasons. First, they are relevant to design socially acceptable ADM systems. Second, the individual evaluation of an algorithm may contribute to how that individual interacts with and acts upon the decision of the ADM system, thereby potentially shaping inequality outcomes.

A comprehensive literature review on fairness perceptions on ADM concludes that perceptions strongly depend on context characteristics, such as the features used by the algorithm and the purpose of the algorithm (Starke et al., 2021). Participants in one study applied some justice principles relevant for human decision-making also to algorithmic decisions, but the concrete style of explaining the algorithm impacted justice perceptions only when the respondent was exposed to multiple styles (Binns et al., 2018). In addition, general trust in ML systems and the features used and *not used* are relevant for fairness judgments (Dodge et al., 2019). Empirically validated frameworks that

define process features relevant to fairness perceptions (Grgić-Hlača et al., 2018) can build the basis for practically applicable guidelines for designing contextually fair algorithms, which is why such work is particularly attractive for future work.

Whether and how individual characteristics such as socio-demographic attributes like age and gender interact with, for example, explanation styles and the impact of the decision situation needs further research and possibly depends on individual affectedness (Pierson, 2018). Experimental evidence suggests that fairness ratings depend on whether respondents’ characteristics are involved in the algorithmic decision, and conservatives were found to be more accepting of using individual characteristics in computer-assisted bail decisions than liberals (Grgić-Hlača et al., 2020).

In conclusion, building fair algorithms is a prerequisite for arriving at fair predictions and, subsequently, decisions. Software toolkits that assist in assessing the fairness of algorithms are available (Bellamy et al., 2019; Saleiro et al., 2019). To advance Fair ML, we can intensify research on fairness perceptions in concrete ADM processes and strengthen the link between distributive justice principles and (fairness in) automated allocation systems (Kuppler et al., 2021). Moreover, Starke et al. (2021) suggest systematizing situation-specific factors such as whether a decision is high-stake or low-stake and the area of application (e.g. decisions in the criminal justice system or hiring) that may shape fairness perceptions.

Research avenues:

- How do contextual information (the purpose of an algorithm) and explanations of algorithm function shape fairness perceptions of ADM processes? How do individual characteristics influence fairness perceptions?
- How can fairness assessment and mitigation techniques be implemented and extended beyond equalizing error rates towards serving context-specific allocation principles?
- How can social science provide domain-specific knowledge to define appropriate, non-discriminatory outcomes for an ADM system, including the consideration of externalities?

Implementation—micro-interaction with ADM and macro-social outcomes

Researchers from different disciplines have demonstrated that the used data and the data analysis performed do not suffice for explaining the social impacts of algorithms (Cowgill and Tucker, 2020; Kleinberg et al., 2018). The question whether the use of an algorithm will produce fair outcomes is not only a question of the fairness of predictions and decisions, but also of their actual impacts in a

social environment (Kusner et al., 2019). In fact, “[...] even fair decisions at the machine learning level may not lead to equitable results in society and the decision-making process may need to compensate for these other inequities” (Rodolfa et al., 2021:304).

The notion of disparate impact helps to understand the difference between the output of an analysis and subsequent societal consequences. Disparate impact refers to effects of practices that result in unintended disadvantages for groups of individuals with certain characteristics (Barocas and Selbst, 2016). For example, even if no discrimination is intended, individuals may be affected differently due to their characteristics. Implementing notions of disparate impact in algorithms is one step to practically achieve fairer results (e.g. Feldman et al., 2015), and computer scientific research developed and applied such extended notions of fairness. Among those are suggestions to ascertain fairness by optimizing how an outcome of interest is expected to be affected in the long term (Liu et al., 2019), choosing fairness metrics that satisfy specific policy goals (Rodolfa et al., 2020), and engaging with the needs of affected population groups to adjust analyses in feedback loops (Noriega-Campero et al., 2018).

These outcome-oriented approaches seem most promising for the development of an encompassing understanding of fairness that contributes to contextually appropriate assessments. A social science perspective can help to analyze the implementation process of ADM in social contexts and to understand interaction processes at the micro-level between algorithms, affected individuals and, in some cases, human deciders, and their macro-social outcomes.

Human versus algorithmic predictions: Empirical evidence from real-life cases. Studying impacts of ADM systems in real-life cases faces the same challenge as other observational social research: it remains unclear *what the outcome would have been* had a decision been taken without an algorithm (see Holland, 1986). Although methods for tackling such problems of causal inference are well known to social scientists, there is so far only little research applying them to the study of social impacts of ADM (such as Cowgill and Tucker, 2017). One notable exception are recidivism prediction algorithms in the USA, where studies find mixed effects regarding the reduction of crime and racial disparity through algorithms (Berk, 2017; Kleinberg et al., 2018; Stevenson, 2018). Stevenson (2018) suggests that even if algorithms made better predictions, they might not necessarily improve relevant outcomes, and judges’ own biases could lead to a sub-optimal use of algorithmic predictions, emphasizing the need to study how human decision-makers rely on algorithms.

