Responsible Data Management

Julia Stoyanovich
New York Univerisity
New York, NY, USA
stoyanovich@nyu.edu

Bill Howe
Univerisity of Washington
Seattle, WA, USA
billhowe@uw.edu

H.V. Jagadish
Univerisity of Michigan
Ann Arbor, MI, USA
jag@umich.edu

ABSTRACT

The need for responsible data management intensifies with the growing impact of data on society. One central locus of the societal impact of data are Automated Decision Systems (ADS), socio-legal-technical systems that are used broadly in industry, non-profits, and government. ADS process data about people, help make decisions that are consequential to people's lives, are designed with the stated goals of improving efficiency and promoting equitable access to opportunity, involve a combination of human and automated decision making, and are subject to auditing for legal compliance and to public disclosure. They may or may not use AI, and may or may not operate with a high degree of autonomy, but they rely heavily on data.

In this article, we argue that the data management community is uniquely positioned to lead the responsible design, development, use, and oversight of ADS. We outline a technical research agenda that requires that we step outside our comfort zone of engineering for efficiency and accuracy, to also incorporate reasoning about values and beliefs. This seems high-risk, but one of the upsides is being able to explain to our children what we do and why it matters.

PVLDB Reference Format:
Julia Stoyanovich, Bill Howe, H.V. Jagadish. Responsible Data Management. *PVLDB*, 13(12): 3474 - 3488, 2020.
DOI: https://doi.org/10.14778/3415478.3415570

1. INTRODUCTION

We are in the midst of a global trend to regulate algorithms, artificial intelligence, and automated decision systems. This flurry of activity hardly comes as a surprise. As reported by the recent One Hundred Year Study on Artificial Intelligence [58]: "AI technologies already pervade our lives. As they become a central force in society, the field is shifting from simply building systems that are intelligent to building intelligent systems that are human-aware and trustworthy." In the European Union, the General Data Protection Regulation (GDPR) [66] offers protections to individuals regarding the collection, processing, and movement of their personal data, and applies broadly to the use of such data by governments and private-sector entities. Regulatory activity in several countries outside of the EU, notably, Japan [48] and Brazil [32], is in close alignment with the GDPR.

In the US, many major cities, a handful of states, and even the Federal government are establishing task forces and issuing guidelines about responsible development and use of technology, often starting with its use in government itself——rather than in the private sector——where there is, at least in theory, less friction between organizational goals and societal values. Case in point: New York City rightfully prides itself on being a trendsetter—in architecture, fashion, the performing arts and, as of late, in its very publicly made commitment to opening the black box of the government's use of technology: In May 2018, an Automated Decision Systems (ADS) Task Force was convened, the first such in the nation, and charged with providing recommendations to New York City's agencies about becoming transparent and accountable in their use of ADS. The Task Force issued its report in November 2019, making a commitment to using ADS *where* they are beneficial, reducing potential harm across their lifespan, and promoting fairness, equity, accountability, and transparency in their use [5].

Can the principles of the responsible use of ADS — of *socio-legal-technical systems* that may or may not use AI, and may or may not operate with a high degree of autonomy, but that rely heavily on data — be operationalized as a matter of policy [2]? Can this be done in the face of a crisis of trust in government, which extends to the lack of trust in the government's ability to manage modern technology in the interest of the public [73]? What will it take to instill responsible ADS practices beyond government?

In this article, we hope to convince you that the data management community should play a central role in the responsible design, development, use, and oversight of ADS. By engaging in this work, we have a critical opportunity to help make society more equitable, inclusive, and just; make government operations more transparent and accountable; and encourage public participation in ADS design and oversight. To make progress, we may need to step outside our engineering comfort zone and start reasoning in terms of values and beliefs, in addition to checking results against known ground truths and optimizing for efficiency objectives. This seems high-risk, but one of the upsides is being able to explain to our children what we do and why it matters.

Proceedings of the VLDB Endowment, Vol. 13, No. 12
ISSN 2150-8097.
DOI: https://doi.org/10.14778/3415478.3415570

Outline. In the remainder of this paper, we will first illustrate the issues under discussion with an example from the domain of hiring and employment (Section 2). We will go on to position these issues, and possible solutions, within the broader context of bias, worldviews, and equality of opportunity frameworks (Section 4). Then, we will discuss recent technical work by us and others on embedding responsibility into data lifecycle management (Sections 5) and on interpretability of data and models for a range of stakeholders (Section 6). In the technical sections, we will point out specific opportunities for contributions by the data management community. We will conclude in Section 7.

2. AUTOMATED HIRING SYSTEMS

To make our discussion concrete, let us focus on hiring and employment. Since the 1990s, and increasingly so in the last decade, commercial tools are being used by companies large and small to hire more efficiently: source and screen candidates faster and with less paperwork, and successfully select candidates who will perform well on the job. These tools are also meant to improve efficiency for the job applicants, matching them with relevant positions, allowing them to apply with a click of a button, and facilitating the interview process. According to Jenny Yang, former Commissioner of the US Equal Employment Opportunity Commission (EEOC), "Automated hiring systems act as modern gatekeepers to economic opportunity. [...] Across industries, major employers including Unilever, Hilton, and Delta Air Lines are using data-driven, predictive hiring tools." [68]

The hiring funnel. Bogen and Rieke [9] describe the hiring process from the point of view of an employer as a series of decisions that form a funnel (Figure 1): "Employers start by *sourcing* candidates, attracting potential candidates to apply for open positions through advertisements, job postings, and individual outreach. Next, during the *screening* stage, employers assess candidates——both before and after those candidates apply——by analyzing their experience, skills, and characteristics. Through *interviewing* applicants, employers continue their assessment in a more direct, individualized fashion. During the *selection* step, employers make final hiring and compensation determinations."

The hiring funnel is an example of an ADS: a socio-legal-technical system operationalized as a sequence of data-driven, algorithm-assisted steps, in which a series of decisions culminates in job offers to some candidates and rejections to others. While potentially beneficial, the use of ADS in hiring is also raising concerns that pertain, broadly speaking, to the decisions made by these systems and to the process by which these decisions are made.

Discrimination. One set of concerns relates to *discrimination*. As pointed out by Bogen and Rieke [9], "The hiring process starts well before anyone submits an actual job application, and jobseekers can be disadvantaged or rejected at any stage. Importantly, while new hiring tools rarely make affirmative hiring decisions, they often automate rejections."

Because of how impactful hiring decisions are for individuals and population groups, and because of a history of discrimination, hiring practices are subject to antidiscrimination laws in many countries. In the US, Title VII of the Civil Rights Act of 1964 broadly prohibits hiring discrimination by employers and employment agencies on the basis of *protected characteristics* that include "race, color, religion, sex, and national origin." This law is supplemented by other federal laws that extend similar protections based on age and disability status, and by a patchwork of other federal, state, and local laws.

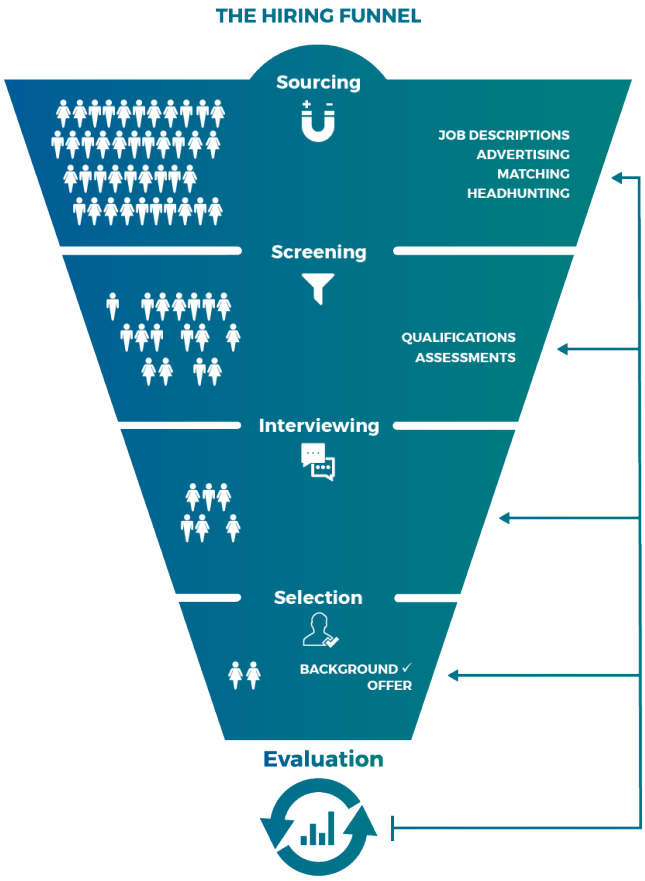

Figure 1: The hiring funnel, reproduced with permission from Bogen and Rieke [9], is an example of an Automated Decision System (ADS): a data-driven, algorithm-assisted process in which a series of decisions culminates in job offers to some applicants and rejections to others.

Are existing legal protections against discrimination sufficient today, when ADS are reshaping, streamlining, and scaling up hiring? Or is the use of ADS reviving and reinforcing historical discrimination, and giving rise to new forms of discrimination? Is discrimination going undetected, due, for example, to legal constraints on the types of demographic data that a potential employer can collected, or to applicants declining to disclose their demographic group membership? Can attempts to de-bias datasets and models be effective, or do they amount to *fairwashing*—covering up, and even legitimizing, discrimination with the help of technological solutions?

