"How Much Do You Want to Play God in Controlling The Dataset?": Data Scientists' Perspectives on Modifying Data to Address Fairness Concerns

Kevin Bryson[1], **Galen Harrison**[2], **Margaret Jennings**[1], **Alex Kale**[1], **Blase Ur**[1]

[1]University of Chicago, [2]University of Virginia

kbryson@uchicago.edu, gh7vp@virginia.edu, jenningsm@uchicago.edu, kalea@uchicago.edu, blase@uchicago.edu

Abstract

Machine-learning (ML) models are often trained on data that embed discrimination and biases. We wondered under what circumstances data-science practitioners would want to modify a model's underlying training data directly, instead of attempting to fix a model learned from biased data. We also wondered what tools and practices they would need to do so. To answer these questions, we conducted scenario-based interviews with 20 practitioners about using data modification methods (DMMs)—pre-processing methods that directly change a model's underlying training data—to address fairness issues. Participants acknowledged pressing fairness issues in the scenario we presented, but were reluctant to violate what they perceived as data science and ML norms by modifying training data. Participants frequently felt that their role (both in the scenario and in their own work contexts) meant that they had little responsibility or agency over the ethical outcomes of their models. Our findings suggest that norms within data science and ML practice can act as a barrier to fairness processes. This work calls for the fairness community to reevaluate the community norms and values that create these barriers and to reconsider the range of valid fairness interventions.

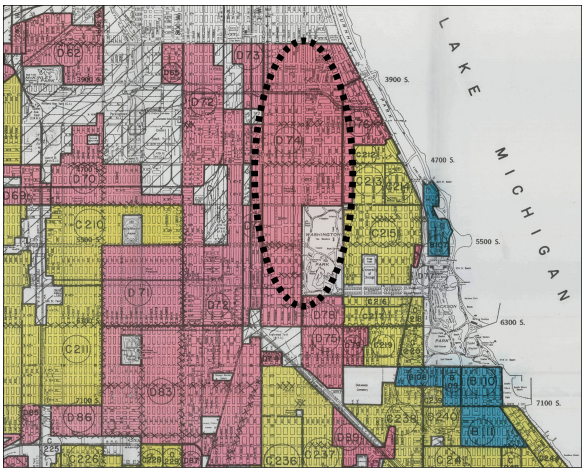
1 Introduction

Society's reliance on data-driven machine learning (**ML**) systems has frequently raised concerns about the fairness of these systems (Barabas et al. 2020; Miceli, Posada, and Yang 2022; Barocas and Selbst 2016; Sambasivan et al. 2021; Lima et al. 2025). Repeatedly, ML systems have refracted broader issues of racism into specific patterns of decisions: Black men are predicted to have higher recidivism risk (Angwin et al. 2016); Black students are predicted to have higher high school dropout risk (Feathers 2023; Perdomo et al. 2023); and immigrants are predicted as more likely to commit welfare fraud (Burgess 2023; Amnesty International 2024). In each case, populations that are already marginalized face amplified discrimination from algorithmic systems.

Though "garbage in, garbage out" is a well-known maxim, fairness work does not often directly address the inequity embedded in data (Jacobs and Wallach 2021; Rismani et al. 2025; Simson, Fabris, and Kern 2024). Instead, both academic papers and industry practice typically focus on the inequity captured in downstream models trained on that data. Likewise, fairness toolkits (Bellamy et al. 2018; Bird et al. 2020; Saleiro et al. 2018; Wexler et al. 2020; Delaney et al. 2024; Hardt et al. 2021) focus primarily on modifying the model that has been learned from data, rather than directly addressing critical, underlying issues in the data (Deng et al. 2022; Richardson et al. 2021; Balayn et al. 2023; Rismani et al. 2025; Wachter, Mittelstadt, and Russell 2021). Though data decisions affect fairness outcomes, those decisions are widely overlooked in academic research and by practitioners (Harrison et al. 2024; Simson, Fabris, and Kern 2024; Guha et al. 2024; Lakkaraju et al. 2017). Similarly, previous work has found that practitioners often focus on models, not data, eventually cascading to produce downstream harms (Sambasivan et al. 2021). In light of this paradoxical behavior towards data, we wondered why data-science practitioners do not seem to engage with the underlying issues in their data to the same degree they seem to be trying to modify their trained models. If training data is regarded as immutable "ground truth," yet also demonstrates overt discrimination and biases, what would practitioners do?