Previous research reports mixed findings regarding differences in the accuracy of predictions between algorithms and human deciders. Some find that humans perform worse

than algorithms (Green and Chen, 2019), while others find comparable accuracy and fairness in predictions (Bansak, 2019; Dressel and Farid, 2018; Tan et al., 2018). In addition, in the context of recidivism prediction tasks, it is likely that the characteristics of the defendants will matter: given a risk assessment, human deciders deviated more strongly to unfavorable predictions for black defendants than for white defendants (Green and Chen, 2019).

Overall, the question whether an algorithm can outperform a human decider will have to be evaluated on a case-by-case basis. In those cases where the final decision remains in the hands of a human decider, we also must consider whether and how a human decider is involved and *influenced* by an algorithmic decision or recommendation.

From “automation bias” to “algorithmic aversion”—How human deciders (do not) adopt algorithmic recommendations. Many ADM systems, particularly those in which the stakes are high, involve a human decider who may consider algorithmic predictions in her decisions. While a machine-assisted decision may deviate from a purely human decision, human deciders will not always follow the algorithmic recommendation. Therefore, potential biases inherent in the algorithmic prediction may be alleviated or corrected by human deciders, but humans may also introduce or reinforce discrimination in the process. The interaction of a human decider with an ADM system is likely complex and requires detailed investigation. Research on “human factors” and human-computer interaction provides valuable work that can be applied to the study of ADM systems (Zerilli et al., 2019). The communication between algorithmic recommendations, human deciders, and affected individuals is likely shaped by the complexity of the underlying model. If we want the involved individuals to understand how ADM systems arrive at decisions and to uphold accountability, we need algorithms that can be explained—either by making use of inherently interpretable methods or by employing post-hoc interpretation techniques (Molnar, 2019). Differential social impacts may arise, for example, if explanations are differently effective for social groups and shape the reliance on or compliance with algorithmic recommendations.

Here, we focus on the specific problem of circumstances under which a human decider will be more likely to adopt (or override) an algorithmic recommendation. Two central phenomena characterize human reliance on algorithmic predictions: automation bias and algorithmic aversion. Automation bias refers to errors stemming from human reliance on automated systems such as ADM: while errors of omission refer to cases where someone relies on a flawed algorithmic prediction (false negatives), errors of commission refer to falsely assuming an error (false positives) (Wickens et al., 2015).

Empirical evidence for automation bias has been found, for example, in clinical decision support systems (Goddard

et al., 2012). Research shows that factors such as trust and own experience shape reliance on automated systems (Burton et al., 2019; Cepera et al., 2018; Lee and See, 2004; Logg et al., 2019; Weyer et al., 2018). Moreover, Parasuraman and Manzey (2010) note that errors of commission are lower when a system serves information integration and analysis as compared to providing concrete recommendations for actions.

There also is evidence for algorithmic aversion, that is, individuals becoming less likely to rely on algorithmic predictions after experiencing false predictions (Burton et al., 2019). Experimental studies show that confidence in algorithms is lowered when algorithms make a mistake (Dietvorst et al., 2015). Moreover, humans tend to adjust their predictions more often based on human advice than based on statistical forecasting (Önkal et al., 2009). However, Grgić-Hlača et al. (2019) report that machine advice does affect participants' predictions in the case of criminal recidivism and Araujo et al. (2020) even find evidence for algorithmic *appreciation*, that is, a preference for automated decisions compared to human decisions.

Empirical studies of actual adoptions of algorithmic recommendations and consequences for inequality are scarce and mostly investigate the judicial context. Results show that higher recidivism scores lead to longer sentences but judges also seem to rely less on risk scores over time (Stevenson and Doleac, 2019). If risk scores are transformed into a categorical scale (low, medium, and high risk), individuals who are placed just above a threshold value receive on average one to four additional weeks of detention before trial compared to those placed just below the threshold (Cowgill, 2018). Again, individual characteristics seem to play an important role as this effect was more pronounced for black defendants than for white defendants.

Social sciences add domain-specific knowledge and tools for understanding macro-level outcomes of human-ADM interactions. Social sciences contribute to developing fair ADM systems by bringing in their domain-specific expertise on individual behavior and social practices across social environments. A thorough analysis of ADM impacts requires such domain-specific knowledge, for example, on labor market behavior. For instance, social sciences can help to answer questions such as: how will an individual adjust her behavior when an employment agency employee decides for a specific (or no) training program based on an ADM recommendation? Could this decrease motivation as an individual feels more constrained by algorithmic decisions than by human decisions? First research documents how individuals evaluate algorithmic decisions compared to human decisions, finding both similarities and differences (Araujo et al., 2020; Binns et al., 2018; Plane et al., 2017; see section “Data preparation and analysis—from fairness in algorithmic output to fairness in social impact”). Due to potential context-dependency, more

research is needed to gain a better understanding of human interpretations of algorithmic decisions.