Due process. Another set of concerns relates to *due process*, also known as *procedural fairness* or *procedural regularity*. As explained by Kroll *et al.* [34]: "A baseline requirement in most contexts is procedural regularity: each participant will know that the same procedure was applied to her and that

the procedure was not designed in a way that disadvantages her. This baseline requirement draws on the Fourteenth Amendment [to the US Constitution] principle of procedural due process. Ever since a seminal nineteenth century case, the [US] Supreme Court has articulated that procedural fairness or due process requires rules to be generally applicable and not designed for individual cases."

Notably, research demonstrates that, as long as a process is seen as fair, people will accept outcomes that may not benefit them. This finding is supported in numerous domains, including hiring and employment, legal dispute resolution and citizen reactions to police and political leaders, and it remains relevant when decisions are made with the assistance of algorithms [63].

Citron and Pasquale [12] discuss the need for due process safeguards in scoring systems: "The act of designating someone as a likely credit risk (or bad hire, or reckless driver) raises the cost of future financing (or work, or insurance rates), increasing the likelihood of eventual insolvency or un-employability. When scoring systems have the potential to take a life of their own, contributing to or creating the situation they claim merely to predict, it becomes a normative matter, requiring moral justification and rationale." Score-based selection and ranking are indeed in broad use at all stages of the hiring funnel, and can amount to self-fulfilling prophecy if left unchecked.

An immediate interpretation of due process for the hiring ADS is that the employer ought to be able to show that the same decision making procedure was used for all job candidates. Yet, simply stating that the same code was executed for everyone does not get to the heart of the issue, precisely because individuals and population groups may be represented differently in the data. For example, groups that are historically under-represented in the workforce will also be under-represented in the data record, which may in turn reduce generalizability of predictive models for those groups [11]. Further, values of a particular feature may be missing more frequently for one sub-population than for another (e.g., age may be unspecified for women more frequently than for men), also leading to disparate predictive accuracy. Finally, it has been documented that survey data can be noisier for minority groups than for others [28]. (Lehr and Ohm [36] give additional examples of the impact of data on discrimination and due process in machine learning.)

Feature selection. An important dimension of due process, closely linked to discrimination, is substantiating the use of particular features in decision-making. Regarding the use of predictive analytics to screen candidates, Yang [68] states: "Algorithmic screens do not fit neatly within our existing laws because algorithmic models aim to identify statistical relationships among variables in the data whether or not they are understood or job related.[...] Although algorithms can uncover job-related characteristics with strong predictive power, they can also identify correlations arising from statistical noise or undetected bias in the training data. Many of these models do not attempt to establish cause-and-effect relationships, creating a risk that employers may hire based on arbitrary and potentially biased correlations." That is, identifying features that are impacting a decision is important, but it is insufficient to alleviate due process and discrimination concerns. The employer should also show that these features are relevant for performance on the job.

An extreme case of feature selection gone wrong is when tools claim to predict job performance by analyzing an interview video for body language and speech patterns. In his recent talk, Arvind Narayanan refers to tools of this kind as "fundamentally dubious" and places them in the category of *AI snake oil* [44]. The premise of such tools, that (a) it is possible to predict social outcomes based on a person's appearance or demeanor and (b) it is ethically defensible to try, reeks of scientific racism and is at best an elaborate random number generator.

Even features that can legitimately be used for hiring may capture information differently for different individuals and groups. For example, it has been documented that the mean score of the math section of the SAT (Scholastic Assessment Test, used broadly in the US) differs across racial groups, as does the shape of the score distribution [50]. These disparities are often attributed to racial and class inequalities encountered early in life, and are thought to present persistent obstacles to upward mobility and opportunity.

Auditing and disclosure. Because of the wide-spread use of commercial ADS in hiring, and because of the discrimination and due process concerns they raise, there is a push to strengthen the accountability structure in this domain. The gist of most proposals is to develop new legal and regulatory mechanisms—and the supporting technical methods—to facilitate auditing of these systems and public disclosure.

For example, Yang [68] advocates that "A federal *explainability standard* that sets forth the parameters for what it means to explain an algorithm to different audiences (such as workers, employers, or technologists) would be valuable to ensure these considerations are built into the design of an algorithmic system from the outset." She also speaks to the importance of *the right to an explanation*—that "employers should explain the rationale for a decision in terms that a reasonable worker could understand. Standards could be established to include disclosure of the material variables considered and the types of inferences the algorithm is making to score the individuals."

As another example, New York City Commission on Technology is entertaining a bill "in relation to the sale of automated employment decision tools" that would require auditing such tools for bias and disclosing to the candidate the job qualifications or characteristics used for assessment [67].

3. WHAT IS AN ADS? AND WHY US?

We have been referring to Automated Decision Systems (ADS) throughout this paper. Yet, there is currently no consensus as to what is, and is not, an ADS. In fact, the need to define this term for the purpose of regulation has been the subject of much debate. As a representative case, Chapter 6 of the NYC ADS Task Force report [5] summarizes their months-long struggle to, somewhat ironically, define their own mandate—come up with a definition that is sufficiently broad to capture the important concerns discussed earlier in this section, yet sufficiently specific to be practically useful.

If an intentional definition is out of reach, we may attempt to define ADS by extension. An automated resume screening tool seems like a natural example of an ADS, as does a tool that matches job applicants with positions in which they are predicted to do well. But is a calculator an ADS? (No!) What about a formula in a spreadsheet? (Depends on what it's used for [23].)

The hiring funnel in Figure 1, *as well as each component of the funnel*, are ADS examples. These systems (1) process data about people, some of which may be sensitive or proprietary; (2) help make decisions that are consequential to people's lives and livelihoods; (3) are designed with the stated goals of improving efficiency and promoting, or at least not hindering, equitable access to opportunity; (4) involve a combination of human and automated decision making; and (5) are subject to auditing for legal compliance and, at least potentially, to public disclosure.

Central to this ADS definition is the placing of technical decision-making components—a spreadsheet formula, a matchmaking algorithm, or a predictive analytic—within *the lifecycle of data collection and analysis*. Much excellent work on algorithmic fairness and transparency goes on in the machine learning, data mining, and algorithms communities. Yet, a critical shortcoming of that work is their focus on the last mile of data analysis. In contrast, and precisely because of the importance of a lifecycle view of ADS, the data management community is uniquely positioned to deliver true practical impact in the responsible design, development, use, and oversight of these systems.

- Because data management technology offers a natural centralized point for enforcing policies, we can develop methodologies to transparently and explicitly enforce requirements through the ADS lifecycle.
- Because of the unique blend of theory and systems in our methodological toolkit, we can help inform regulation by studying the feasible trade-offs between different classes of legal and efficiency requirements.
- Because of our pragmatic approach, we can support compliance by developing standards for effective and efficient auditing and disclosure, and developing protocols for embedding these standards in systems.

Importantly, the ADS lifecycle discussed in this section is itself embedded within the societal context of ADS purpose and impacts. We elaborate on this point in the next section.

4. FRAMING TECHNICAL SOLUTIONS

Before diving into specific research directions, let us step back and think carefully about the role that technological interventions, such as data management solutions, can play in supporting the responsible use of ADS. This discussion is necessary to help us find a pragmatic middle ground between the harmful extremes of techno-optimism—a belief that technology can single-handedly fix deep-seated societal problems like structural discrimination in hiring, and techno-bashing—a belief that any attempt to operationalize ethics and legal compliance in ADS will amount to fairwashing and so should be dismissed outright.

4.1 Data: a Mirror Reflection of the World

We often hear that ADS, such as automated hiring systems, operate on biased data and result in biased outcomes. What is the meaning of the term "bias" in this context? Informally, data is a mirror reflection of the world. More often than not, this reflection is distorted. One reason for this may be that the mirror itself (the measurement process) is distorted: it faithfully reflects some portions of the world, while amplifying or diminishing others. Another reason may be that *even if* the mirror was perfect, it may be reflecting a distorted world — a world such as it is, and not as it could or should be. The mirror metaphor helps us make several simple but important observations, on which we will elaborate more formally (and less poetically) in Section 4.2.

1. *A reflection cannot know whether it is distorted.* Based on the reflection alone, and without knowledge about the properties of the mirror and of the world it reflects, we cannot know whether the reflection is distorted, and, if so, for what reason. That is, data alone cannot tell us whether it is a distorted reflection of a perfect world, a perfect reflection of a distorted world, or whether these distortions compound.
2. *Beauty is in the eye of the (human) beholder.* It is up to people — individuals, groups, and society at large — and not up to data or algorithms, to come to a consensus about whether the world is how it should be, or if it needs to be improved and, if so, how we should go about improving it.
3. *Changing the reflection does not change the world.* If the reflection itself is used to make important decisions, and we agree that it is distorted and explicitly state the assumed or verified nature of such distortions, then compensating for the distortions is worthwhile. But the mirror metaphor only takes us so far. We have to work much harder—usually going far beyond technological solutions—to propagate the changes back into the world, not merely brush up the reflection.

