DOI:10.1145/3488717

Perspectives on the role and responsibility of the data-management research community in designing, developing, using, and overseeing automated decision systems.

BY JULIA STOYANOVICH, SERGE ABITEBOUL, BILL HOWE, H.V. JAGADISH, AND SEBASTIAN SCHELTER

Responsible Data Management

INCORPORATING ETHICS AND legal compliance into data-driven algorithmic systems has been attracting significant attention from the computing research community, most notably under the umbrella of fair[8] and interpretable[16] machine learning. While important, much of this work has been limited in scope to the "last mile" of data analysis and has disregarded both the *system's design, development, and use life cycle* (What are we automating and why? Is the system working as intended? Are there any unforeseen consequences post-deployment?) and the *data life cycle* (Where did the data come from? How long is it valid and appropriate?). In this article, we argue two points. First, the decisions we make during data collection and preparation profoundly impact the robustness, fairness, and interpretability of the systems we build. Second, our responsibility for the operation of these systems does not stop when they are deployed.

Example: Automated hiring systems. To make our discussion concrete, consider the use of predictive analytics in hiring. Automated hiring systems are seeing ever broader use and are as varied as the hiring practices themselves, ranging from resume screeners that claim to identify promising applicants[a] to video and voice analysis tools that facilitate the interview process[b] and game-based assessments that promise to surface personality traits indicative of future success.[c] Bogen and Rieke[5] describe the hiring process from the employer's point of view as a series of decisions that forms a funnel, with stages corresponding to

a https://www.crystalknows.com
b https://www.hirevue.com
c https://www.pymetrics.ai

sourcing, screening, interviewing, and selection. (Figure 1 depicts a slightly reinterpreted version of that funnel.)

The popularity of automated hiring systems is due in no small part to our collective quest for efficiency. In 2019 alone, the global market for artificial intelligence (AI) in recruitment was valued at $580 million.[d] Employers choose to use these systems to source and screen candidates faster, with less paperwork, and, in the post-COVID-19 world, as little in-person contact as is practical. Candidates are promised a more streamlined job-search experience, although they rarely have a say in whether they are screened by a machine.

The flip side of efficiency afforded by automation is that we rarely understand how these systems work and, indeed, whether they work. Is a resumé screener identifying promising candidates or is it picking up irrelevant—or even discriminatory—patterns from historical data, limiting access to essential economic opportunity for entire segments of the population and potentially exposing an employer to legal liability? Is a job seeker participating in a fair competition if she is being systematically screened out, with no opportunity for human intervention and recourse, despite being well-qualified for the job?

If current adoption trends are any indication, automated hiring systems are poised to impact each one of us—as employees, employers, or both. What's

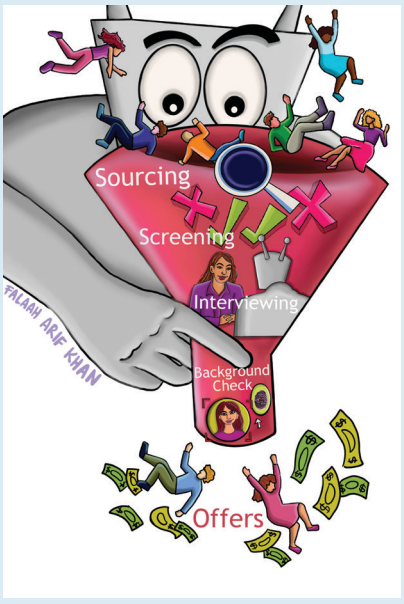
» key insights

- **Responsible data management involves incorporating ethical and legal considerations across the life cycle of data collection, analysis, and use in all data-intensive systems, whether they involve machine learning and AI or not.**
- **Decisions during data collection and preparation profoundly impact the robustness, fairness, and interpretability of data-intensive systems. We must consider these earlier life cycle stages to improve data quality, control for bias, and allow humans to oversee the operation of these systems.**
- **Data alone is insufficient to distinguish between a distorted reflection of a perfect world, a perfect reflection of a distorted world, or a combination of both. The assumed or externally verified nature of the distortions must be explicitly stated to allow us to decide whether and how to mitigate their effects.**

IMAGE BY PAVLO PLAKHOTIA

d https://www.industryarc.com/Report/19231/artificial-intelligence-in-recruitmentmarket.html

Figure 1. The hiring funnel is an example of an automated decision system—a data-driven, algorithm-assisted process that culminates in job offers to some candidates and rejections to others.



more, many of us will be asked to help design and build such systems. Yet, their widespread use far outpaces our collective ability to understand, verify, and oversee them. This is emblematic of a broader problem: the widespread and often rushed adoption of *automated decision systems* (ADSs) without an appropriate prior evaluation of their effectiveness, legal compliance, and social sustainability.