Social sciences can help to predict and to assess outcomes of ADM processes by providing domain-specific knowledge in the fields of the ADM application (Bertelsmann Stiftung, 2020). This includes knowledge on which goals human decision-makers may follow, which factors they consider, and how these differ from the purely ADM process (see Kleinberg et al., 2018). This also entails how characteristics of the individual and its environment shape the severity of the impact of a decision derived by an ADM (see Abebe et al., 2020). Moreover, as institutional and organizational contexts may react to the implementation of ADM systems in (unintended) ways (Selbst et al., 2019), social sciences also provide methods and previous research to understand established practices in specific contexts and anticipate potential reactions. These methods can also be used to investigate established practices that shape ADM implementation. In the case of comparing algorithmic and human decisions, understanding the goals programmed into an algorithm and analyzing the goals human deciders consider when taking a decision is crucial (Kleinberg et al. 2018; Stevenson 2018).

Additional to in-depth case studies that investigate ADM in concrete contexts (e.g. Elish and Watkins, 2020), experimental research and observational studies along the lines of the research presented in this section improve our understanding of interactions within ADM systems. To show whether and how algorithmic literacy and subsequent behavior impact social inequality, we need to study how these competencies, awareness (Gran et al., 2021), and knowledge related to algorithms are distributed across social groups—for example, by age and education (Fischer and Petersen, 2018)—, and then how this knowledge translates into behavior (e.g. adjusting to the algorithm's “preferences,” see Freeman Engstrom et al. 2020).

Furthermore, social scientists can contribute to investigating how individual decisions made by ADM systems influence inequality and discrimination on the societal macro-level, that is, how single decisions accumulate to overall patterns of inequality in a population akin to the micro-macro model of sociological explanation (Coleman, 1994). Agent-based modeling (ABM) is a promising method to study how interaction on the micro-level produces macro-outcomes as it allows researchers to simulate, for example, interactions of technical and social elements of an ADM process (Gilbert, 2008). ABM could be used to model an interactional setting with three types of agents: affected individuals, algorithms, and deciders. Each affected individual has, for example, certain demographic characteristics and attitudes towards technology. Results of algorithmic predictions based on different fairness strategies can be presented to the decider. The human decider—if applicable—may consider the algorithmic decision and

the affected individual's characteristics to arrive at a decision and weigh both according to her own experience, for example. The affected individual may then adapt her behavior according to her characteristics and the decision.

ABM presents many advantages as it allows researchers to represent the interplay of human and machine actors (see Calero Valdez and Ziefle, 2018) in ADM systems and dynamics over time. For example, fairness implications may only show when considering *long-term* effects on macro-outcomes in the population (Heidari et al., 2019; Liu et al., 2019), and simulations can be run for hundreds or thousands of rounds. Furthermore, ABM responds to calls for a stronger integration of the social environment of ADM systems to grasp their impact appropriately. ABM has already been used to study the governance of socio-technical systems (Adelt et al., 2018), and Cruz Cortés and Ghosh (2019: 3), for example, apply ABM in the context of criminal recidivism risk for a “[s]ystematic analysis [...] [which] implies analyzing the data generating process, the decision-making stage, and its consequences all under the same framework.” In conclusion, simulations are promising tools to assess macro-level outcomes of ADM applications from a social science perspective.

Research avenues:

- How do individuals adapt behavior preemptively or as a reaction towards an algorithmic decision? Which individual resources affect interactional behavior?
- Which situational and individual characteristics determine reliance on ADM systems across social contexts?
- How do these individual decisions and interactions aggregate to macro-social inequality outcomes, and how can researchers study such impacts using simulation techniques?

Conclusion

Synthesizing several theoretical and empirical advances in the research on the consequences of ADM systems for social inequality, this paper provides an overview geared towards social science research, with a focus on data generation, analysis, and implementation challenges. For each part of the ADM pipeline, we highlighted possible inequality issues and how social sciences can contribute to their study. Put briefly, (1) the data used may be biased, (2) the algorithm itself might rely on contextually problematic conceptualizations and formalizations of fairness—or may not consider fairness at all—and (3) the inequality outcomes depend on concrete interactional settings that can result in cumulative disadvantages, particularly for those who have been historically disadvantaged. We summarize potential sources of inequality, related social science topics, example papers, and research avenues in Table S1 (in Supplementary Materials).

Social sciences can draw on established research to contribute to these efforts by bringing in expertise on methods, concrete social contexts, and human (inter)action to investigate how ADM systems affect (macro-)social inequality outcomes. To study algorithmic bias, social scientists can contribute to developing context-aware fairness notions, and to evaluate the scale of actual impacts that ADM systems produce in practice.