In their seminal 1996 paper, Friedman and Nissenbaum identified three types of bias that can arise in computer systems: pre-existing, technical, and emergent bias [19]. In the remainder of this section we will use this classification to structure our discussion on bias, worldviews, and mitigation strategies.

4.2 Pre-existing Bias

Pre-existing bias exists independently of an algorithm itself and has its origins in society. Often, the presence or absence of pre-existing bias cannot be scientifically verified, but rather is postulated based on a belief system. We already discussed that disparities in math SAT scores have been observed among ethnic groups [50]. If we believed that the test measures an individual's academic potential, we would not consider this an indication of pre-existing bias. If, on the other hand, we believed that standardized test scores are sufficiently impacted by preparation courses that the score itself says more about socio-economic conditions than an individual's academic potential, then we would consider the data to be biased.

Worldviews. Friedler *et al.* [18] reflect on the impossibility of a purely objective interpretation of algorithmic fairness (in the sense of a lack of bias): "In order to make fairness mathematically precise, we tease out the difference between beliefs and mechanisms to make clear what aspects of this debate are opinions and which choices and policies logically follow from those beliefs." They model the decision pipeline of a task as a sequence of mappings between three metric spaces: construct space (CS), observed space (OS), and decision space (DS), and define worldviews (belief systems) as assumptions about the properties of these mappings.

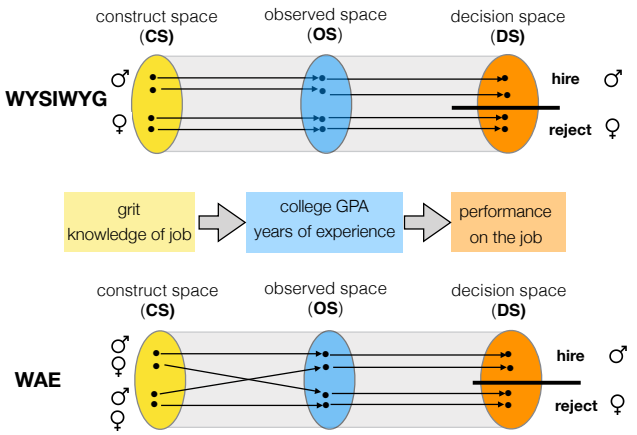

Figure 2: An illustration of worldviews from Frieder et al. [18] for hiring. "What you see is what you get" (WYSIWIG) assumes that the mapping from the construct space (*CS*) to the observed space (*OS*) shows low distortion, while "We are all equal" (WAE) assumes that this mapping shows structural bias, leading to a distortion in group structure.

The spaces and the mappings between them are illustrated in Figure 2 for the hiring ADS. Individuals are represented by points. *CS* represents the "true" properties of an individual (e.g., grit and knowledge of the job for the hiring ADS), *OS* represents the properties that we can measure (e.g., college GPA as a proxy for grit, years of experience as a proxy for knowledge of the job). *OS* is the feature space of a component in the decision pipeline, such as a classifier, a score-based selection procedure, or a human hiring manager. Finally, *DS* is the space of outcomes of that component.

When considering mappings, we are concerned with whether they preserve pair-wise distances between individuals. Importantly, because both *CS* and the mapping from *CS* to *OS* are, by definition, unobservable, a belief about the properties of the mapping has to be postulated. Friedler *et al.* [18] describe two extreme cases: WYSIWYG ("what you see is what you get") assumes low distortion from *CS* to *OS*, while WAE ("we are all equal") assumes the presence of structural bias—a systematic distortion in group structure.

While in general we cannot confirm the presence of pre-existing bias in a dataset, we are sometimes able to use another dataset, or contextual knowledge about the dataset or about the world itself, to corroborate or challenge the claim of pre-existing bias. For example, Lum and Isaac [39] showed that two areas with non-white and low-income populations in Oakland, CA experience 200 times more drug-related arrests than other areas. Yet, based on the 2011 National Survey on Drug Use and Health, the estimated number of drug users is distributed essentially uniformly across Oakland, with variation driven primarily by differences in population density. This information can be combined with our knowledge about policing practices, namely, that low-income neighborhoods are patrolled more frequently than other neighborhoods, and influence our belief about the presence of bias in the drug-related arrests dataset.

If pre-existing bias in a dataset is postulated, perhaps with corroboration from other datasources and with background knowledge about data collection practices, yet we are still interested in using this data in decision-making, then we need to identify an appropriate bias mitigation strategy. The WAE worldview justifies mitigations that enforce equality of outcomes, which are most intuitively operationalized as *statistical parity*, a requirement that the demographics of individuals receiving any outcome (positive or negative, in the case of binary classification) is the same as their demographics in the input. For example, if half of the job applicants are women, then half of those selected for in-person interviews should be women even if they appear less qualified by conventional metrics. (See Mitchell *et al.* [41] for a recent survey of fairness measures, of which statistical parity is an example.) This mitigation is simple to enact, but it's a blunt instrument: it does not tell us which women to select or, more generally, whether and how to look for useful signal in the data under the assumption of pre-existing bias. Next, we look at an alternative framework that brings more nuance into the treatment of pre-existing bias and can help inform the design of mitigation strategies.

Equality of opportunity. Heidari *et al.* [22] show an application of equality of opportunity (EOP) frameworks to algorithmic fairness. EOP emphasizes the importance of personal qualifications, and seeks to minimize the impact of circumstances and arbitrary factors on individual outcomes. "At a high level, in these models an individual's outcome/position is assumed to be affected by two main factors: his/her circumstance c and effort e. Circumstance c is meant to capture all factors that are deemed irrelevant, or for which the individual should not be held morally accountable; for instance c could specify the socio-economic status they were born into. Effort e captures all accountability factors—those that can morally justify inequality." [22]

Several conceptions of EOP have been proposed, differing in what features they consider to be relevant (or morally acceptable to use) and which they deemed irrelevant. So, libertarian EOP allows all features to be used in decision-making, while formal EOP prohibits the use of sensitive features like gender and race but can still use proxy features.

In contrast, substantive EOP, notably, Rawlsian [49] and luck egalitarian [51], seeks to offer equal opportunity in access to positions by providing fair access to the necessary qualifications for the positions. Both conceptions concede that opportunity is only equal relative to one's effort, but they differ in how effort is modeled: Rawlsian EOP asserts that equal effort should imply equal opportunity (represented as a utility distribution), regardless of circumstances. Luck egalitarian EOP considers effort relative to one's demographic group ("type" in their terms): two individuals are considered to have exercised the same level of effort if "they sit at the same quantile or rank of the effort distribution for their corresponding types." [22]

4.3 Technical Bias

Technical bias can be introduced at any stage of the ADS lifecycle, and it may exacerbate pre-existing bias. The bad news is that risks of introducing technical bias stemming from data management components abound. The good news is that, unlike with pre-existing bias, there is no ambiguity about whether a technical fix should be attempted: if technical systems we develop are introducing bias, then we should be able to instrument these systems to measure it and understand its cause. It may then be possible to mitigate this bias and to check whether the mitigation was effective.

Importantly, as we instrument our systems, we must once again take the lifecycle view. The goal is to understand how properties of individual components compose, and whether we can make guarantees about the presence or absence of technical bias in the pipeline overall based on what we know about individual components. In what follows, we discuss potential sources of technical bias in several lifecycle stages that are within our (data management) purview.

Data cleaning. Methods for missing value imputation that are based on incorrect assumptions about whether data is missing at random may distort protected group proportions. Consider a form that gives job applicants a binary choice of gender and also allows to leave gender unspecified. Suppose that about half of the users identify as men and half as women, but that women are more likely to omit gender. Then, if mode imputation (replacing a missing value with the most frequent value for the feature, the default in scikit-learn) is used, then all (predominantly female) unspecified gender values will be set to male. More generally, multi-class classification for missing value imputation typically only uses the most frequent classes as target variables [8], leading to a distortion for small population groups, because membership in these groups will never be imputed.

Next, suppose that some individuals identify as non-binary. Because the system only supports male, female, and unspecified as options, these individuals will leave gender unspecified. If mode imputation is used, then their gender will be set to male. A more sophisticated imputation method will still use values from the active domain of the feature, setting the missing values of gender to either male or female. This example illustrates that bias can arise from an incomplete or incorrect choice of data representation.

Finally, consider a form that has home address as a field. A homeless person will leave this value unspecified, and it is incorrect to attempt to impute it. While dealing with `null` values is known to be difficult and is already considered among the issues in data cleaning, the needs of responsible data management introduce new problems. As we pointed out in Section 2 under *due process*, data quality issues often disproportionately affect members of historically disadvantaged groups, and we risk compounding technical bias due to data representation with bias due to statistical concerns.

Filtering. Selections and joins can arbitrarily change the proportion of protected groups (e.g., female gender) even if they do not directly use the sensitive attribute (e.g., gender) as part of the predicate or of the join key. This change in proportion may be unintended and is important to detect, particularly when this happens during one of many preprocessing steps in the ADS pipelines.

Another potential source of technical bias is the usage of pre-trained word embeddings. For example, a pipeline may replace a textual name feature with the corresponding vector from a word embedding that is missing for rare, non-western names. If we then filter out records for which no embedding was found, we may disproportionately remove individuals from specific ethnic groups.