Defining ADSs. There is currently no consensus as to what an ADS is or is not, though proposed regulation in the European Union (EU), several U.S. states, and other jurisdictions are beginning to converge on some factors to consider: the degree of human discretion in the decision, the level of impact, and the specific technologies involved. As an example of the challenges, Chapter 6 of the New York City ADS Task Force report[e] summarizes a months-long struggle to, somewhat ironically, define its own mandate: to craft a definition that captures the breadth of ethical and legal concerns, yet remains practically useful. Our view is to lean towards breadth, but to tailor operational requirements and oversight mechanisms for an ADS depending on application domain and context of use, level of impact,[34] and relevant legal and regulatory requirements. For example, the use of ADSs in hiring and employment is subject to different concerns than their use in credit and lending. Further, the potential harms will be different depending on whether an ADS is used to advertise employment or financial opportunities or to help make decisions about whom to hire and to whom a loan should be offered.

To define ADS, we may start with some examples. Figure 1's hiring funnel and associated components, such as an automated resume screening tool and a tool that matches job applicants with positions, are natural examples of ADSs. But is a calculator an ADS? No, because it is not qualified with a context of use. Armed with these examples, we propose a pragmatic definition of ADSs:

- They process data about people, some of which may be sensitive or proprietary
- They help make decisions that are consequential to people's lives and livelihoods
- They involve a combination of human and automated decision-making
- They are designed to improve efficiency and, where applicable, promote equitable access to opportunity

In this definition, we deliberately direct our attention toward systems in which the ultimate decision-making responsibility is with a human and away from fully autonomous systems, such as self-driving cars. Advertising systems are ADSs; while they may operate autonomously, the conditions of their operation are specified and reviewed via negotiations between platform providers and advertisers. Further, regulation is compelling ever closer human oversight and involvement in the operations of such systems. Actuarial models, music recommendation systems, and health screening tools are all ADSs as well.

Why responsible data management? The placement of technical components that assist in decision-making—a spreadsheet formula, a matchmaking algorithm, or predictive analytics—within the *life cycle of data collection and analysis* is central to defining an ADS. This, in turn, uniquely positions the data-management community 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 enforce requirements transparently and explicitly through the life cycle of an ADS. Due to the unique blend of theory and systems in our methodological toolkit, we can help inform regulation by studying the feasible tradeoffs between different classes of legal and efficiency requirements. Our pragmatic approach enables us to support compliance by developing standards for effective and efficient auditing and disclosure, and by developing protocols for embedding these standards in systems.

In this article, we assert that the data-management community should play a central role in responsible ADS design, development, use, and oversight. Automated decision systems may or may not use AI, and they may or may not operate with a high degree of autonomy, but they all rely heavily on data. To set the stage for our discussion, we begin by interpreting the term "bias" (Section 2). We then discuss the data management-related challenges of ADS oversight and embedding responsibility into ADS life cycle management, pointing out specific opportunities for novel research contributions. Our focus is on specific issues where there is both a well-articulated need and strong evidence that technical interventions are possible. Fully addressing all the issues we raise requires socio-technical solutions that go beyond the scope of what we can do with technology alone. Although vital, since our focus is on technical data-management interventions, we do not discuss such socio-technical solutions in this article.

Crucially, the data-management problems we seek to address are not purely technical. Rather, they are socio-legal-technical. It is naïve to expect that purely technical solutions will suffice, so we must 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

e https://www1.nyc.gov/site/adstaskforce/index.page

high-risk, but one of the upsides is being able to explain to our children what we do and why it matters.

All About That Bias

We often hear that an ADS, such as an automated hiring system, operates on "biased data" and results in "biased outcomes." What is the meaning of the term "bias" in this context, how does it exhibit itself through the ADS life cycle, and what does data-management technology have to offer to help mitigate it?

Bias in a general sense refers to systematic and unfair discrimination against certain individuals or groups of individuals in favor of others. In their seminal 1996 paper, Friedman and Nissenbaum identified three types of bias that can arise in computer systems: *preexisting*, *technical*, and *emergent*.[12] We discuss each of these in turn in the remainder of this section, while also drawing on a recent fine-grained taxonomy of bias, with insightful examples that concern social media platforms, from Olteanu et al.[26]

Preexisting bias. This type of bias has its origins in society. In data-science applications, it exhibits itself in the input data. Detecting and mitigating preexisting bias is the subject of much research under the heading of algorithmic fairness.[8] Importantly, the presence or absence of this type of bias cannot be scientifically verified; rather, it must be postulated based on a belief system.[11] Consequently, the effectiveness—or even the validity—of a technical attempt to mitigate preexisting bias is predicated on that belief system. To explain preexisting bias and the limits of technical interventions, such as data debiasing, we find it helpful to use the mirror reflection metaphor, depicted in Figure 2.

The mirror metaphor. Data is a mirror reflection of the world. When we think about preexisting bias in the data, we interrogate this reflection, which is often distorted. One possible reason is that the mirror (the measurement process) introduces distortions. It faithfully represents some portions of the world, while amplifying or diminishing others. Another possibility is that even a perfect mirror can only reflect 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. First, 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. The assumed or externally verified nature of the distortions must be explicitly stated, to allow us to decide whether and how to mitigate their effects. Our second observation is that it is up to people—individuals, groups, and society at large—and not 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. The third and final observation is that, if data is used to make important decisions, such as who to hire and what salary to offer, then compensating for distortions is worthwhile. But the mirror metaphor only takes us so far. We must work much harder—usually going far beyond technological solutions—to propagate the changes back into the world and not merely brush up the reflection.[37]

As an example of preexisting bias in hiring, consider the use of an applicant's Scholastic Assessment Test (SAT) score during the screening stage. It has been documented that the mean score of the math section of the SAT, as well as the shape of the score distribution, differs across racial groups.[28] If we believed that standardized test scores were sufficiently impacted by preparation courses and that the score itself says more about socioeconomic conditions than an individual's academic potential, then we would consider the data to be biased. We may then seek to correct for that bias before using the feature, for example, by selecting the top-performing individuals of each racial group, or by using a more sophisticated *fair ranking method* in accordance with our beliefs about the nature of the bias and with our bias mitigation goals.[40] Alternatively, we may disregard this feature altogether.