Social science research on inequality and ADM systems as well as interactions between algorithms and humans goes far beyond what we were able to cover here (e.g. Joyce et al., 2021; Liu, 2021). Other challenges range from, for example, accounting for the agency of algorithms (Lange et al., 2019), social and political challenges with respect to regulation (Mittelstadt, 2019), privacy (Anthony et al., 2017), and governance (Danaher et al., 2017), to artificial intelligence shifting power relationships (Kalluri, 2020), or other social impacts beyond inequality outcomes. Moreover, we need to understand the contexts in which ADM are applied, including established practices and interactions between human and technical elements. To this end, researchers can draw on a variety of qualitative approaches, such as ethnography (see, e.g. Lange et al., 2019) within the respective social contexts or expert interviews with individuals involved in ADM implementation.

Finally, “impacts” of ADM on social inequality do not necessarily equal to *increases* in disparities. Human decisions are oftentimes also biased and flawed, and algorithmic decisions could potentially display *less* bias than humans (Mayson, 2019) and reduce social inequality overall. However, social implications need to be thought of when designing and implementing ADM applications. We hope that this paper will assist in the development of a research framework and that it will help to enhance concrete guidelines for creating socially responsible ADM systems. Such guidelines are currently discussed and urgently needed as the supervision, assessment, and even necessity of approval of ADM is an ongoing policy debate (e.g. AlgorithmWatch, 2019).

Acknowledgments

We thank the participants of the doctoral colloquium of the Center of Doctoral Studies in the Social and Behavioral Sciences (Sociology) at the University of Mannheim, the members of the Kreuter-Keusch research lab, as well as the anonymous reviewers for their helpful feedback and comments on the paper. We acknowledge funding from the VolkswagenStiftung for the project “Consequences of Artificial Intelligence for Urban Societies” (CAIUS) and the Baden-Württemberg Stiftung for the project “Fairness in Automated Decision making” (FairADM). The publication of this article was funded by the Mannheim Centre for European Social Research (MZES). This work was supported by the University of Mannheim’s Graduate School of Economic and Social Sciences. FG led the development of the paper, RB and CK contributed to research and writing, FK

conceptualized the underlying research projects and contributed to writing.

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 work was supported by the Baden-Württemberg Stiftung, Volkswagen Foundation; The publication of the article was funded by Mannheim Centre for European Social Research (MZES).

ORCID iDs

Frederic Gerdon  <https://orcid.org/0000-0003-4442-6698>

Ruben L Bach  <https://orcid.org/0000-0001-5690-2829>

Christoph Kern  <https://orcid.org/0000-0001-7363-4299>

Frauke Kreuter  <https://orcid.org/0000-0002-7339-2645>

Supplemental material

Supplemental material for this article is available online.