Ranking. Technical bias can arise when results are presented in ranked order, such as when a hiring manager is considering potential candidates to invite for in-person interviews. The main reason is the inherent position bias — the geometric drop in visibility for items at lower ranks compared to those at higher ranks, which arises because in Western cultures we read from top to bottom, and from left to right, and so items in the top-left corner of the screen attract more attention [7]. A practical implication is that, even if two candidates are equally suitable for the job, only one of them can be placed above the other, suggesting that it should be prioritized. Depending on the needs of the application and on the level of technical sophistication of the decision-maker, this problem can be addressed by suitably randomizing the ranking, showing results with ties, or plotting the score distribution.

4.4 Emergent Bias

Emergent bias arises in a context of use and may be present if a system was designed with different users in mind or when societal concepts shift over time. For ranking and recommendation in e-commerce, emergent bias arises most notably because searchers tend to trust the systems to indeed show them the most suitable items at the top positions [46], which in turn shapes a searcher's idea of a satisfactory answer, leading to a "rich-get-richer" situation.

This example immediately translates to hiring and employment. If hiring managers trust recommendations from an ADS, and if these recommendations systematically prioritize applicants of a particular demographic profile, then a feedback loop will be created, further diminishing workforce diversity over time. Bogen and Rieken [9] illustrate this problem: "For example, an employer, with the help of a third-party vendor, might select a group of employees who meet some definition of success—for instance, those who 'outperformed' their peers on the job. If the employer's performance evaluations were themselves biased, favoring men, then the resulting model might predict that men are more likely to be high performers than women, or make more errors when evaluating women. This is not theoretical: One resume screening company found that its model had identified having the name 'Jared' and playing high school lacrosse as strong signals of success, even though those features clearly had no causal link to job performance."

4.5 Summary

In summary, (1) We must clearly state the beliefs against which we are validating fairness. Technical interventions to improve fairness should be consistent with these beliefs. Beliefs cannot be checked empirically or falsified, as they are not hypotheses; they can only be stated axiomatically. (2) We cannot fully automate responsibility, particularly because many of the concerns we are looking to address are themselves a consequence of automation. We embrace the idea that technical interventions are only part of an over-all mitigation strategy, and should verify that they are even an effective step — there is no guarantee that is the case. (3) We need to broaden the scope of data management research beyond manipulations of properties of either a dataset or an algorithm; ADS are datasets together with algorithms together with contexts of use: the calculator is not discriminatory, but its context of use may be.

5. MANAGING THE ADS LIFECYCLE

As we discussed in Section 3, ADS critically depend on data and so should be seen through the lens of the *data*

lifecycle [26]. Responsibility concerns, and important decision points, arise in data sharing, annotation, acquisition, curation, cleaning, and integration. Several lines of recent work argue that opportunities for improving data quality and representativeness, controlling for bias, and allowing humans to oversee the process, are missed if we do not consider these earlier lifecyle stages [30, 36, 61].

Database systems centralize correctness constraints to simplify application development via schemas, transaction protocols, etc.; algorithmic fairness and interpretability are now emerging as first-class requirements. But unlike research in the machine learning community, we need generalized requirements and generalized solutions that work across a range of applications. In what follows, we give examples of our own recent and ongoing work that is motivated by this need. These examples underscore that tangible technical progress is possible, and also that much work remains to be done to offer systems support for the responsible management of the ADS lifecycle.

5.1 Data Acquisition

Data used for analysis is often originally created for a different purpose, and therefore is frequently not representative of the true distribution. Even if the data is explicitly collected for the purpose of analysis, it can be hard to obtain a representative sample. Consider, for example, a website with reviews of products (or restaurants or hotels or movies). The point of collecting reviews and scores is to provide users with a distribution of opinion about the product, including not only the average score, but also the variance, and other aspects in the detailed reviews. Yet, we know that not every customer leaves a review—in fact only a very small fraction do. There is no reason to believe that this small fraction is a random sample of the population. It is likely that the sample skews young and well educated, potentially leading to a substantial bias in the aggregate opinions recorded.

While bias in restaurant reviews may not be a socially critical issue, similar bias could manifest itself in many other scenarios as well. Consider the use of ADS for pre-screening employment applications. As discussed above, historical under-representation of some minorities in the workforce can lead to minorities being under-represented in the training set, which in turn could push the ADS to reject more minority applicants or, more gennerally, to exhibit disparate predictive accuracy [11]. It is worth noting that the problem here is not only that some minorities are proportionally under-represented, but also that the absolute representation of some groups is low. Having 2% African Americans in the training set is a problem when they constitute 13% of the population. But it is also a problem to have only 0.2% Native Americans in the training set, even if that is representative of their proportion in the population. Such a low number can lead to Native Americans being ignored by the ADS as a small "outlier" group.

To address the problem of low absolute representation, Asudeh *et al.* [4] proposed methods to assess the coverage of a given dataset over multiple categorical features and to mitigate inadequate coverage. An important question for the data owner is what they can do about the lack of coverage. The proposed answer is to direct the data owner to acquire more data, in a way that is cognizant of the cost of data acquisition. Further, because some combinations of features are invalid or unimportant, a human expert helps identify regions of the feature space that are of interest and sets coverage goals for these regions.

Asudeh *et al.* [4] use a threshold to determine an appropriate level of coverage. Experimental results in the paper demonstrate an improvement in classifier accuracy for minority groups when additional data is acquired. This work addresses a step in the ADS lifecycle upstream from model training, and shows how improving data representativeness can improve accuracy and fairness, in the sense of disparate predictive accuracy [11]. As we will discuss in Section 5.5, there is an opportunity to integrate coverage-enhancing interventions more closely into ADS lifecycle management, both to help orchestrate the pipelines and, perhaps more importantly, to make data acquisition task-aware, setting coverage objectives based on performance requirements for the specific predictive analytics downstream, rather than based on a global threshold.

5.2 Preprocessing for Fair Classification

Even when the acquired data satisfies representativeness requirements, it may still be subject to pre-existing bias, as discussed in Section 4.2. Further, preprocessing operations, including data cleaning, filtering, and ranking, can exhibit technical bias in subtle ways, as discussed in Section 4.3. We may thus be interested in developing fairness-enhancing interventions to mitigate these effects.

In this section, we assume that data acquisition and preprocessing are preparing data for a prediction task that involves training a classifier. In most contexts, there are many prediction tasks associated with a given dataset, each representing a separate application requiring distinct domain knowledge. We first we briefly describe *associational fairness* measures, and then present methods that use *causal models* to capture this domain knowledge, and intervene on the data at the preprocessing stage to manage unfairness for a specific downstream prediction task.

Associational fairness. Most treatments of algorithmic fairness rely on statistical correlations in the observed data. A prominent example is statistical parity (discussed in Section 4.2), a requirement that the demographics of individuals receiving any outcome is the same as their demographics in the input. *Conditional statistical parity* [13] controls for a set of admissible factors to avoid some spurious correlations.

Equalized odds requires protected and privileged groups to have the same false positive rates and the same false negative rates [21]. This notion is consistent with Rawlsian equality of opportunity (EOP), discussed in Section 4.2, under the assumption that all individuals with the same true label have the same effort-based utility. As a final example, *predictive value parity* (a weaker version of calibration [31]) requires the equality of positive and negative predictive values across different groups and is consistent with luck egalitarian EOP if the predicted label is assumed to reflect an individual's effort-based utility. (See Heidari *et al.* [22] for details.)

Associational fairness measures are based on data alone, without reference to additional structure or context [41]. Consequently, these measures can be fooled by anomalies such as Simpson's paradox [47].

Causal fairness. Avoiding anomalous correlations motivates work based on causal models [29, 35, 43, 52, 53, 74]. These approaches capture background knowledge as causal

relationships between variables, usually represented as causal DAGs: directed graphs in which nodes represent variables and edges represent potential causal relationships. Discrimination is measured as the causal influence of the protected attribute on the outcome along particular causal paths that are deemed to be socially unacceptable.

An important concept in causal modeling is a *counterfactual* — an intervention where we modify the state of a set of variables $\mathbf{X}$ in the real world to some value $\mathbf{X} = \mathbf{x}$ and observe the effect on some output Y. For example, we may ask "Would this applicant have been hired if they had (or had not) been female?" Kusner *et al.* [35] define fairness in terms of counterfactuals for an individual, which in general cannot be estimated from observational data [47]. Kilbertus *et al.* [29] define fairness as equal outcome distributions for the whole population under counterfactuals for a different value of the protected attribute, however, the distributions can be equal even when there is discrimination [54].

Salimi *et al.* [54] introduced a measure called *interventional fairness* that addresses these issues, and also showed how to achieve it based on observational data, without requiring the complete causal model. The user specifies a set of *admissible* and *inadmissible* variables, indicating through which paths in the causal model influence is allowed to flow from the protected attribute to the outcome. The Markov boundary (MB) (parents, children, children's other parents) of a variable Y describes those nodes that can potentially influence Y. A key result is that, if the MB of the outcome is a subset of the MB of the admissible variables (i.e., admissible variables "shield" the outcome from the influence of sensitive and inadmissible variables), then the algorithm satisfies interventional fairness.