Technical bias. This type of bias arises due to the operation of the technical system itself, and it can amplify

Figure 2. Data as a mirror reflection of the world,[37] illustrated by Falaah Arif Khan.

(a) Data is an image of the world, its mirror reflection. When we think about bias in the data, we interrogate this reflection. Does the data systematically over-represent or under-represent some parts of the world, or does it distort reality in some other way?

(b) Even if we were able to reflect the world perfectly in the data, it would still be a reflection of the world as it is, not how it could or should be. People—not data or algorithms—must decide what world we want to live in.

(c) Debiasing data may be worthwhile if we are about to base consequential decisions on that data. But we must be careful to propagate the changes back into the world, not merely touch up its reflection.

preexisting bias. Technical bias, particularly when it is due to preprocessing decisions or post-deployment issues in data-intensive pipelines, has been noted as problematic,[23,26,33] but it has so far received limited attention when it comes to diagnostics and mitigation techniques. We now give examples of potential sources of technical bias in several ADS life cycle stages, which are particularly relevant to data management.

Data cleansing. 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 gender choice but also allows gender to be unspecified. Suppose that about half of the applicants identify as men and half as women, but that women are more likely to omit gender. If mode imputation—replacing a missing value with the most frequent value for the feature, a common setting in scikit-learn—is applied, then all (predominantly female) unspecified gender values will be set to male. More generally, multiclass classification for missing-value imputation typically only uses the most frequent classes as target variables,[4] leading to a distortion for small groups, because membership in these groups will not 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. While dealing with null values is known to be difficult and is already considered among the issues in data cleansing, the needs of responsible data management introduce new problems. It has been documented that data-quality issues often disproportionately affect members of historically disadvantaged groups,[20] so we risk compounding technical bias due to data representation with bias due to statistical concerns.

The flip side of efficiency afforded by automation is that we rarely understand how these systems work and, indeed, whether they work.

Other data transformations that can introduce skew include text normalization, such as lowercasing, spell corrections, or stemming. These operations can be seen as a form of aggregation, in effect collapsing terms with different meanings under the same representation. For example, lowercasing "Iris," a person's name, as "iris" will make it indistinguishable from the name of a flower or from the membrane behind the cornea of the eye, while stemming the terms "[tree] leaves" and "[he is] leaving" will represent both as "leav."[26]

Other examples of aggregation that can lead to data distribution changes include "zooming out" spatially or temporally: replacing an attribute value with a coarser geographic or temporal designation or mapping a location to the center of the corresponding geographical bounding box.[26]

Filtering. Selections and joins are commonly used as part of data preprocessing. A selection operation checks each data record against a predicate—for instance, U.S. address ZIP code is 10065 or age is less than 30—and retains only those records that match the predicate. A join combines data from multiple tables—for example, creating a record that contains a patient's demographics and clinical records using the social security number attribute contained in both data sources as the join key. These operations can arbitrarily change the proportion of protected groups (for example, female gender) even if they do not directly use the sensitive attribute (for example, gender) as part of the predicate or the join key. For example, selecting individuals whose mailing address ZIP code is 10065—one of the most affluent locations on Manhattan's Upper East Side—may change the data distribution by race. Similarly, joining patient demographic data with clinical records may introduce skew by age, with fewer young individuals having matching clinical records. These changes in proportion may be unintended but are important to detect, particularly when they occur during one of many preprocessing steps in the ADS pipeline.

Another potential source of techni-

cal bias is the use of pretrained 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 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: Items in the top-left corner of the screen attract more attention.[3] 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, which implies prioritization. Depending on the application's needs and on the decision-maker's level of technical sophistication, this problem can be addressed by suitably randomizing the ranking, showing results with ties, or plotting the score distribution.

Emergent bias. This type of bias arises in the context of use of the technical system. In Web ranking and recommendation in e-commerce, a prominent example is "rich-get-richer": searchers tend to trust systems to show them the most suitable items at the top positions, which in turn shapes a searcher's idea of a satisfactory answer.

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 Rieke[5] 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."

Emergent bias is particularly difficult to detect and mitigate, because it refers to the impacts of an ADS outside the systems' direct control. We will cover this in the "Overseeing ADS" section.

Managing the ADS Data Life Cycle

Automated decision systems critically depend on data and should be seen through the lens of the *data life cycle*.[19] Responsibility concerns, and important decision points, arise in data sharing, annotation, acquisition, curation, cleansing, and integration. Consequently, substantial 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 life cycle stages.