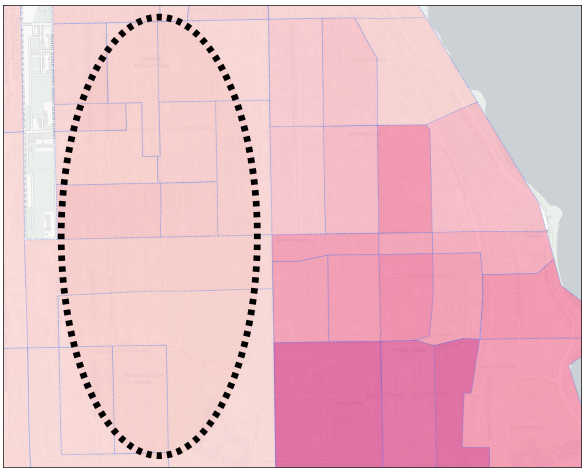
In this work, we explore this line of questioning through scenario-based interviews with 20 data-science practitioners. Specifically, we seek to understand their perceptions of using **data modification methods (DMMs)**, an umbrella term encapsulating pre-processing methods that directly alter the underlying training data to address a fairness issue before constructing a model. For instance, the DMMs we study aim to address fairness concerns by transforming training data according to fairness criteria (Zemel et al. 2013), re-weighting training data instances (Kamiran and Calders 2012), modifying the target labels for supervised learning (Kamiran and Calders 2012), or creating synthetic data. Though these methods fall under the pre-processing umbrella, characterizing them solely as pre-processing minimizes the distinction that we hypothesize exists for practitioners (Section 2). We were motivated to study DMMs as a way to understand why fairness interventions and data-science work seem to focus disproportionately on models and their outputs, rather than data (Jacobs and Wallach 2021; Rismani et al. 2025; Simson, Fabris, and Kern 2024; Sambasivan et al. 2021).

To motivate why DMMs present a uniquely revealing lens to analyze this question, consider a scenario in which a predictive model uses data drawn from a city known to be impacted



id	HOLC_area	loan_outcome
54321	D74	Denied

(a) HOLC redlined area D74 from Chicago's South Side. Source: Mapping Inequality (Nelson et al. 2021).



id	redlining_flag	loan_outcome	modified_outcome
12345	True	Denied	Originated (Approved)

(b) 2020 census tracts aligning with area D74. Lighter pink tracts indicate a proportionally higher non-White population.

Figure 1: An example of **changing target labels** in training data in response to discrimination from historical redlining.

by exclusionary zoning and redlining (Taylor 2019; Aaronson, Hartley, and Mazumder 2021; Derenoncourt et al. 2024; Nelson et al. 2021). Redlining in the U.S. refers to the practice of denying or restricting access to financial services (e.g., insurance or loans) for certain communities, typically based on racial characteristics. Being denied access to housing in this manner has contributed to a widening racial wealth gap among a broader set of harms (Derenoncourt et al. 2024).

The degree of intervention needed to address issues like this in data depends on the theory of harm and responsibility. One ethical theory might point to demographic disparities and hold the practitioner responsible for minimizing those disparities. Another theory might focus more directly on specific people who were denied loans or who were dissuaded from even applying for those loans, as well as on the specific properties whose values were affected by redlining. Under the first theory, training a model with equalized odds as a constraint might be appropriate. Under the second, one might need to target interventions to specific people or specific groups who may have been directly affected by the decision.