This condition on MB is used to design database repair algorithms, through a connection between the independence constraints encoding fairness and multi-valued dependencies (MVD). Several repair algorithms are described, and the results show that, in addition to satisfying interventional fairness, the classifier trained on repaired data performs well against associational fairness metrics.

5.3 Preprocessing for Fair Ranking

In Section 5.2 we discussed fairness-enhancing interventions for classification. We now turn to ranking, another common operation in automated hiring systems. Ranking may be invoked as part of preprocessing, with results passed to a predictive analytic; alternatively, its output may be presented directly to a human decision-maker.

Algorithmic *rankers* take a collection of candidates as input and produce a ranking (permutation) of the candidates as output. The simplest kind of a ranker is *score-based*; it computes a score of each candidate independently and returns the candidates in score order (e.g., from higher to lower, with suitably specified tie-breaking). Another common kind of a ranker is *learning-to-rank* (LTR), where supervised learning is used to predict the ranking of unseen candidates. In both score-based ranking and LTR, we may output the entire permutation, or, more often, only the highest scoring k candidates, the top-k, where k is much smaller than the size of the input n. Set selection is a special case of ranking that ignores the relative order among the top-k.

Associational fairness. Yang and Stoyanovich [71] were the first to propose associational fairness measures for ranking. Their formulation is based on an adaptation of equality-of-outcomes fairness measures, such as statistical parity (see Section 4.2) to account for position bias, a kind of technical bias that is prominent in rankings (see Section 4.3). The intuition is that, because it is more likely that a higher-ranked candidate will be selected, it is also more important to achieve statistical parity at higher ranks.

For example, suppose that there is a single job opening, that half of the applicants are women, and that at most 10 of the applicants will be invited for in-person interviews. It is insufficient to guarantee that 5 women are among the top-10, because they may end up in positions 6 through 10. Rather, men and women should alternate at the top-10, and it is particularly important to see both genders in equal proportion in earlier prefixes. To operationalize this intuition, Yang and Stoyanovich [71] place proportional representation fairness within the NDCG framework [27], imposing proprotionality constraint over every prefix of the ranking and accounting for position bias with a logarithmic discount.

Fairness measures of this kind can be used in supervised learning to train a fair LTR model. They can also be used to formulate a fairness objective that a ranking—score-based or learned—must meet to be legally or ethically admissible. Asudeh *et al.* [3] develop methods to *design fair score-based rankers* that rely on such fairness objectives. These methods query a fairness oracle that, given a ranking, returns true if it meets fairness criteria. If the ranking is found inadmissible, an alternative ranking is suggested that is both fair and close to the original, in the sense of being generated by a score-based ranker with similar feature weights.

For example, if a job applicant's score is computed as $0.5x_1 + 0.5x_2$, where x_1 is their years of experience and x_2 is their college GPA (both suitably normalized), and the resulting ranking turns out to be unfair, then the system may suggest to the hiring manager a satisfactory ranking, computed as $0.55x_1 + 0.45x_2$ instead.

Causal intersectional fairness. Much previous research on algorithmic fairness, including also on fairness in raking, considers a single sensitive attribute, such as either gender *or* race, or allows constraints on the combinations of sensitive attribute values. In all these cases, the set of sensitive attribute values induces a partitioning on the set of candidates. However, this treatment may be insufficient because we often need to impose fairness constraints on gender *and* on race, *and* on some combinations of gender and race. For example, we may be interested in detecting discrimination with respect to women, Blacks, and Black women. This is because, as noted by Crenshaw [14], it is possible to give the appearance of being fair with respect to each sensitive attribute such as race and gender separately, while being unfair with respect to *intersectional* subgroups.

Yang *et al.* [70] developed a causal framework for intersectionally fair ranking. Consider the task of selecting (and ranking) job applicants at a moving company (this example is inspired by Datta *et al.* [15]), and the corresponding causal model in Figure 3. Applicants are hired based on their qualification score Y, computed from weight-lifting ability X, and affected by gender G and race R, either directly or through X. By representing relationships between features in a causal DAG, we gain an ability to postulate which relationships between features and outcomes are legitimate, and which are

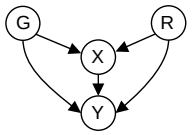

Figure 3: Causal model that includes sensitive attributes G (gender), R (race), utility score Y, other covariates $\mathbf{X}$.

potentially discriminatory.

In our example, we may state that the impact of gender G on score Y through weight-lifting ability X is legitimate (because men are on average better at lifting weights than women), but that direct impact of gender on score Y is discriminatory. Further, we may state that the impact of race R on score Y is discriminatory, both directly and through X. Technically, we can encode these constraints by treating X as a *resolving mediator* [29] for gender but not for race.

If the qualification score Y is lower for female applicants and for Blacks, then the intersectional group Black females faces greater discrimination than either the Black or the female group. The gist of the methods of Yang *et al.* [70] is to rank on counterfactual scores to achieve intersectional fairness. From the causal model, they compute model-based counterfactuals to answer the question, "What would this person's score be if they had (or had not) been a Black woman (for example)?" By ranking on counterfactual scores, they are treating every individual in the sample as though they had belonged to one specific intersectional subgroup.

This method can be justified by a connection to luck egalitarian EOP in that it considers the fine-grained impacts of group membership on the effort-based utility distribution Y.

5.4 Diversity in Set Selection and Ranking

The term *diversity* captures the quality of a collection of candidates $\mathcal{S} \subset \mathcal{C}$ of size k with regards to the variety of its constituent elements [16]. Diversity constraints on the output of an ADS may be imposed for legal reasons, such as for compliance with Title VII of the Civil Rights Act of 1964. Beyond legal requirements, benefits of diversity in hiring and elsewhere are broadly recognized [45, 65]. Further, when set selection or ranking are used as part of preprocessing, improving diversity of the training set can improve performance of the predictive analytic upstream.

A popular measure of diversity is *coverage*, which ensures representation of the demographic categories of interest in $\mathcal{S}$, or in every prefix of a ranking $\boldsymbol{\tau}(\mathcal{S})$. Coverage diversity is closely related to proportional representation fairness: a unifying formulation is to specify a lower bound ℓ_v for each sensitive attribute value v, and to enforce it as the minimum cardinality of items satisfying v in the selected set $\mathcal{S}$ [64]. If the k selected candidates need to also be ranked in the output, this formulation can be extended to specify a lower bound $\ell_{v,p}$ for every attribute v and every prefix p of the returned ranked list, with $p \leq k$ [10]. Then, at least $\ell_{v,p}$ items satisfying v should appear in the top p positions of the output. Given a set of diversity constraints, one can then seek to maximize the score utility of $\mathcal{S}$ (the sum of utility scores of the elements of $\mathcal{S}$), subject to these constraints.

Stoyanovich *et al.* [64] consider *on-line set selection*. Their work extends the classic Secretary problem [17, 37], and it's more recent k-choice variant [6], to account for diversity over a single sensitive attribute. In on-line set selection, candidates are interviewed one-by-one, their utility is revealed during the interview, the decision is made to hire or reject the candidate, and this decision is irreversible. The goal is to hire k candidates to maximize the *expected utility* of the selected set. The strategy is to (1) estimate the expected scores by observing and, initially, not hiring any candidates; then (2) hire candidates whose utility meets or exceeds the estimate. Stoyanovich *et al.* [64] estimate expected scores independently for different demographic groups to meet the ℓ_v constraints, thus deriving a relative view of utility, which is consistent with luck egalitarian EOP.

Table 1: 12 candidates with sensitive attributes `race` and `gender`. Each cell lists an individual's id, and score in parentheses.

	Male		**Female**	
White	A (99)	B (98)	C (96)	D (95)
Black	E (91)	F (91)	G (90)	H (89)
Asian	I (87)	J (87)	K (86)	L (83)

Yang *et al.* [69] also take a relative view of utility. They consider set selection and ranking in presence of *multiple* sensitive attributes, with diversity constraints on each. They observe an intersectional issue — that utility loss is non-uniform across groups, and that groups with systematically lower scores suffer the loss disproportionately. They address this by placing additional constraint on the selection procedure, balancing utility loss across groups.

For example, consider 12 candidates in Table 1 who are applying for $k = 4$ positions, and suppose that we wish to hire two candidates of each gender, and at least one candidate from each race. The set that maximizes utility while satisfying diversity is {A, B, G, K} (utility 373). This outcome selects the highest-scoring male and White candidates (A and B), but misses the highest-scoring Black (E and F) and Asian (I and J) candidates. This type of unfairness is unavoidable, but it can be distributed this unfairness in a more balanced way: the set {A, C, E, K} (utility 372) contains the top female, male, White, and Black candidates.

5.5 Holistic View of the Pipeline

In Sections 5.1-5.4, we discussed fairness and diversity considerations at different lifecycle stages. We now show how components such as these can be treated holistically.

Schelter *et al.* [56] developed FairPrep, a design and evaluation framework for fairness-enhancing interventions in machine learning pipelines that treats data as a first-class citizen. The framework implements a modular data lifecycle, enables re-use of existing implementations of fairness metrics and interventions, and integration of custom feature transformations and data cleaning operations from real world use cases. FairPrep pursues the following goals:

- Expose a *developer-centered design* throughout the lifecycle, which allows for low effort customization and composition of the framework's components.
- Surface *discrimination* and *due process* concerns, including disparate error rates, failure of a model to fit the data, and failure of a model to generalize.