Database systems centralize correctness constraints to simplify application development with the help of schemas, standards, and transaction protocols. As algorithmic fairness and interpretability emerge as first-class requirements, there is a need to develop generalized solutions that embed them as constraints and that work across a range of applications. In what follows, we highlight promising examples of our own recent and ongoing work that is motivated by this need. These examples underscore that tangible technical progress is possible and that much work remains to be done to offer systems support for the responsible management of the ADS life cycle. These examples are not intended to be exhaustive, but merely illustrate technical approaches that apply to different points of the data life cycle. Additional examples, and research directions, are discussed in Stoyanovich et al.[37] Before diving into the details, we recall the previously discussed mirror-reflection metaphor, as a reminder of the limits of technical interventions.

Data acquisition. Consider the use of an ADS for pre-screening employment applications. Historical underrepresentation of women and minorities in the workforce can lead to an underrepresentation of these groups in the training set, which in turn could push the ADS to reject more minority applicants or, more generally, to exhibit disparate predictive accuracy.[7] 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 mitigate low absolute representation, Asudeh et al.[2] assess the coverage of a given dataset over multiple categorical features. An important question for an ADS vendor is, then, what can it do about the lack of coverage. The proposed answer is to direct them to acquire more data, in a way that is cognizant of the cost of data acquisition. Asudeh et al.[2] use a threshold to determine an appropriate level of coverage and experimentally demonstrate an improvement in classifier accuracy for minority groups when additional data is acquired.

This work addresses a step in the ADS life cycle upstream from model training and shows how improving data representativeness can improve accuracy and fairness, in the sense of disparate predictive accuracy.[7] There are clear future opportunities to integrate coverage-enhancing interventions more closely into ADS life cycle 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.

Data preprocessing. Even when the acquired data satisfies representativeness requirements, it may still be subject to preexisting bias, as discussed in the "Preexisting bias" section. We may thus be interested in developing interventions to mitigate these effects. The algorithmic fairness community has

developed dozens of methods for data and model de-biasing, yet the vast majority of these methods take an *associational interpretation of fairness* that is solely based on data, without reference to additional structure or context. In what follows, we present two recent examples of work that take a causal interpretation of fairness: a database repair framework for fair classification by Salimi et al.[29] and a framework for fair ranking that mitigates intersectional discrimination by Yang et al.[38] We focus on examples of causal fairness notions here because they correspond very closely to the methodological toolkit of data management by making explicit the use of structural information and constraints.

Causal fairness approaches—for example, Kilbertus et al.[21] and Kusner et al.[22]—capture background knowledge as causal relationships between variables, usually represented as causal DAGs, or directed acyclic graphs, in which nodes represent variables, and edges represent potential causal relationships. Consider the task of selecting job applicants at a moving company and the corresponding causal model in Figure 3, an example inspired by Datta et al.[10] 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 potentially discriminatory. In our example, the impact of gender (G) on the decision to hire an individual for a position with a moving company (Y) may be considered admissible if it flows through the node representing weight-lifting ability (X). On the other hand, the direct impact of gender on the decision to hire would constitute direct discrimination and would thus be considered inadmissible.

Salimi et al.[29] introduced a measure called *interventional fairness* for classification and showed how to achieve it based on observational data, without requiring the complete causal model. The authors consider the Markov boundary (MB)—parents, children, children's other parents—of a variable Y, which describes whether those nodes can potentially influence Y. Their key result is that the algorithm satisfies interventional fairness if the MB of the outcome is a subset of the MB of the admissible variables—that is, admissible variables "shield" the outcome from the influence of sensitive and inadmissible variables. This condition on the MB is used to design *database repair algorithms*, through a connection between the independence constraints encoding fairness and multivalued dependencies (MVD) that can be checked using the training data. 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.

The data management problems we are looking to address are not purely technical. Rather, they are socio-legal-technical.

As another example of a data preprocessing method that makes explicit use of structural assumptions, Yang et al.[38] developed a causal framework for *intersectionally fair ranking*. Their motivation is that 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.[9] For example, if fairness is taken to mean proportional representation among the top-k, it is possible to achieve proportionality for each gender subgroup (for instance, men and women) and for each racial subgroup (for example, Black and White), while still having inadequate representation for a subgroup defined by the intersection of both attributes (for example, Black women). The gist of the methods of Yang et al.[38] is to use a causal model to compute model-based *counterfactuals*, answering the question: "What would this person's score be if she had been a Black woman (for example)?" and then ranking on counterfactual scores to achieve intersectional fairness.

Data-distribution debugging. We now return to our discussion of technical bias and consider data-distribution shifts, which may arise during data preprocessing and impact machine learning-model performance downstream. In contrast to important prior work on data-distribution shift detection in deployed models—for instance, Rabanser et al.[27]—our focus

is explicitly on data manipulation, a cause of data-distribution shifts that has so far been overlooked. We will illustrate how this type of bias can arise and will suggest an intervention: a data-distribution debugger that helps surface technical bias, allowing a data scientist to mitigate it.[33]

Consider Ann, a data scientist at a job-search platform that matches profiles of job seekers with openings for which they are well-qualified and in which they may be interested. A job seeker's interest in a position is estimated based on several factors, including the salary and benefits being offered. Ann uses applicants' resumes, self-reported demographics, and employment histories as input. Following her company's best practices, she starts by splitting her dataset into training, validation, and test sets. Ann then uses pandas, scikit-learn, and accompanying data transformers to explore the data and implement data preprocessing, model selection, tuning, and validation. Ann starts preprocessing by computing value distributions and correlations for the features in the dataset and identifying missing values. She will use a default imputation method in scikit-learn to fill these in, replacing missing values with the mode value for that feature. Finally, Ann implements model selection and hyperparameter tuning, selecting a classifier that displays sufficient accuracy.