Figure 1 demonstrates how one DMM—changing target labels—could perhaps address that issue. By changing labels in areas known to be affected by historical discrimination, the practitioner explicitly defines a relevant affected population and remediating action, which may span from limiting future harms to reparative fairness (Davis, Williams, and Yang 2021; So et al. 2022). Using a demographic-based model criterion would also imply a set of judgments, though the link between the theory of responsibility and its operationalization in the decision system would be much harder to trace.

While this example describes one theoretical understanding of DMM usage, practitioners could plausibly perceive these methods in diverse and contrary ways. To this end, we asked 20 participants to consider how they would address potentially problematic mortgage data for an automated predictive model (Section 3). In response, participants articulated a number of fairness concerns within the data, recognizing many plausible theories of harm. However, the majority of participants considered modifying data to violate values fundamental to their identity as data scientists and thus rejected DMM usage. Accepting that some fairness issues in the data were reproduced in the model, however, was seen as a necessary evil in satisfying the norm of not modifying the data. At the core of this perspective was a desire to maintain supposed objectivity and to align with data-science norms (Section 4.2). This discomfort extended beyond the personal; practitioners anticipated barriers in communicating the impact and rationale of modifying data to stakeholders, highlighting a lack of infrastructure to measure and describe fairness issues in data.

Our findings point towards a need for data-science norms to evolve to highlight the ways data scientists exercise agency. In particular, data science often treats data as entirely objective—"ground truth"—and de-emphasizes the ways in which data is shaped through interpretation by the actors collecting, processing, and analyzing the data. From this perspective, overt interventions, such as DMMs, are impermissible. However, as we find in our study, this perspective presents a barrier to addressing significant structural fairness issues and can contribute to maintaining the status quo.

The rest of the paper is organized as follows. Section 2 contextualizes DMMs within the space of fairness interventions and Section 3 describes our methods. Section 4 presents how participants articulated fairness concerns within our scenario and how they assessed DMM efficacy. Section 5 discusses and interprets the implications of our findings, while Section 6 identifies design opportunities for DMM adoption.

2 Background and Related Work

2.1 Fairness Interventions

Fairness interventions necessarily involve modification from the "ground truth" of the data. They are often characterized by the phase in which they intervene: **pre-processing** methods, including DMMs, intervene on training data (Zemel et al. 2013; Feldman et al. 2015; Kamiran and Calders 2012; van Breugel et al. 2021; Roh et al. 2023; Jain et al. 2024; Calmon et al. 2017; Wang et al. 2025); **in-processing** methods intervene by constraining or regularizing the model, or by specifying a learning algorithm to induce specific changes (Wang et al. 2022; Agarwal et al. 2018; Zafar et al. 2019; Grgić-Hlača et al. 2018; De Paolis Kaluza et al. 2025; Luo et al. 2025); and **post-processing** methods adjust a trained model to satisfy distributional fairness properties (Hardt, Price, and Srebro 2016; Dwork et al. 2012; Pleiss et al. 2017; Hansen et al. 2024; Petersen et al. 2021; Chouldechova 2017; Alghamdi et al. 2022).

We define DMMs as methods that modify data to carry out some fairness goal. DMMs may be motivated by specific fairness metrics like equalized prediction rates or causal fairness criteria (Zemel et al. 2013; Feldman et al. 2015; van Breugel et al. 2021; Chen and Zhu 2025). They can also modify data in a variety of ways. For example, DMMs might adjust specific data instances (Feldman et al. 2015), produce alternative representations of the data (Zemel et al. 2013), reweight the data (Kamiran and Calders 2012), or generate synthetic data (van Breugel et al. 2021).