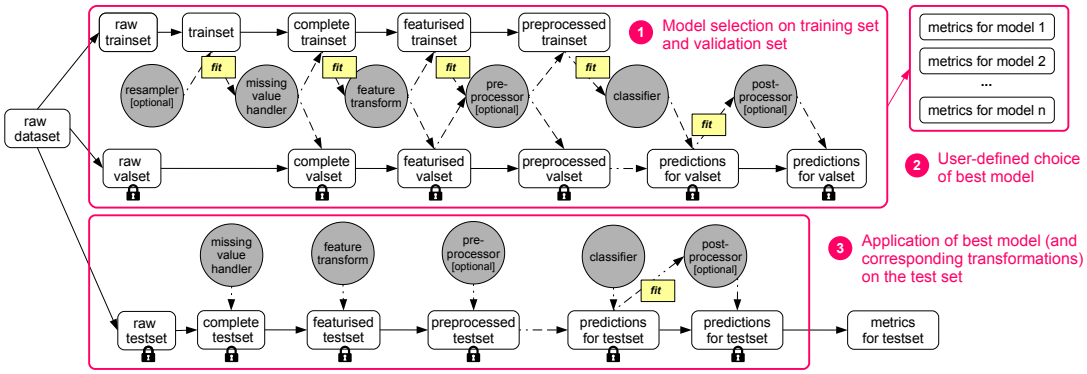

Figure 4: Data life cycle in FairPrep [56], designed to enforce isolation of test data, and to allow for customization through user-provided implementations of different components. An evaluation run consists of three different phases: (1) Learn different models, and their corresponding data transformations, on the training set; (2) Compute performance / accuracy-related metrics of the model on the validation set, and allow the user to select the 'best' model according to their setup; (3) Compute predictions and metrics for the user-selected best model on the held-out test set.

- Follow software engineering and machine learning best practices to reduce the *technical debt* of incorporating fairness-enhancing interventions into an already complex development and evaluation scenario [55, 57].

Figure 4 summarizes the architecture of FairPrep that is based on three main principles: *Data isolation* — to avoid target leakage, user code should only interact with the training set, and never be able to access the held-out test set. *Componentization* — different data transformations and learning operations should be implementable as single, exchangable standalone components; the framework should expose simple interfaces to users, supporting low effort customization. *Explicit modeling of the data lifecycle* — the framework defines an explicit, standardized data lifecycle that applies a sequence of data transformations and model training in a predefined order.

FairPrep currently focuses on data cleaning (including different methods for data imputation), and model selection and validation (including hyperparameter tuning), and can be extended to accommodate earlier lifecycle stages, such as data acquisition, integration, and curation. Schelter *et al.* [56] measured the impact of sound best practices, such as hyperparameter tuning and feature scaling, on the fairness and accuracy of the resulting classifiers, and also showcased how FairPrep enables the inclusion of incomplete data into studies and helps analyze the effects.

6. INTERPRETABILITY

Interpretability—allowing people to understand the process and the decisions of an ADS—is critical to responsibility. Interpretability is needed because it allows people, including software developers, decision-makers, auditors, regulators, individuals who are affected by ADS decisions, and members of the public, to *exercise agency* by accepting or challenging algorithmic decisions and, in the case of decision-makers, to *take responsibility* for these decisions.

Making ADS interpretable is difficult, both because they are complex (multiple steps, models with implicit assumptions), and because they rely on datasets that are often re-purposed—used outside of the original context for which they were intended. For these reasons, humans need to be able to determine the "fitness for use" of a given model or dataset, and to assess the methodology that was used to produce it.

To address this need, we have been developing interpretability tools based on the concept of a *nutritional label*, drawing an analogy to the food industry, where simple, standard labels convey information about the ingredients and production processes [60, 72]. Short of setting up a chemistry lab, the consumer would otherwise have no access to this information. Similarly, consumers of data products cannot be expected to reproduce the computational procedures just to understand fitness for their use. Nutritional labels, in contrast, are designed to support specific decisions rather than provide complete information.

6.1 Properties of a Nutritional Label

The data management community has been studying systems and standards for metadata, provenance, and transparency for decades [24, 1, 42]. We are now seeing renewed interest in these topics, and clear opportunities for this community to contribute.

Several recent projects, including the Dataset Nutrition Label project [25], Datasheets for Datasets [20], and Model Cards [40], are proposing to use metadata to support interpretability. Notably, all these method rely on manually constructed annotations. In contrast, our goal is to *generate labels automatically or semi-automatically* as a side effect of the computational process itself, embodying the paradigm of *interpretability-by-design*.

To differentiate a nutritional label from more general forms of metadata, we articulate several properties.

- *Comprehensible*: The label is not a complete (and therefore overwhelming) history of every processing step applied to produce the result. This approach has its place and has been extensively studied in the literature on scientific workflows, but is unsuitable for the applications we target. The information on a nutritional label must be short, simple, and clear.

- *Consultative*: The label should provide actionable information, not just descriptive metadata. Based on this information, consumers may cancel unused credit

Figure 5: Ranking Facts for the CS departments dataset.

cards to improve their credit score and job applicants may take a certification exam to improve their chances of being hired.

- *Comparable*: Labels should enable comparisons between related products, implying a standard. The IEEE is developing a series of ethics standards, known as the IEEE P70xx series, as part of its Global Initiative on Ethics of Autonomous and Intelligent Systems. These standards include "IEEE P7001: Transparency of Autonomous Systems" and "P7003: Algorithmic Bias Considerations" [33]. The work on nutritional labels is synergistic with these efforts.

- *Concrete*: The label must contain more than just general statements about the source of the data; such statements do not provide sufficient information to make technical decisions about fitness for use.

- *Computable*: Although primarily intended for human consumption, nutritional labels should be machine-readable to enable data discovery, integration, and automated warnings of potential misuse.

- *Composable*: Datasets are frequently integrated to construct training data; the nutritional labels must be similarly integratable. In some situations, the composed label is simple to construct: the union of sources. In other cases, the biases may interact in complex ways: a group may be sufficiently represented in each source dataset, but underrepresented in their join.

- *Concomitant*: The label should be carried with the dataset; systems should be designed to propagate labels through pipelines, modifying them as appropriate.

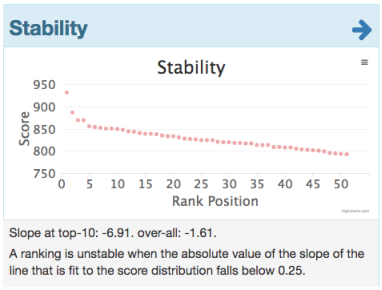

Figure 6: Stability: detailed widget.

6.2 A Nutritional Label for Rankings

To make our discussion more concrete, we now describe Ranking Facts, a system that automatically derives nutritional labels for rankings, developed by Yang *et al.* [72].

Figure 5 presents Ranking Facts that explains a ranking of Computer Science departments. Ranking Facts is made up of a collection of visual widgets. Each widget addresses an essential aspect of interpretability, and is based on our recent technical work on fairness, diversity, and stability in algorithmic rankers. We spoke about fairness and diversity in Section 5.3, and will now briefly describe the remaining components of the tool.

Features and methodology. The Recipe and Ingredients widgets help explain the ranking methodology. Recipe succinctly describes the ranking algorithm. For example, for a linear scoring formula, each attribute would be listed together with its weight. Ingredients lists attributes most material to the ranked outcome, in order of importance. For example, for a linear model, this list could present the attributes with the highest learned weights. Put another way, the explicit intentions of the designer of the scoring function about which attributes matter, and to what extent, are stated in the Recipe, while Ingredients may show attributes that are actually associated with high rank. Such associations can be derived with linear models or with other methods, such as rank-aware similarity in our prior work [59].

Stability. The Stability widget explains whether the ranking methodology is robust on the given dataset. An unstable ranking is one where slight changes to the data (e.g., due to uncertainty or noise), or to the methodology (e.g., by slightly adjusting the weights in a score-based ranker) could lead to a significant change in the output. This widget can report whether the ranking is sufficiently stable according to some pre-specified criterion, or give a score that indicates the extent of the change required for the ranking to change.

A detailed Stability widget complements the overview widget. An example is shown in Figure 6, where the stability of a ranking is quantified as the slope of the line that is fit to the score distribution, at the top-10 and over-all. A score distribution is unstable if scores of items in adjacent ranks are close to each other, and so a very small change in scores will lead to a change in the ranking. In this example the score distribution is considered unstable if the slope is 0.25 or lower. Alternatively, stability can be computed with respect to each scoring attribute, or it can be assessed using a model of uncertainty in the data. In these cases, stability quantifies

the extent to which a ranked list will change as a result of small *changes to the underlying data*. A complementary notion of stability, quantifies the magnitude of change as a result of small changes to the ranking model.