References

- Abebe R, Kleinberg J and Weinberg SM (2020) Subsidy allocations in the presence of income shocks. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(5): 7032–7039.
- Adelt F, Weyer J, Hoffmann S, et al. (2018) Simulation of the governance of complex systems (SimCo): basic concepts and experiments on urban transportation. *Journal of Artificial Societies and Social Simulation* 21: 2.
- Aghaei S, Azizi MJ and Vayanos P (2019) *Learning Optimal and Fair Decision Trees for Non-Discriminative Decision-Making*. Available at: <https://arxiv.org/abs/1903.10598> (accessed 10 May 2021).
- AlgorithmWatch (2019) *Atlas of Automation. Automated Decision-Making and Participation in Germany*. Available at: <https://atlas.algorithmwatch.org/en> (accessed 10 May 2021).
- Allhutter D, Mager A, Cech F, et al. (2020) *Der AMS Algorithmus. Eine Soziotechnische Analyse des Arbeitsmarkchancen-Assistenz-Systems (AMAS)*. Available at: <https://dx.doi.org/10.1553/ITA-pb-2020-02> (accessed 18 February 2022).
- Anthony D, Campos-Castillo C and Horne C (2017) Toward a sociology of privacy. *Annual Review of Sociology* 43(1): 249–269.
- Araujo T, Helberger N, Kruike-meier S, et al. (2020) In AI we trust? Perceptions about automated decision-making by artificial intelligence. *AI & SOCIETY* 35: 611–623.
- Baker R, Brick JM, Bates NA, et al. (2013) Summary report of the AAPOR task force on non-probability sampling. *Journal of Survey Statistics and Methodology* 1(2): 90–143.
- Bansak K (2019) Can nonexperts really emulate statistical learning methods? A comment on “the accuracy, fairness, and limits of predicting recidivism”. *Political Analysis* 27(3): 370–380.
- Barocas S and Selbst AD (2016) Big data’s disparate impact. *California Law Review* 104(3): 671–732.
- Bartlett R, Morse A, Stanton R, et al. (2019) *Consumer-Lending Discrimination in the FinTech Era*. Available at: <https://faculty.haas.berkeley.edu/morse/research/papers/discrim.pdf> (accessed 10 May 2021).
- Bellamy RKE, Mojsilovic A, Nagar S, et al. (2019) AI Fairness 360: an extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development* 63(4/5): 4:1–4:15.
- Benthall S and Haynes BD (2019) Racial Categories in Machine Learning. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 29–31 January 2019, pp. 289–298. New York: Association for Computing Machinery.
- Berk R (2017) An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology* 13(2): 193–216.
- Berk R, Heidari H, Jabbari S, et al. (2017) *A Convex Framework for Fair Regression*. Available at: <https://arxiv.org/abs/1706.02409> (accessed 10 May 2021).
- Berk R, Heidari H, Jabbari S, et al. (2018) Fairness in criminal justice risk assessments. *Sociological Methods & Research* 104(6): 1–42.
- Bertelsmann Stiftung (2020) *Praxisleitfaden zu den Algo.Rules. Orientierungshilfen für Entwickler:innen und ihre Führungskräfte*. Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/alg/Algo.Rules_Praxisleitfaden.pdf (accessed 10 May 2021).
- Binns R, van Kleek M, Veale M, et al. (2018) It’s Reducing a Human Being to a Percentage. In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*, Montreal QC, Canada, 21–26 April 2018, pp. 1–14. New York: Association for Computing Machinery.
- Boeschoten L, van Kesteren E-J, Bagheri A, et al. (2020) *Fair Inference on Error-Prone Outcomes*. Available at: <https://arxiv.org/abs/2003.07621> (accessed 10 May 2021).
- Burton JW, Stein M-K and Jensen TB (2019) A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 27(11): 1309.
- Calero Valdez A and Zieffle M (2018) Human factors in the age of algorithms. Understanding the human-in-the-loop using agent-based modeling. In: *Social Computing and Social Media. Technologies and Analytics: 10th International Conference, SCSSM 2018, Held as Part of HCI International 2018, Proceedings, Part II*, Las Vegas, NV, USA, 15–20 July 2018, pp. 357–371. Cham: Springer International Publishing.
- Cepera KP, Konrad J and Weyer J (2018) Trust in algorithms. An empirical study of users’ Willingness to change behaviour. In: *Critical Issues in Science, Technology and Society Studies: Conference proceedings of the 17th STS Conference Graz 2018*, Graz, Austria, 7–8 May 2018, pp. 38–47. Graz: Verlag der Technischen Universität Graz.
- Coleman JS (1994) *Foundations of Social Theory*. Cambridge: Belknap Press of Harvard University Press.
- Corbett-Davies S and Goel S (2018) *The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning*. Available at: <https://arxiv.org/abs/1808.00023> (accessed 10 May 2021).
- Cowgill B (2018) *The Impact of Algorithms on Judicial Discretion: Evidence from regression discontinuities*. Available at:

- www.columbia.edu/~bc2656/papers/RecidAlgo.pdf (accessed 10 May 2021).
- Cowgill B and Tucker CE (2017) *Algorithmic bias: A counterfactual perspective*. Available at: <https://bitlab.cas.msu.edu/trustworthy-algorithms/whitepapers/Bo%20Cowgill.pdf> (accessed 10 May 2022).
- Cowgill B and Tucker CE (2020) *Algorithmic fairness and economics*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3361280 (accessed 10 May 2021).
- Cruz Cortés E and Ghosh D (2019) *A Simulation based dynamic evaluation framework for system-wide Algorithmic Fairness*. Available at: <https://arxiv.org/abs/1903.09209> (accessed 10 May 2021).
- Danaher J, Hogan MJ, Noone C, et al. (2017) Algorithmic governance: developing a research agenda through the power of collective intelligence. *Big Data & Society* 4(2): 1–21.
- Daumé III H (2017) *A Course in Machine Learning*. Available at: <http://ciml.info/> (accessed 10 May 2021).
- 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.
- Dodge J, Liao QV, Zhang Y, et al. (2019) Explaining models: An empirical study of how explanations impact fairness judgment. In: *Proceedings of the 24th International Conference on Intelligent User Interfaces - IUI '19 the 24th International Conference*, Marina del Ray, California, USA, 17–20 March 2019, pp. 275–285. New York: Association for Computing Machinery.
- Dressel J and Farid H (2018) The accuracy, fairness, and limits of predicting recidivism. *Science Advances* 4(1): eaao5580.
- Dwork C, Hardt M, Pitassi T, et al. (2012) Fairness through Awareness. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*, Cambridge, MA, USA, 8–10 January 2012, pp. 214–226. New York: Association for Computing Machinery.
- Elish MC and Watkins EA (2020) *Repairing Innovation: A Study of Integrating AI in Clinical Care*. Available at: <https://datasociety.net/pubs/repairing-innovation.pdf> (accessed 16 February 2022).
- Elliott MR and Valliant R (2017) Inference for nonprobability samples. *Statistical Science* 32(2): 249–264.
- European Parliament, Directorate General for Parliamentary Research Services, Castelluccia C and Le Métayer D (2019) *Understanding Algorithmic Decision-Making: Opportunities and Challenges*. Luxembourg: Publications Office. Available at: <https://data.europa.eu/doi/10.2861/536131> (accessed 8 February 2022).
- Feldman M, Friedler SA, Moeller J, et al. (2015) Certifying and removing disparate impact. In: *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Sydney, NSW, Australia, 10–13 August 2015, pp. 259–268. New York: Association for Computing Machinery.
- Fischer S and Petersen T (2018) *Was Deutschland über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage*. Available at: https://www.bertelsmann-stiftung.de/fileadmin/files/BSSt/Publikationen/GrauePublikationen/Was_die_Deutschen_ueber_Algorithmen_denken.pdf (accessed 10 May 2021).
- Freeman Engstrom D, Ho DE, Sharkey CM, et al. (2020) *Government by Algorithm: Artificial Intelligence in Federal Administrative Agencies*. Available at: <https://www.acus.gov/sites/default/files/documents/Government%20by%20Algorithm.pdf> (accessed 16 February 2022).
- Friedler SA, Scheidegger C, Venkatasubramanian S, et al. (2018) *A Comparative Study of Fairness-Enhancing Interventions in Machine Learning*. Available at: <https://arxiv.org/abs/1802.04422> (accessed 10 May 2021).
- Gamper J, Kernbeiß G and Wagner-Pinte M (2020) *Das Assistenzsystem AMAS. Zweck, Grundlagen, Anwendung*. Available at: https://www.ams-forschungsnetzwerk.at/download/pub/2020_Assistenzsystem_AMAS-dokumentation.pdf (accessed 18 February 2022).
- Gilbert GN (2008) *Agent-Based Models*. Los Angeles: Sage.
- Goddard K, Roudsari A and Wyatt JC (2012) Automation bias: A systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19(1): 121–127.
- Gran A-B, Booth P and Bucher T (2021) To be or not to be algorithm aware: A question of a new digital divide? *Information, Communication & Society* 24(12): 1779–1796.
- Green B and Chen Y (2019) Disparate interactions. An algorithm-in-the-loop analysis of fairness in risk assessments. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 29–31 January 2019, pp. 90–99. New York: Association for Computing Machinery.
- Green B and Viljoen S (2020) Algorithmic Realism: Expanding the Boundaries of Algorithmic Thought. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency - FAT* '20*, Barcelona, Spain, 27–30 January 2020, pp. 19–31. New York: Association for Computing Machinery.
- Grgić-Hlača N, Engel C and Gummadi KP (2019) Human decision making with machine assistance. An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3(CSCW): 1–15.
- Grgić-Hlača N, Redmiles EM, Gummadi KP, et al. (2018) Human perceptions of fairness in algorithmic decision making. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*, Lyon, France, 23–27 April 2018, pp. 903–912. New York: Association for Computing Machinery.
- Grgić-Hlača N, Weller A and Redmiles EM (2020) *Dimensions of Diversity in Human Perceptions of Algorithmic Fairness*. Available at: <https://arxiv.org/abs/2005.00808> (accessed 10 May 2021).
- Groves RM (2004) *Survey Errors and Survey Costs*. New York: Wiley.
- Hardt M, Price E and Srebro N (2016) Equality of opportunity in supervised learning. In: *Advances in Neural Information Processing Systems 29*, Barcelona, Spain, 5–10 December 2016, 3315–3323. Curran Associates, Inc.
- Hargittai E and Hsieh YP (2013) Digital Inequality. In: Dutton WH (ed) *The Oxford Handbook of Internet Studies*. Oxford: Oxford University Press, pp. 129–150.
- Hebert-Johnson U, Kim MP, Reingold O, et al. (2018) Multicalibration: Calibration for the (Computationally-Identifiable) Masses. In: *Proceedings of the 35th International Conference on*