We focus on DMMs in this study because they are more overt, "bias transforming" methods (Wachter, Mittelstadt, and Russell 2021). Unlike specifying a different objective function or adding a regularization parameter, modifying the data runs directly counter to the notion that the data is "ground truth" (Muller et al. 2019; Wachter, Mittelstadt, and Russell 2021; Lakkaraju et al. 2017). In contrast, post-processing methods can be oblivious to individual features and "aim to avoid" what they deem as "subjective interpretation" (Hardt, Price, and Srebro 2016). As we explore in Section 4.2, this property aligns with the norms expressed by the practitioners we interviewed, while the properties of DMMs do not.

In the data-management literature, approaches that address invalid, duplicate, or missing data as violations of integrity constraints have been formalized as *data repair*, though they resemble data modification in function (Rekatsinas et al. 2017; Krishnan et al. 2017; Liu et al. 2021; Gilad, Deutch, and Roy 2020; Tae et al. 2019). While those approaches address data cleanliness, others have explicitly oriented data-repair methods towards addressing fairness concerns (Salimi et al. 2019; Salimi, Howe, and Suciu 2020). Balayn, Lofi, and Houben (2021) survey this intersection of data repair and fairness, arguing for new data-centric approaches to investigate how database management and traditional fairness techniques can be synthesized to overcome the shortcomings of existing algorithmic-focused methods. Though data repair can be repurposed for fairness issues, we argue that differences in practitioner intent give reason to formalize a distinction between DMMs and data repair writ large.

2.2 Studies of Fairness Interventions

In practice, data scientists operate under a number of constraints that affect fairness outcomes (Smith et al. 2025; Rakova et al. 2021; Holstein et al. 2019). Several studies have demonstrated how pre-processing and data work can be neglected in data-science workflows, often to the detriment of fairness outcomes (Sambasivan et al. 2021; Harrison et al. 2024; Simson, Fabris, and Kern 2024; Paullada et al. 2021; Schelter et al. 2019; Guha et al. 2024). This prior literature led us to hypothesize that practitioners may be entirely ignorant of using pre-processing methods for fairness.

To the best of our knowledge, no prior work has specifically studied practitioners' perceptions of DMMs. Furthermore, there have been only limited empirical investigations into the effects of pre-processing decisions on fairness outcomes (Simson, Fabris, and Kern 2024; Harrison et al. 2024; Schelter et al. 2019). Our examination of pedagogical materials (Barocas, Hardt, and Narayanan 2023) and widely cited surveys (Verma and Rubin 2018) suggests the primacy in the fairness research community of in-processing and post-processing fairness definitions like equality of opportunity (Hardt, Price, and Srebro 2016) and demographic parity (Dwork et al. 2012). Surveys of evaluation practices and fairness methods similarly find a disproportionate focus on models and their outputs relative to data-centric methods (Rismani et al. 2025; Black et al. 2023; Schelter et al. 2019; Hort et al. 2024). Our work primarily questions the motivating factors behind this apparent predisposition across research, education, and data-science practice.

3 Methods

Prior to this study, it had been an open question how prevalent DMM usage is among ML and data-science practitioners, as well as what factors influence DMM usage (or non-usage). Hypothesizing that methods like DMMs that overtly modify data are not widely used by practitioners, we primarily sought to identify rationales for why this is the case. We also sought to uncover the characteristics of practitioners or DMMs that influence decisions to use—or not to use—DMMs.

3.1 Participants and Recruitment

We interviewed 20 data-science practitioners with work experience in data pre-processing or analysis, as well as experience in training, deploying, or monitoring ML models. Our eligibility criteria included people who live in the U.S. with experience working as a "data scientist, data analyst, data engineer, machine learning engineer, machine learning researcher, or similar position." We did not use familiarity with algorithmic fairness as a criterion in recruitment, though eight participants described experiences with fairness in their own work. Throughout interviews, participants reflexively described how their work experiences influenced their assessment and decisions in the hypothetical scenario. After each interview, we discussed major themes raised by the participant with the entire research team, enabling us to stop interviewing once we reached theoretical saturation, meaning we were no longer hearing substantially new themes from recent participants (Saunders et al. 2018; Blandford 2013).