Asudeh *et al.* [3] develped methods for quantifying and improving the stability of a score-based ranker with respect to a given dataset, and focused on a notion of stability that quantifies whether the output ranking will change due to a small change in the attribute weights. This notion of stability is natural for consumers of a ranked list (i.e., those who use the ranking to prioritize items and make decisions), who should be able to assess the magnitude of the *region in the weight space* that produces the observed ranking. If this region is large, then the same ranked order would be obtained for many choices of weights, and the ranking is stable. But if this region is small, then we know that only a few weight choices can produce the observed ranking. This may suggest that the ranking was "cherry-picked" by the producer to obtain a specific outcome.

6.3 Interpretability in the Service of Trust

Interpretability means different things to different stakeholders, including individuals being affected by decisions, individuals making decisions with the help of algorithmic tools, policy-makers, regulators, auditors, vendors, data scientists who develop and deploy the systems, and members of the general public. Stoyanovich *et al.* [63] proposed a framework that connects interpretability of ADS with *trust*, which was one of the starting points of our discussion in Section 1. Indeed, remarkably little is known about how humans perceive and evaluate algorithms and their outputs, what makes a human trust or mistrust an algorithm, and how we can empower humans to exercise agency —– to adopt or challenge an algorithmic decision.

The authors argued that designers of nutritional labels should explicitly consider *what* they are explaining, *to whom*, and *for what purpose*. Further, to design effective explanations, it will be helpful to rely on concepts from social psychology such as procedural justice (that links with due process, discussed in Section 2), moral cognition, and social identity. Finally, it is necessary to experimentally validate the effectiveness of explanations, because information disclosure does not always have the intended effect.

For example, although the nutritional and calorie labelling for food are in broad use today, the information conveyed in the labels does not always affect calorie consumption. A plausible explanation is that "When comparing a $3 Big Mac at 540 calories with a similarly priced chicken sandwich with 360 calories, the financially strapped consumer [. . .] may well conclude that the Big Mac is a better deal in terms of calories per dollar" [38]. It is therefore important to understand, with the help of experimental studies, what kinds of disclosure are effective, and for what purpose.

7. CONCLUSIONS

In this article, we gave a perspective on the role that the data management research community can play in the responsible design, development, use, and oversight of Automated Decision Systems (ADS). We intentionally grounded our discussion in automated hiring tools, a specific use case that gave us ample opportunity to both appreciate the potential benefits of data science and AI in an important domain, and to get a sense of the ethical and legal risks.

We also intentionally devoted half of this paper to setting the stage — bringing in concepts from law, philosophy and social science, and grounding them in data management questions, before discussing technical research. This breakdown underscores that we (technologists) must think carefully about where in the ADS lifecycle a technical solution is appropriate, and where it simply won't do.

On a related note, an important thread that runs through this paper is that we *cannot fully automate responsibility*. While some of the duties of carrying out the task of, say, legal compliance can in principle be assigned to an algorithm, the accountability for the decisions being made by an ADS always rests with a person. This person may be a decision maker or a regulator, a business leader or a software developer. For this reason, we see our role as researchers in helping build systems that "expose the knobs" or responsibility to people, for example, in the form of explicit fairness constrains or interpretability mechanisms.

Those of us in academia have an additional responsibility to teach students about the social implications of the technology they build. A typical student is driven to develop technical skills and has an engineer's desire to build useful artifacts, such as a classification algorithm with low error rates. A typical student may not have the awareness of historical discrimination, or the motivation to ask hard questions about the choice of a model or of a metric. This typical student will soon become a practising data scientist, influencing how technology companies impact society. It is critical that the students we send out into the world have at least a rudimentary understanding of responsible data science and AI.

Towards this end, we are developing educational materials on responsible data science. Jagadish launched the first Data Science Ethics MOOC on the EdX platform in 2015 (`https://www.edx.org/course/data-science-ethics`). This course has since been ported to Coursera (`https://www.coursera.org/learn/data-science-ethics`) and to Futurum, and has been taken by thousands of students worldwide. More importantly, individual videos, including case study videos, have been individually licensed under Creative Commons and can be freely incorporated in your own teaching where appropriate.

Stoyanovich has a highly visible technical course on Responsible Data Science [62], with all materials publicly available online. In a pre-course survey, in response to the prompt, "Briefly state your view of the role of data science and AI in society", one student wrote: "It is something we cannot avoid and therefore shouldn't be afraid of. I'm glad that as a data science researcher, I have more opportunities as well as more responsibility to define and develop this 'monster' under a brighter goal." Another student responded, "Data Science [DS] is a powerful tool and has the capacity to be used in many different contexts. As a responsible citizen, it is important to be aware of the consequences of DS/AI decisions and to appropriately navigate situations that have the risk of harming ourselves or others."

8. ACKNOWLEDGEMENTS

The work of Julia Stoyanovich was supported in part by NSF Grants No. 1926250, 1934464, and 1922658. The work of Bill Howe was supported in part by NSF Grants No. 1740996 and 1934405. The work of H.V. Jagadish was supported in part by NSF Grants No. 1741022 and 1934565.

9. REFERENCES

[1] Open provenance. https://openprovenance.org. [Online; accessed 14-August-2019].

[2] S. Abiteboul and J. Stoyanovich. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. *J. Data and Information Quality*, 11(3):15:1–15:9, 2019.

[3] A. Asudeh, H. V. Jagadish, G. Miklau, and J. Stoyanovich. On obtaining stable rankings. *PVLDB*, 12(3):237–250, 2018.

[4] A. Asudeh, Z. Jin, and H. V. Jagadish. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 554–565. IEEE, 2019.

[5] Automated Decision Systems Task Force. New York City Automated Decision Systems Task Force Report. https://www1.nyc.gov/assets/adstaskforce/downloads/pdf/ADS-Report-11192019.pdf, 2019. [Online; accessed 14-August-2019].

[6] M. Babaioff, N. Immorlica, D. Kempe, and R. Kleinberg. Online auctions and generalized secretary problems. *SIGecom Exchanges*, 7(2), 2008.

[7] R. Baeza-Yates. Bias on the web. *Commun. ACM*, 61(6):54–61, 2018.

[8] F. Biessmann, D. Salinas, S. Schelter, P. Schmidt, and D. Lange. Deep learning for missing value imputation in tables with non-numerical data. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2017–2025. ACM, 2018.

[9] M. Bogen and A. Rieke. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn*, 2018.

[10] L. E. Celis, D. Straszak, and N. K. Vishnoi. Ranking with fairness constraints. In *45th International Colloquium on Automata, Languages, and Programming, ICALP*, pages 28:1–28:15, 2018.

[11] I. Y. Chen, F. D. Johansson, and D. A. Sontag. Why is my classifier discriminatory? In S. Bengio, H. M. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, pages 3543–3554, 2018.

[12] D. K. Citron and F. A. Pasquale. The scored society: Due process for automated predictions. *Washington Law Review*, 89, 2014.

[13] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 797–806. ACM, 2017.

[14] K. Crenshaw. Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. *University of Chicago Legal Forum*, (1):139–167, 1989.

[15] A. Datta, S. Sen, and Y. Zick. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 598–617. IEEE Computer Society, 2016.

[16] M. Drosou, H. V. Jagadish, E. Pitoura, and J. Stoyanovich. Diversity in big data: A review. *Big Data*, 5(2):73–84, 2017.

[17] E. Dynkin. The optimum choice of the instant for stopping a markov process. *Sov. Math. Dokl.*, 4, 1963.

[18] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im)possibility of fairness. *CoRR*, abs/1609.07236, 2016.

[19] B. Friedman and H. Nissenbaum. Bias in computer systems. *ACM Trans. Inf. Syst.*, 14(3):330–347, 1996.

[20] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. M. Wallach, H. D. III, and K. Crawford. Datasheets for datasets. *CoRR*, abs/1803.09010, 2018.

[21] M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016.

[22] H. Heidari, M. Loi, K. P. Gummadi, and A. Krause. A moral framework for understanding fair ML through economic models of equality of opportunity. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 181–190. ACM, 2019.

[23] T. Herndon, M. Ash, and R. Pollin. Does high public debt consistently stifle economic growth? a critique of Reinhart and Rogof. *Political Economy Research Institute working Paper Series*, (322), 2013.

[24] M. Herschel, R. Diestelkämper, and H. Ben Lahmar. A survey on provenance: What for? what form? what from? *VLDB J.*, 26(6):881–906, 2017.

[25] S. Holland, A. Hosny, S. Newman, J. Joseph, and K. Chmielinski. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR*, abs/1805.03677, 2018.

[26] H. V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J. M. Patel, R. Ramakrishnan, and C. Shahabi. Big data and its technical challenges. *Commun. ACM*, 57(7):86–94, 2014.

[27] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[28] J. Kappelhof. *Total Survey Error in Practice*, chapter Survey Research and the Quality of Survey Data Among Ethnic Minorities. 2017.

[29] N. Kilbertus, M. R. Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*, pages 656–666, 2017.

[30] K. Kirkpatrick. It's not the algorithm, it's the data. *Commun. ACM*, 60(2):21–23, 2017.

[31] J. M. Kleinberg, S. Mullainathan, and M. Raghavan. Inherent trade-offs in the fair determination of risk scores. In C. H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*,

volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[32] R. Koch. What is the LGPD? Brazil's version of the GDPR. `https://gdpr.eu/gdpr-vs-lgpd/`, 2018. [Online; accessed 14-August-2019].