- Machine Learning, ICML 2018*, Stockholm, Sweden, 10-15 July 2018. PMLR.
- Heidari H, Nanda V and Gummedi KP (2019) On the long-term impact of algorithmic decision policies: Effort unfairness and feature, segregation through social learning. In: *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, 9–15 June 2019, pp. 2692–2701. PMLR.
- Holland PW (1986) Statistics and causal inference. *Journal of the American Statistical Association* 81(396): 945–960.
- Jacobs AZ and Wallach H (2021) Measurement and fairness. In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Canada, 3-10 March 2021, pp. 375–385. New York: Association for Computing Machinery.
- Japac L, Kreuter F, Berg M, et al. (2015) Big data in survey research. *Public Opinion Quarterly* 79(4): 839–880.
- Johndrow JE and Lum K (2017) *An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction*. Available at: <https://arxiv.org/abs/1703.04957> (accessed 10 May 2021).
- Joyce K, Smith-Doerr L, Alegria S, et al. (2021) Toward a sociology of artificial intelligence: A call for research on inequalities and structural change. *Socius: Sociological Research for a Dynamic World* 7: 1–11.
- Kalluri P (2020) Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature* 583(7815): 69.
- Kasy M and Abebe R (2021) Fairness, equality, and power in algorithmic decision-making. In: *FACCT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Canada, pp. 576–586. New York: Association for Computing Machinery.
- Keusch F, Bähr S, Haas G-C, et al. (2020) Coverage error in data collection combining mobile surveys with passive measurement using apps: data from a German national survey. *Sociological Methods & Research*: 0049124120914924.
- Kim MP, Ghorbani A and Zou J (2019) Multiaccuracy: Black-box post-processing for fairness in classification. In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, HI, USA, 27–28 January 2019, pp. 247–254. New York: Association for Computing Machinery.
- Kim MP, Kern C, Goldwasser S, et al. (2022) Universal adaptability: target-independent inference that competes with propensity scoring. *Proceedings of the National Academy of Sciences* 119(4): e2108097119.
- Kleinberg J, Lakkaraju H, Leskovec J, et al. (2018) Human decisions and machine predictions. *The Quarterly Journal of Economics* 133(1): 237–293.
- Kohler-Hausmann I (2011) *Discrimination*. Available at: <https://www.oxfordbibliographies.com/view/document/obo-9780199756384/obo-9780199756384-0013.xml> (accessed 10 May 2021). <https://doi.org/10.1093/OBO/9780199756384-0013>
- Kopf J (2019) *Ein kritischer Blick auf die AMS-Kritiker*. Available at: <https://www.derstandard.de/story/2000109032448/ein-kritischer-blick-auf-die-ams-kritiker> (accessed 10 May 2021).
- Kuppler M, Kern C, Bach RL, et al. (2021) *Distributive Justice and Fairness Metrics in Automated Decision-making: How Much Overlap Is There?* Available at: <https://arxiv.org/abs/2105.01441> (accessed 1 December 2021).
- Kusner M, Russell C, Loftus J, et al. (2019) Making decisions that reduce discriminatory impacts. In: *Proceedings of the 36th International Conference on Machine Learning*, Long Beach, CA, USA, 9-15 June 2019, pp. 3591–3600. PMLR.
- Kusner MJ and Loftus JR (2020) The long road to fairer algorithms. *Nature* 578(7793): 34–36.
- Lange A-C, Lenglet M and Seyfert R (2019) On studying algorithms ethnographically: making sense of objects of ignorance. *Organization* 26(4): 598–617.
- Lee JD and See KA (2004) Trust in automation: designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 46(1): 50–80.
- Lerman J (2013) Big data and its exclusions. *Stanford Law Review Online* 66: 55–63.
- Lind K and Wallentin L (2020) *Central Authorities Slow to React as Sweden's Cities Embrace Automation of Welfare Management*. Available at: <https://algorithmwatch.org/en/story/trelleborg-sweden-algorithm/> (accessed 10 May 2021).
- Liu LT, Dean S, Rolf E, et al. (2019) *Delayed Impact of Fair Machine Learning*. Available at: <https://arxiv.org/abs/1803.04383> (accessed 10 May 2021).
- Liu Z (2021) Sociological perspectives on artificial intelligence: A typological reading. *Sociology Compass* 15(3): 1–13.
- Logg JM, Minson JA and Moore DA (2019) Algorithm appreciation: people prefer algorithmic to human judgment. *Organizational Behavior and Human Decision Processes* 151: 90–103.
- Lopez J (2019) Reinforcing intersectional inequality via the AMS algorithm in Austria. In: *Conference Proceedings of the 18th STS Conference Graz 2019: Critical Issues in Science, Technology and Society Studies*, Graz, Austria, 6-7 May 2019, pp. 289–309. Graz: Verlag der Technischen Universität Graz.
- Lutz C (2019) Digital inequalities in the age of artificial intelligence and big data. *Human Behavior and Emerging Technologies* 1(2): 141–148.
- Makhlouf K, Zhioua S and Palamidessi C (2020) *On The Applicability of ML Fairness Notions*. Available at: <https://arxiv.org/abs/2006.16745> (accessed 10 May 2021).
- Mann M and Matzner T (2019) Challenging algorithmic profiling: The limits of data protection and anti-discrimination in responding to emergent discrimination. *Big Data & Society* 6(2): 1–11.
- Mayson SG (2019) Bias in, bias out. *The Yale Law Journal* 128(8): 2218–2300.
- Mehrabi N, Morstatter F, Saxena N, et al. (2019) *A Survey on Bias and Fairness in Machine Learning*. Available at: <https://arxiv.org/abs/1908.09635> (accessed 11 May 2021).
- Miller AP (2018) *Want Less-Biased Decisions? Use Algorithms*. Available at: <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms> (accessed 10 May 2021).
- Mittelstadt BD (2019) Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1(11): 501–507.
- Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: mapping the debate. *Big Data & Society* 3(2): 1–21.
- Molnar C (2019) *Interpretable Machine Learning. A Guide for Making Black Box Models Explainable*. <https://christophm.github.io/interpretable-ml-book/> (accessed 2 December 2021).
- Noriega-Campero A, Bakker MA, Garcia-Bulle B, et al. (2018) *Active Fairness in Algorithmic Decision Making*. Available at: <https://arxiv.org/abs/1810.00031> (accessed 10 May 2021).
- Obermeyer Z, Powers B, Vogeli C, et al. (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464): 447–453.