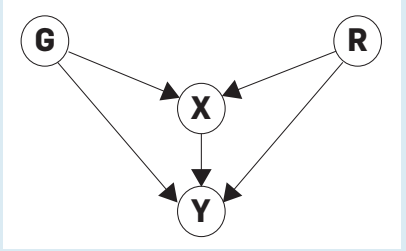
When Ann more closely considers the performance of the classifier, she observes a disparity in predictive accuracy:[7] Accuracy is lower for older job seekers, who are frequently matched with lower-paying positions than they would expect. Ann now needs to understand why this is the case, whether any of her technical choices during pipeline construction contributed to this disparity, and what she can do to mitigate this effect.

It turns out that this issue was the result of a *data-distribution bug*—a shift in the values of a feature that is important for the prediction and that is the result of a technical choice during pre-processing. Here, that feature is the number of years of job experience. The bug was introduced because of Ann's assumption that the values of this feature are *missing at random* and because of her choice to use mode imputation, which is consistent with this assumption. In fact, values were missing more frequently for older job seekers: They would not enter a high value in "years of experience" because they might be afraid of age discrimination. This observation is consistent with the intuition that individuals are more likely to withhold information that may disadvantage them. Taken together, these two factors resulted in imputed years-of-experience values skewing lower, leading to a lower salary-requirement estimate and impacting older applicants more than younger ones.

Data-distribution bugs are difficult to catch. In part, this is because different pipeline steps are implemented using different libraries and abstractions, and the data representation often changes from relational data to matrices during data preparation. Further, preprocessing often combines relational operations on tabular data with estimator/transformer pipelines, a composable and nestable abstraction for combining operations on array data which originates from scikit-learn and is executed in a hard-to-debug manner with nested function calls.

Grafberger et al. designed and implemented mlinspect,[15] a lightweight data-distribution debugger that supports automated inspection of data-intensive pipelines to detect the accidental introduction of statistical bias and linting for best practices. The mlinspect library extracts logical query plans—modeled as DAGs of preprocessing operators—from pipelines that use popular libraries, such as pandas and scikit-learn, and combines relational operations and estimator/transformer pipelines. The library automatically instruments the code and traces the impact of operators on properties, such as the distribution of sensitive groups in the data. mlinspect is a necessary first step in what we hope will be a long line of work in collectively developing data-science best practices and the tooling to support their broad adoption. Much important work remains to allow us to start treating data as a first-class citizen in software development.

Figure 3. Causal model includes sensitive attributes: *G* (gender), *R* (race), *X* (weight-lifting ability), and *Y* (utility score).



Overseeing ADS

We are in the midst of a global trend to regulate the use of ADSs. In the EU, the General Data Protection Regulation (GDPR) offers individuals protections 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 and Brazil, is in close alignment with the GDPR. In the U.S., many major cities, a handful of states, and the Federal government are establishing task forces and issuing guidelines about responsible development and technology use. With its focus on data rights and data-driven decision-making, the GDPR is, without a doubt, the most significant piece of technology regulation to date, serving as a "common denominator" for the oversight of data collection and usage, both in the EU and worldwide. For this reason, we will discuss the GDPR in some depth in the remainder of this section.

The GDPR aims to protect the rights and freedoms of natural persons with regard to how their personal data is processed, moved, and exchanged (Article 1). The GDPR is broad in scope and applies to "the processing of personal data wholly or partly by automated means" (Article 2), both in the private and public sectors. Personal data is broadly construed and refers to any information relating to an identified or identifiable natural person, called the *data subject* (Article 4). The GDPR aims to give data subjects insight into, and control over, the collection and processing of their personal data. Providing such insight, in response to the "right to be informed," requires

technical methods for interpretability, discussed in the following section, "Interpretability for a range of stakeholders." We will also highlight, in the upcoming section, "Removing personal data," the right to erasure as a representative example of a regulatory requirement that raises a concrete data-management challenge. Additional details can be found in Abitebout and Stoyanovich.[1]

As we have done throughout this article, we highlight specific challenges within the broad topic of ADS oversight and outline promising directions for technical work to address these challenges. It is important to keep in mind that ADS oversight will not admit a purely technical solution. Rather, we hope that technical interventions will be part of a robust distributed infrastructure of accountability, in which multiple stakeholder groups participate in ADS design, development, and oversight.

Interpretability for a range of stakeholders. Interpretability—allowing people to understand the process and decisions of an ADS—is critical to the responsible use of these systems. Interpretability means different things to different stakeholders, yet the common theme is that it allows people, including software developers, decision-makers, auditors, regulators, individuals who are affected by ADS decisions, and members of the public at large, to exercise agency by accepting or challenging algorithmic decisions and, in the case of decision-makers, to take responsibility for these decisions.