[33] A. R. Koene, L. Dowthwaite, and S. Seth. IEEE p7003™ standard for algorithmic bias considerations: work in progress paper. In *Proceedings of the International Workshop on Software Fairness, FairWare@ICSE 2018, Gothenburg, Sweden, May 29, 2018*, pages 38–41, 2018.

[34] J. A. Kroll, J. Huey, S. Barocas, E. W. Felten, J. R. Reidenberg, D. G. Robinson, and H. Yu. Accountable algorithms. *University of Pennsylvania Law Review*, 165, 2017.

[35] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva. Counterfactual fairness. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 4066–4076, 2017.

[36] D. Lehr and P. Ohm. Playing with the data: What legal scholars should learn about machine learning. *UC Davis Law Review*, 51(2):653–717, 2017.

[37] D. V. Lindley. Dynamic programming and decision theory. *Journal of the Royal Statistical Society*, 10(1):39–51, 03 1961.

[38] G. Loewenstein. Confronting reality: pitfalls of calorie posting. *The American Journal of Clinical Nutrition*, 93(4):679–680, 2011.

[39] K. Lum and W. Isaac. To predict and serve? *Significance*, 13(5), 2016.

[40] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 220–229, 2019.

[41] S. Mitchell, E. Potash, S. Barocas, A. D'Amour, and K. Lum. Prediction-based decisions and fairness: A catalogue of choices, assumptions, and definitions. *CoRR*, abs/1811.07867, 2020.

[42] L. Moreau, B. Ludäscher, I. Altintas, R. S. Barga, S. Bowers, S. P. Callahan, G. C. Jr., B. Clifford, S. Cohen, S. C. Boulakia, S. B. Davidson, E. Deelman, L. A. Digiampietri, I. T. Foster, J. Freire, J. Frew, J. Futrelle, T. Gibson, Y. Gil, C. A. Goble, J. Golbeck, P. T. Groth, D. A. Holland, S. Jiang, J. Kim, D. Koop, A. Krenek, T. M. McPhillips, G. Mehta, S. Miles, D. Metzger, S. Munroe, J. Myers, B. Plale, N. Podhorszki, V. Ratnakar, E. Santos, C. E. Scheidegger, K. Schuchardt, M. I. Seltzer, Y. L. Simmhan, C. T. Silva, P. Slaughter, E. G. Stephan, R. Stevens, D. Turi, H. T. Vo, M. Wilde, J. Zhao, and Y. Zhao. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.

[43] R. Nabi and I. Shpitser. Fair inference on outcomes. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans , Louisiana, USA, February 2-7, 2018*, pages 1931–1940. AAAI Press, 2018.

[44] A. Narayanan. How to recognize ai snake oil. `https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf`, 2019. Arthur Miller lecture on science and ethics, MIT.

[45] S. E. Page. *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies-New Edition*. Princeton University Press, 2008.

[46] B. Pan, H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. In google we trust: Users' decisions on rank, position, and relevance. *Journal of computer-mediated communication*, 12(3):801–823, 2007.

[47] J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[48] Personal Information Protection Commission, Japan. Amended Act on the Protection of Personal Information. 2016.

[49] J. Rawls. *A theory of justice*. Harvard University Press, 1971.

[50] R. V. Reeves and D. Halikias. Race gaps in sat scores highlight inequality and hinder upward mobility. `https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality\-and-hinder-upward-mobility`, 2017. [Online; accessed 14-August-2019].

[51] J. E. Roemer. Equality of opportunity: a progress report. *Social Choice and Welfare*, 19(2):405–471, 2002.

[52] C. Russell, M. J. Kusner, J. Loftus, and R. Silva. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*, pages 6414–6423, 2017.

[53] B. Salimi, H. Parikh, M. Kayali, L. Getoor, S. Roy, and D. Suciu. Causal relational learning. In *Proceedings of the 2020 International Conference on Management of Data, SIGMOD Conference 2020, June 14-19, 2020*, pages 241–256. ACM, 2020.

[54] B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In P. A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, editors, *Proceedings of the 2019 International Conference on Management of Data, SIGMOD Conference 2019, Amsterdam, The Netherlands, June 30 - July 5, 2019*, pages 793–810. ACM, 2019.

[55] S. Schelter, F. Biessmann, T. Januschowski, D. Salinas, S. Seufert, G. Szarvas, M. Vartak, S. Madden, H. Miao, A. Deshpande, et al. On challenges in machine learning model management. *IEEE Data Eng. Bull.*, 41(4):5–15, 2018.

[56] S. Schelter, Y. He, J. Khilnani, and J. Stoyanovich. Fairprep: Promoting data to a first-class citizen in studies on fairness-enhancing interventions. In A. Bonifati, Y. Zhou, M. A. V. Salles, A. Böhm, D. Olteanu, G. H. L. Fletcher, A. Khan, and B. Yang,

editors, *Proceedings of the 23nd International Conference on Extending Database Technology, EDBT 2020, Copenhagen, Denmark, March 30 - April 02, 2020*, pages 395–398. OpenProceedings.org, 2020.

[57] D. Sculley, G. Holt, D. Golovin, E. Davydov, T. Phillips, D. Ebner, V. Chaudhary, M. Young, J.-F. Crespo, and D. Dennison. Hidden technical debt in machine learning systems. In *Advances in neural information processing systems*, pages 2503–2511, 2015.

[58] P. Stone, R. Brooks, E. Brynjolfsson, R. Calo, O. Etzioni, G. Hager, J. Hirschberg, S. Kalyanakrishnan, E. Kamar, S. Kraus, K. Leyton-Brown, D. Parkes, W. Press, A. A. Saxenian, J. Shah, M. Tambe, and A. Teller. One hundred year study on artificial intelligence: Report of the 2015-2016 study panel. *Stanford University*, 2016.

[59] J. Stoyanovich, S. Amer-Yahia, and T. Milo. Making interval-based clustering rank-aware. In A. Ailamaki, S. Amer-Yahia, J. M. Patel, T. Risch, P. Senellart, and J. Stoyanovich, editors, *EDBT 2011, 14th International Conference on Extending Database Technology, Uppsala, Sweden, March 21-24, 2011, Proceedings*, pages 437–448. ACM, 2011.

[60] J. Stoyanovich and B. Howe. Nutritional labels for data and models. *IEEE Data Eng. Bull.*, 42(3):13–23, 2019.

[61] J. Stoyanovich, B. Howe, S. Abiteboul, G. Miklau, A. Sahuguet, and G. Weikum. Fides: Towards a platform for responsible data science. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 26:1–26:6. ACM, 2017.

[62] J. Stoyanovich and A. Lewis. Teaching responsible data science: Charting new pedagogical territory. *CoRR*, abs/1912.10564, 2019.

[63] J. Stoyanovich, J. J. Van Bavel, and T. V. West. The imperative of interpretable machines. *Nature Machine Intelligence*, 2:197–199, 2020.

[64] J. Stoyanovich, K. Yang, and H. V. Jagadish. Online set selection with fairness and diversity constraints. In M. H. Böhlen, R. Pichler, N. May, E. Rahm, S. Wu, and K. Hose, editors, *Proceedings of the 21th International Conference on Extending Database Technology, EDBT 2018, Vienna, Austria, March 26-29, 2018*, pages 241–252. OpenProceedings.org, 2018.

[65] J. Surowiecki. *The wisdom of crowds*. Anchor, 2005.

[66] The European Union. Regulation (EU) 2016/679: General Data Protection Regulation (GDPR). 2016.

[67] The New York City Council. A local law to amend the administrative code of the city of new york, in relation to the sale of automated employment decision tools. `https://legistar.council.nyc.gov/LegislationDetail.aspx?ID=4344524&GUID=B051915D-A9AC-451E-81F8-6596032FA3F9`, 2020.

[68] J. Yang. The future of work: Protecting workres' civil rights in the digital age. *Testimony before the Education and Labor Committee, United States House of Representatives*, 2020.

[69] K. Yang, V. Gkatzelis, and J. Stoyanovich. Balanced ranking with diversity constraints. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6035–6042. ijcai.org, 2019.

[70] K. Yang, J. R. Loftus, and J. Stoyanovich. Causal intersectionality for fair ranking. *CoRR*, abs/2006.08688, 2020.

[71] K. Yang and J. Stoyanovich. Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management, Chicago, IL, USA, June 27-29, 2017*, pages 22:1–22:6. ACM, 2017.

[72] K. Yang, J. Stoyanovich, A. Asudeh, B. Howe, H. V. Jagadish, and G. Miklau. A nutritional label for rankings. In G. Das, C. M. Jermaine, and P. A. Bernstein, editors, *Proceedings of the 2018 International Conference on Management of Data, SIGMOD Conference 2018, Houston, TX, USA, June 10-15, 2018*, pages 1773–1776. ACM, 2018.

[73] B. Zhang and A. Dafoe. Artificial intelligence: American attitudes and trends. *Center for the Governance of AI, Future of Humanity Institute, University of Oxford*, 2019.

[74] J. Zhang and E. Bareinboim. Fairness in decision-making - the causal explanation formula. In S. A. McIlraith and K. Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 2037–2045. AAAI Press, 2018.