Table 1: Participants' job titles and their relevant work and educational background.

Participant	Field	Job Title	Gender	Age	Race
P1	Health	Data Scientist	Man	32	Black
P2	Academic Research	Data Scientist	Man	34	White
P3	Sociology	ML Engineer	Woman	40	White
P4	Digital Humanities	Researcher	Non-binary	25	—
P5	Public Health	Data Scientist	Man	44	White
P6	Health Tech	Data Analyst	Woman	29	White
P7	Health Tech	Data Scientist	Man	28	White
P8	Academic Research	PhD Student	Man	29	Asian
P9	Freelance	Data Scientist	—	—	—
P10	Health Research	Data Engineer	Woman	35	White
P11	Academic Research	Researcher	Man	43	White
P12	Industry	Data Scientist	—	—	—
P13	Consulting	Data Scientist	Man	28	White
P14	Public Policy	Data Scientist	—	—	—
P15	Freelance	ML Engineer	Man	—	—
P16	Industry	Data Scientist	Woman	50	Asian
P17	Industry	Data Scientist	Woman	32	White
P18	Public Policy	Data Scientist	—	—	—
P19	Freelance	ML Engineer	Man	29	Black
P20	Industry	ML Engineer	Man	23	Asian

'—' indicates the participant chose not to disclose this information

As detailed in Table 1, most participants worked as data scientists. They had varied educational backgrounds—primarily in traditional computer science or data science, though some participants' primary training was in sociology or public policy. Participants were recruited through the freelance platform Upwork and local tech industry hack nights. We prioritized recruiting participants with relevant work experience from a diverse set of backgrounds. Interviews typically lasted between 60 and 90 minutes. We compensated participants $50 via Upwork or an electronic gift card.

3.2 Task Context

We chose to study DMMs in a hypothetical scenario because we believed that participants would be unfamiliar with the use of these methods in a fairness context, resulting in participants disregarding DMMs or finding them counterintuitive to implement (Rosson and Carroll 2007; Carroll et al. 2026). While scenarios are necessarily distinct from actual work practices, they enabled us to ensure three key characteristics: the chosen context of mortgage lending in the U.S. would sufficiently motivate fairness concerns; participants would not be bogged down in implementation details; and the interview would focus on the ethical implications of DMM usage.

The interviews focused on a scenario in which participants imagined building an automated mortgage decision-making model and were in the process of exploring and identifying discriminatory trends within the 2019 U.S. Home Mortgage Disclosure Act data (Consumer Financial Protection Bureau 2019). The 2019 dataset is among the first to provide important predictive variables like debt-to-income ratio and combined loan-to-value ratio, as well as avoiding any Covid-pandemic-influenced confounders. We wanted participants to understand the national scope and also to direct their attention towards smaller geographic areas. Thus, many of the figures we showed participants contrast national trends with those in Chicago, a city with a persistent legacy of segregation (Loury 2023; Lutton, Fan, and Loury 2020; Nightingale 2012) and history of redlining (Taylor 2019; Aaronson, Hartley, and Mazumder 2021; Nelson et al. 2021; Derenoncourt et al. 2024). The context of the scenario had the advantage of engaging with interrelated questions of fairness and historical injustice while being open-ended enough for participants to identify a number of plausible hypotheses and interpretations.