Interpretability rests on making explicit the interactions between the computational process and the data on which it acts. Understanding how code and data interact is important both when an ADS is interrogated for bias and discrimination, and when it is asked to explain an algorithmic decision that affects an individual.

To address the interpretability needs of different stakeholders, several recent projects have been developing tools based on the concept of a nutritional label—drawing an analogy to the food industry, where simple, standard labels convey information about ingredients and production processes. Short of setting up a chemistry lab, a food consumer would otherwise have no access to this information. Similarly, consumers of data products or individuals affected by ADS decisions cannot be expected to reproduce the data collection and computational procedures. These projects include the Dataset Nutrition Label,[18] Datasheets for Datasets,[13] Model Cards,[25] and Ranking Facts,[39] which all use specific kinds of metadata to support interpretability. Figure 4 offers an example of a nutritional label; it presents Ranking Facts[39] to explain a ranking of computer science departments.

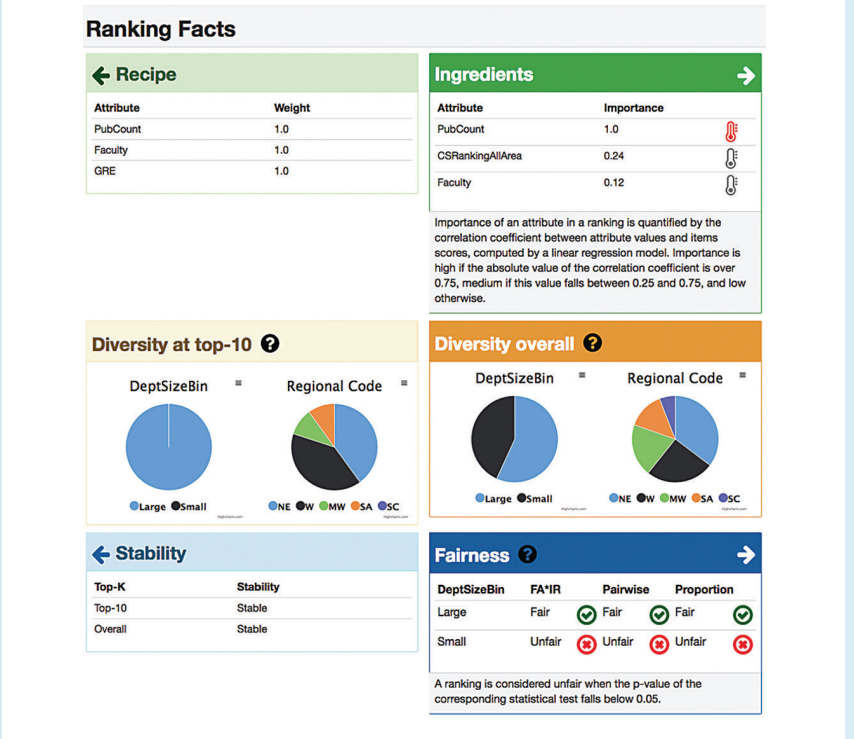
In much of this work, nutritional labels are manually constructed, and they describe a single component in the data life cycle, typically a dataset or a model. Yet, to be broadly applicable, and to faithfully represent the computational process and the data on which it acts, nutritional labels should be generated *automatically* or *semiautomatically* as a side effect of the computational process itself, embodying the paradigm of *interpretability by design*.[36] This presents an exciting responsible data-management challenge.

The data-management community has been studying systems and standards for metadata and provenance for decades.[17] This includes work on fine-grained provenance, where the goal is to capture metadata associated with a data product and propagate it through a series of transformations, to explain its origin and history of derivation, and to help answer questions about the robustness of the computational process and the trustworthiness of its results. There is now an opportunity to revisit many of these insights and to extend them to support the interpretability needs of different stakeholders, both technical and non-technical.

Removing personal data. The right to be forgotten is originally motivated by the desire of individuals not to be perpetually stigmatized by something they did in the past. Under pressure from despicable social phenomena such as revenge porn, it was turned into law in 2006 in Argentina, and since then in the EU, as part of the GDPR (Article 17), stating that data subjects have the right to request the timely erasure of their personal data.

An important technical issue of clear relevance to the data-management community is deletion of infor-

Figure 4. Ranking Facts for the CS department's dataset.



mation in systems that are designed explicitly to accumulate data. Making data-processing systems GDPR-compliant has been identified as one of the data-management community's key research challenges.[35] The requirement of efficient deletion is in stark contrast with the typical requirements for data-management systems, necessitating substantial rethinking and redesign of the primitives, such as enhancing fundamental data structures with efficient delete operations.[30]

Data deletion must be both permanent and deep, in the sense that its effects must propagate through data dependencies. To start, it is difficult to guarantee that all copies of every piece of deleted data have actually been deleted. Further, when some data is deleted, the remaining database may become inconsistent, and may, for example, include dangling pointers. Additionally, production systems typically do not include a strong provenance mechanism, so they have no means of tracking the use of an arbitrary data item (one to be deleted) and reasoning about the dependencies on that data item in derived data products. Although much of the data-management community's attention over the years has been devoted to tracking and reasoning about provenance, primarily in relational contexts and in workflows (see Herschel et al.[17] for a recent survey), there is still important work to be done to make these methods both practically feasible and sufficiently general to accommodate current legal requirements.