- Otte G, Boehle M, Kunißen K, et al. (2021) Social inequalities – empirical focus. In: Hollstein B, Greshoff R and Schimank U (eds) *Soziologie – Sociology in the German-Speaking World*. Berlin, Boston: De Gruyter Oldenbourg, pp.361–380.
- Önkal D, Goodwin P, Thomson M, et al. (2009) The relative influence of advice from human experts and statistical methods on forecast adjustments. *Journal of Behavioral Decision Making* 22(4): 390–409.
- Parasuraman R and Manzey DH (2010) Complacency and bias in human use of automation: an attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 52(3): 381–410.
- Pierson E (2018) *Demographics and Discussion Influence Views on Algorithmic Fairness*. Available at: <https://arxiv.org/abs/1712.09124> (accessed 10 May 2021).
- Plane AC, Redmiles EM and Mazurek ML (2017) Exploring User Perceptions of Discrimination in Online Targeted Advertising. In: *Proceedings of the 26th USENIX Security Symposium*, Vancouver, BC, Canada, 16-18 August 2017, pp. 935–951. Berkeley: USENIX Association.
- Rodolfa K, Saleiro P and Ghani R (2021) Bias and fairness. In: Foster I, Ghani R and Jarmin RS, et al., (eds) *Big Data and Social Science. Data Science Methods and Tools for Research and Practice*. Boca Raton, FL: CRC Press, pp.281–312.
- Rodolfa KT, Salomon E, Haynes L, et al. (2020) Case Study: Predictive Fairness to Reduce Misdemeanor Recidivism through Social Service Interventions. In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency - FAT* '20*, Barcelona, Spain, 27–30 January 2020, pp. 142–153. New York: Association for Computing Machinery.
- Saleiro P, Kuester B, Hinkson L, et al. (2019) *Aequitas: A bias and fairness audit toolkit*. Available at: <https://arxiv.org/abs/1811.05577> (accessed 10 May 2021).
- Seaver N (2019) Knowing algorithms. In: Vertesi J and Ribes D (eds) *DigitalSTS: A Field Guide for Science & Technology Studies*. Princeton: Princeton University Press, pp.412–422.
- Selbst AD, boyd D, Friedler SA, et al. (2019) Fairness and abstraction in sociotechnical systems. In: *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 29-31 January 2019. New York: Association for Computing Machinery.
- Sen I, Floeck F, Weller K, et al. (2019) *A Total Error Framework for Digital Traces of Humans*. Available at: <https://arxiv.org/abs/1907.08228> (accessed 10 May 2021).
- Starke C, Baleis J, Keller B, et al. (2021) *Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature*. <https://arxiv.org/abs/2103.12016> (accessed 10 May 2021).
- Stevenson MT (2018) Assessing risk assessment in action. *Minnesota Law Review* 103(1): 303–384.
- Stevenson MT and Doleac JL (2019) *Algorithmic Risk Assessment in the Hands of Humans*. Available at: <http://ftp.iza.org/dp12853.pdf> (accessed 10 May 2021).
- Suresh H and Guttag JV (2020) *A Framework for Understanding Unintended Consequences of Machine Learning*. Available at: <https://arxiv.org/abs/1901.10002> (accessed 10 May 2021).
- Tan S, Adebayo J, Inkpen K, et al. (2018) *Investigating Human + Machine Complementarity: A Case Study on Recidivism*. Available at: <https://arxiv.org/abs/1808.09123> (accessed 10 May 2021).
- van Doorn M, Pop I and Wolbers MHJ (2011) Intergenerational transmission of education across European countries and cohorts. *European Societies* 13(1): 93–117.
- Wachter S (2020) Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Review* 35(2): 1–74.
- Weyer J, Delisle M, Kappler K, et al. (2018) Big data in soziologischer perspektive. In: Kolany-Raiser B, Heil R, Orwat C and Hoeren T (eds) *Big Data und Gesellschaft: Eine Multidisziplinäre Annäherung*. Wiesbaden: Springer VS, pp.69–149.
- Wickens CD, Clegg BA, Vieane AZ, et al. (2015) Complacency and automation bias in the use of imperfect automation. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 57(5): 728–739.
- Yang S and Kim JK (2020) *Statistical Data Integration in Survey Sampling: A Review*. Available at: <https://arxiv.org/abs/2001.03259> (accessed 10 May 2021).
- Zerilli J, Knott A, Maclaurin J, et al. (2019) Algorithmic decision-making and the control problem. *Minds and Machines* 29(4): 555–578.
- Zou J and Schiebinger L (2018) AI Can be sexist and racist – it’s time to make it fair. *Nature* 559(7714): 324–326.