3.3 DMMs

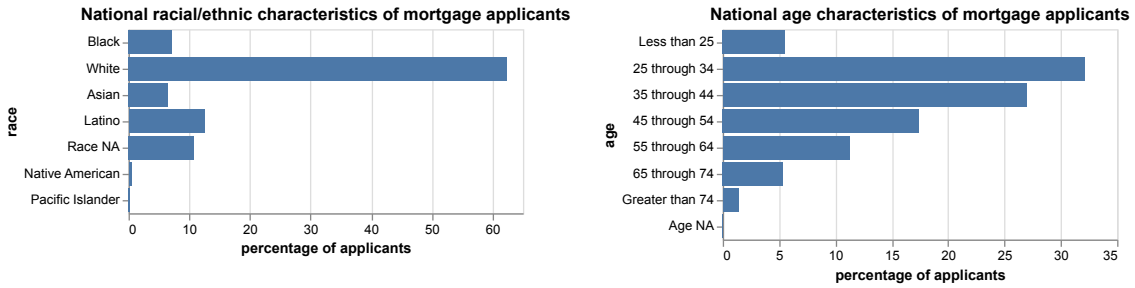
We presented participants with schematized versions of DMMs originating from academic literature. The DMMs spanned output modalities, such as adding feature columns or modifying existing data. We used high-level descriptions to avoid confusion about implementation details and to match the resources practitioners would likely access when coming across the methods. Specifically, the verbiage we used to describe each method was sourced from how they were described in their original manuscripts, their descriptions in fairness toolkits like AI Fairness 360 (Bellamy et al. 2018), or generalized descriptions of the method. Our descriptions of the four DMMs we presented to participants were accompanied by figures demonstrating what those modifications might look like in practice. Our supplementary materials[1] contain the exact text we used verbatim.

Inspired by Zemel et al., we described a pre-processing method that '**transforms**' the training data, obfuscating information about protected group status (e.g., race) while retaining its utility as a data source and achieving group fairness (Zemel et al. 2013). '**Re-weighting**' assigns numeric weights to different data instances to penalize or reward particular model behaviors (Kamiran and Calders 2012; Krasanakis et al. 2018; Chai and Wang 2022; Yan, Seto, and Apostoloff 2022). Kamiran and Calders propose a method to directly '**change target labels**' in order to induce the trained model to satisfy a particular distributional property (Kamiran and Calders 2012). The method of changing labels we described tried to satisfy equal approval rates for disadvantaged and advantaged borrowers. '**Synthetic data**' involves artificially creating additional data as either a partial or complete replacement for real data. The complexity of the synthetic process can range from over-sampling rare data (Chawla et al. 2002) to learning generative models (van Breugel et al. 2021). In interviews, we depicted the latter approach, though the majority of participants described familiarity with the former.

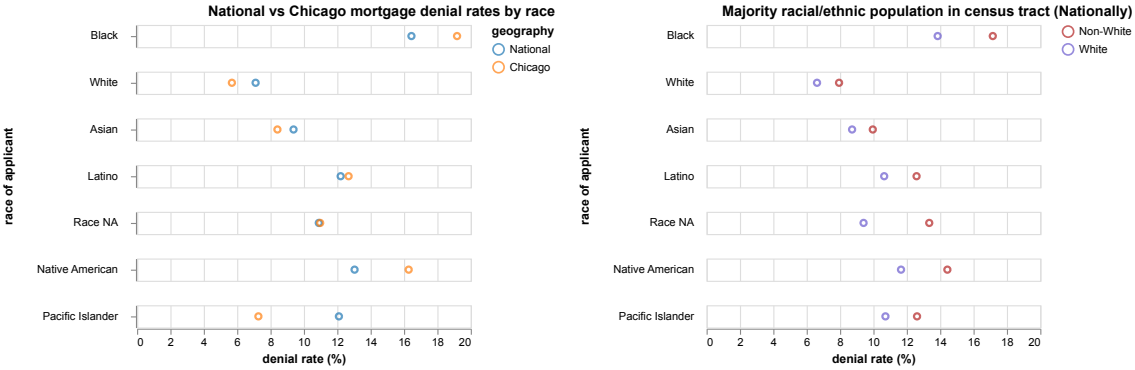
3.4 Interview Structure

First, we described the task context and the participant's role in the scenario (Section 3.2) using visual aids like Figure 2 to imitate exploratory data analysis. Then, we asked participants to describe fairness concerns they expected might exist in the data, as well as what they would do to address those problems. In the second phase of the interview, participants were asked how they would use "methods that augment or modify the data" to address the issues they described, if they were to use these types of methods at all. Next, we elicited participants' reactions to the four DMMs in randomized order. For each DMM, we read a description of the method (Section 3.3) and showed an accompanying diagram demonstrating its effects (see our supplementary materials). We next asked participants to describe their familiarity with the method, to evaluate both

[1] https://www.blaseur.com/papers/aies26-supplementary.zip



(a) Demographic and income characteristics of mortgage applicants in the 2019 Home Mortgage Disclosure Act data.