An important direction that has only recently come into the academic community's focus concerns ascertaining the effects of a deletion on downstream processes that are not purely relational but include other kinds of data analysis tasks, such as data mining or predictive analytics. Recent research[14,31] argues that it is not sufficient to merely delete personal user data from primary data stores such as databases, but that machine-learning models trained on stored data also fall under the regulation. This view is supported by Recital 75 of the GDPR: "The risk to the rights and freedoms of natural persons ... may result from personal data processing ... where personal aspects are evaluated, in particular analyzing or predicting aspects concerning performance at work, economic situation, health, personal preferences or interests, reliability or behavior, location or movements." The machine-learning community has been working on this issue under the umbrella of *machine unlearning*.[6,14] Given a model, its training data, and a set of user data to delete/unlearn, the community proposes efficient ways to accelerate the retraining of the model. However, these approaches ignore the constraints imposed by the complexity of production set-ups (such as redeployment costs) and are thereby hard to integrate into real-world ML applications.[32]

We must learn to step outside our engineering comfort zone and to start reasoning in terms of values and beliefs.

Requests for deletion may also conflict with other laws, such as requirements to keep certain transaction data for some period or requirements for fault tolerance and recoverability. Understanding the impact of deletion requests on our ability to offer guarantees on system resilience and performance, and developing appropriate primitives and protocols for practical use, is another call to action for the data-management community.

Conclusion

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

An important point 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, 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.

Those of us in academia have an

additional responsibility to teach students about the social implications of the technology they build. Typical students are driven to develop technical skills and have an engineer's desire to build useful artifacts, such as a classification algorithm with low error rates. They are also increasingly aware of historical discrimination that can be reinforced, amplified, and legitimized with the help of technical systems. Our students will soon become practicing data scientists, influencing how technology companies impact society. It is our responsibility as educators to equip them with the skills to ask and answer the hard questions about the choice of a dataset, a model, or a metric. It is critical that the students we send out into the world understand responsible data science.

Toward this end, we are developing educational materials and teaching courses on responsible data science. H.V. Jagadish launched the first Data Science Ethics MOOC on the EdX platform in 2015. This course has since been ported to Coursera and FutureLearn, and it has been taken by thousands of students worldwide. Individual videos are licensed under Creative Commons and can be freely incorporated in other courses where appropriate. Julia Stoyanovich teaches highly visible technical courses on Responsible Data Science,[24] with all materials publicly available online. These courses are accompanied by a comic book series, developed under the leadership of Falaah Arif Khan, as supplementary reading.

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."

Acknowledgments

This work was supported in part by NSF Grants No. 1934464, 1934565, 1934405, 1926250, 1741022, 1740996, 1916505, by Microsoft, and by Ahold Delhaize. All content represents the opinion of the authors and is not necessarily shared or endorsed by their respective employers or sponsors. C