(b) Mortgage denial rates by race/ethnicity for applicants across the U.S. relative to those in Chicago (left), as well as denial rates by race/ethnicity for applicants in majority White and non-White census tracts (right).

Figure 2: Figures shown to participants contrasting national and Chicago-specific mortgage trends.

its perceived effectiveness and the fairness/ethics of using it, and to think of what tools or information they would need to apply it in practice. Finally, we prompted participants to evaluate and describe their criteria for choosing to use a DMM, as well as to envision how technical tools could assist in implementing or describing DMMs.

3.5 Data Analysis

We analyzed interview transcripts using a reflexive thematic analysis approach, following best practices for analysis (Braun and Clarke 2019, 2006; McDonald, Schoenebeck, and Forte 2019). The first author conducted all interviews and met regularly with the other authors to discuss themes emerging from interviews, as well as to identify areas to emphasize in future interviews. To create an initial codebook, the first and second authors conducted open coding of three interviews and collaboratively combined their codes. The remaining interviews were analyzed independently by two coders using this codebook as the basis. That said, as interviews progressed, the codebook was updated and modified to account for new themes. The full research team met regularly to discuss the iterative refinement of the codebook, as well as to discuss relationships between themes in the codebook. As new codes were added or existing codes were consolidated, the two coders revisited previous interviews to re-code them. The codebook was organized into higher level themes, broadly corresponding to this study's research questions. After iteratively developing the codebook, we identified 37 distinct codes, grouped into three broad thematic areas. Our supplementary materials include our full codebook.

We emphasize that given the qualitative nature of this work, readers should not try to overgeneralize the proportions of participants who express each theme we present (Braun and Clarke 2019; Blandford 2013). When reporting participant quotes in Section 4, we lightly edit participant quotes for clarity (e.g., by removing insubstantial words like "umm" and "like") without changing their meaning.

3.6 Research Ethics

We received approval for this study from the lead institution's IRB. We did not request any personal identifiers beyond those needed for scheduling the interview. While we recorded audio of the interviews, we took a number of steps to minimize privacy risks, including by redacting identifiers (e.g., names of workplaces, institutions they attended, or cities they work in). Once redacted, authors accessed the transcript via a secure cloud server. All participants verbally consented to being recorded prior to beginning the interview, in addition to reviewing a formal consent form.

4 Results

In this section, we describe participants' conceptions of fairness issues in the scenario, how they thought about addressing them, and their concerns and needs around using DMMs. Throughout, we use the following emphasized descriptors to indicate the number of participants who expressed each

theme: *a few* (fewer than 5 participants), *some* (5-8), *about half* (9-11), and *majority* or *most* (more than 12). We also indicate when participant quotes reference their experiences, previous work, or professional identity (as opposed to the scenario we provided in the study) with the star symbol in the following style: "P99⋆." When quotes were solely in response to the hypothetical scenario, they are left unadorned.

4.1 Articulation of Fairness Concerns

A majority of participants recognized multiple ways in which unfair discrimination might result in incomplete or insufficient data. Housing discrimination can manifest in a variety of ways, including explicit redlining by lenders, predatory inclusion, or systemic issues related to discrimination in other domains (Besbris et al. 2022; Taylor 2019; Edwards 2019; Aaronson, Hartley, and Mazumder 2021; Nelson et al. 2021; Derenoncourt et al. 2024). Issues that participants identified ranged from conventional ML problems (e.g., data imbalance, distribution shift) to more scenario-specific issues like geographic disparities and biased data collection.