References

1. Abiteboul, S. and Stoyanovich, J. Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. *J. of Data and Information Quality 11*, 3 (2019), 15:1–15:9.
2. Asudeh, A., Jin, Z., and Jagadish, H.V. Assessing and remedying coverage for a given dataset. In *35th IEEE International Conference on Data Engineering* (April 2019), 554–565.
3. Baeza-Yates, R. Bias on the web. *Communications of the ACM 61*, 6 (2018), 54–61.
4. Biessmann, F., Salinas, D., Schelter, S., Schmidt, P., and Lange, D. Deep learning for missing value imputation in tables with non-numerical data. In *Proceedings of the 27th ACM Intern. Conf. on Information and Knowledge Management* (2018), 2017–2025.
5. Bogen, M. and Rieke, A. Help wanted: An examination of hiring algorithms, equity, and bias. *Upturn* (2018).
6. Cauwenberghs, G. and Poggio, T. Incremental and decremental support vector machine learning. *NeurIPS* (2001), 409–415.
7. Chen, I., Johansson, F., and Sontag, D. Why is my classifier discriminatory? S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors. In *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018*, 3543–3554.
8. Chouldechova, A. and Roth, A. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM 63*, 5 (2020), 82–89.
9. Crenshaw, K. 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* (1989), 139–167.
10. Datta, A., Sen, S., and Zick, Y. Algorithmic transparency via quantitative input influence: Theory and experiments with learning systems. In *IEEE Symposium on Security and Privacy* (May 2016), 598–617.
11. Friedler, S., Scheidegger, C., and Venkatasubramanian, S. The (im)possibility of fairness: Different value systems require different mechanisms for fair decision making. *Communications of the ACM 64*, 4 (2021), 136–143.
12. Friedman, B. and Nissenbaum, H. Bias in computer systems. *ACM Transactions on Information Systems 14*, 3 (1996), 330–347.
13. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. Datasheets for datasets. *CoRR* (2018), abs/1803.09010.
14. Ginart, A., Guan, M., Valiant, G., and Zou, J. Making AI forget you: Data deletion in machine learning. In *NeurIPS* (2019), 3513–3526.
15. Grafberger, S., Stoyanovich, J., and Schelter, S. Lightweight inspection of data preprocessing in native machine learning pipelines. In *11th Conf. on Innovative Data Sys. Research, Online Proceedings* (January 2021), http://www.cidrdb.org.
16. Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., and Pedreschi, D. A survey of methods for explaining black box models. *ACM Computing Surveys 51*, 5 (2019), 93:1–93:42.
17. Herschel, M., Diestelkämper, R., and Ben Lahmar, H. A survey on provenance: What for? What form? What from? *VLDB Journal 26*, 6 (2017), 881–906.
18. Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. The dataset nutrition label: A framework to drive higher data quality standards. *CoRR* (2018), abs/1805.03677.
19. Jagadish, H.V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J., Ramakrishnan, R., and Shahabi, C. Big data and its technical challenges. *Communications of the ACM 57*, 7 (2014), 86–94.
20. Kappelhof, J. Survey research and the quality of survey data among ethnic minorities. In *Total Survey Error in Practice*, Wiley (2017).
21. Kilbertus, N., Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems* (2017), 656–666.
22. Kusner, M., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. I. Guyon, U. von Luxburg, S. Bengio, H.M. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, In *Advances in Neural Information Processing Systems 30:* (2017), 4066–4076.
23. Lehr, D. and Ohm, P. Playing with the data: What legal scholars should learn about machine learning. *UC Davis Law Review 51*, 2 (2017), 653–717.
24. Lewis, A. and Stoyanovich, J. Teaching responsible data science. *Intern. J. of Artificial Intelligence in Education* (2021).
25. Mitchell, M., et al. Model cards for model reporting. In *Proceedings of the Conf. on Fairness, Accountability, and Transparency 2019*, 220–229.
26. Olteanu, A., Castillo, C., Diaz, F., and Kiciman, E. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers Big Data 2*, 13 (2019).
27. Rabanser, S., Günnemann, S., and Lipton, Z. Failing loudly: An empirical study of methods for detecting dataset shift. H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors. In *Advances in Neural Information Processing Systems 32* (December 2019), 1394–1406.
28. Reeves, R. and Halikias, D. Race gaps in SAT scores highlight inequality and hinder upward mobility. Brookings (2017), https://www.brookings.edu/research/race-gaps-in-sat-scores-highlight-inequality-and-hinder-upward-mobility.
29. Salimi, B., Rodriguez, L., Howe, B., and Suciu, D. Interventional fairness: Causal database repair for algorithmic fairness. P.A. Boncz, S. Manegold, A. Ailamaki, A. Deshpande, and T. Kraska, editors. In *Proceedings of the 2019 Intern. Conf. on Management of Data*, 793–810.
30. Sarkar, S., Papon, T., Staratzis, D., and Athanassoulis, M. Lethe: A tunable delete-aware LSM engine. In *Proceedings of the 2020 Intern. Conf. on Management of Data*.
31. Schelter, S. "Amnesia"–a selection of machine learning models that can forget user data very fast. *Conf. on Innovative Data Systems Research*, 2020.
32. Schelter, S., Grafberger, S., and Dunning, T. HedgeCut: Maintaining randomised trees for low-latency machine unlearning. In *Proceedings of the 2021 Intern. Conf. on Management of Data*.
33. Schelter, S. and Stoyanovich, J. Taming technical bias in machine learning pipelines. *IEEE Data Engineering Bulletin 43*, 4 (2020).
34. Selbst, A. Disparate impact in big data policing. *Georgia Law Review 52*, 109 (2017).
35. Shastri, S., Banakar, V., Wasserman, M., Kumar, A., and Chidambaram, V. Understanding and benchmarking the impact of GDPR on database systems. *PVLDB* (2020).
36. Stoyanovich, J. and Howe, B. Nutritional labels for data and models. *IEEE Data Engineering Bulletin 42*, 3 (2019), 13–23.
37. Stoyanovich, J., Howe, B., and Jagadish, H.V. Responsible data management. In *Proceedings of the VLDB Endowment 13*, 12 (2020), 3474–3488.
38. Yang, K., Loftus, J., and Stoyanovich, J. Causal intersectionality and fair ranking. K. Ligett and S. Gupta, editors. In *2nd Symposium on Foundations of Responsible Computing, Volume 192 of LIPICS, Schloss Dagstuhl–Leibniz Center for Informatics* (June 2021), 7:1-7:20.
39. Yang, K., Stoyanovich, J., Asudeh, A., Howe, B., Jagadish, H.V., and Miklau, G. A nutritional label for rankings. G. Das, C. Jermaine, and P. Bernstein, editors. In *Proceedings of the 2018 Intern. Conf. on Management of Data*, 1773–1776.
40. Zehlike, M., Yang, K., and Stoyanovich, J. Fairness in ranking: A survey. *CoRR* (2021), abs/2103.14000.

Julia Stoyanovich (stoyanovich@nyu.edu) is an associate professor at New York University, New York, NY, USA.

Serge Abiteboul is a researcher at Inria & École Normale Supérieure, Paris, France.

Bill Howe is an associate professor at the University of Washington, Seattle, WA, USA.

H.V. Jagadish is a professor at the University of Michigan, Ann Arbor, MI, USA.

Sebastian Schelter is an assistant professor at the University of Amsterdam, Amsterdam, The Netherlands.