The majority of participants noted the complexity of fairness issues in our mortgage scenario due to enmeshed social and political factors. For example, P8 said, "*I feel like a lot of this stuff is sort of a political decision more than it's even a data one.*" Some participants took issue with the problem formulation, such as by critiquing the choice of target variable (loan approval): "*As a sociologist... I think that loan outcome of actually whether someone gave the loan is toxic to the whole project personally*" (P3⋆). That particular participant noted that data regarding the history of loan approvals and denials did not capture relevant information about loan performance. Other times, participants questioned the automated ML aspect of the scenario.

The majority of participants hypothesized about biases stemming from the way the data was collected. Around half of the participants also hypothesized about the effects that specific lending organizations and even "*the demographics*" (P13) of loan officers might have on outcomes, hypothesizing that particular organizations or loan officers may be specifically biased. Several participants were curious about the frequency with which applicants "*declined to identify*" (P4⋆) their race, proposing that selective decisions about whether to self-report race may explain patterns of missing data. Around half of participants speculated about problems relating to dataset imbalance, and a few raised concerns about how distribution shifts due to differing time periods.

4.2 Recurring Concerns about DMMs

Despite the applicability of data modification and data itself to the fairness concerns participants articulated, participants emphasized how data modification—irrespective of any specific DMM—violated an implicit set of **canonical values** of data science practice. The values we observed included a need to **comply with normative practices** among other practitioners, experts, and stakeholders, as well as a need to **maintain objectivity**. Prior work has critiqued the narratives these values form (Harding 1995; Wachter, Mittelstadt, and Russell 2021; Lakkaraju et al. 2017; Green 2021), and in this work we show how those values arise in response to DMMs and constrain practitioners' range of fairness interventions. As these values appear, we draw attention to them in **bold**.

DMMs were rejected by the majority of participants because they felt that using DMMs would be akin to changing reality. Discomfort with changing reality was sometimes stated as a general principle like "*it's important to model reality*" (P6⋆) or "*because at a certain point, you're just rearranging the data to fit your beliefs, which is definitely something a data scientist shouldn't do... And it's like, how much do you want to play God in controlling the dataset, you know*" (P19⋆). Other times, specific DMMs were viewed as violating data science and ML norms, rendering them too invasive for use: "*It makes me a bit worried because I always...I guess it's just the data science machine learning kind of thing where if you mess with the ground truth labels, like, that's a problem*" (P5).

Worries about "*rearranging data to fit your beliefs*" sometimes halt engagement with DMMs without substantial justification. For example, in response to changing labels, P16⋆ short-circuits further consideration of the method: "*We don't do that... We haven't done that before. No reason to do this.*" For a few other participants, what they viewed as ethical overlapped with others' descriptions of practitioner norms, indicating the interdependence of personal and professional ethics: "*I think this is the ethical side of me talking, but I'm more neutral towards this method [synthetic data] because it is not tweaking the existing data*" (P14⋆). Participants repeatedly made clear that the value of "reality" or **ground truth** data supersedes other considerations, including fairness.

In this way, DMMs make evident the tensions between modifying data to achieve fairness goals, what practitioners view as ethical, and data science norms. On the one hand, participants recognize disparities and fairness issues across many areas. On the other hand, altering the data that reflects those issues would **violate normative practices**. Crossing that invisible boundary is potentially powerful, yet could leave practitioners precariously responsible for any negative effects, echoing concerns about practitioner ability to implement fairness interventions in practice (Deng et al. 2022, 2023; Deng, Barocas, and Wortman Vaughan 2025; Widder et al. 2023; Rakova et al. 2021; Cunningham et al. 2025).

The tensions practitioners face between ethics, normative practices, and addressing fairness issues have no straightforward resolution. In balancing those tensions, about half of participants wonder, for example, "*is it possible to take away human judgment?*" (P15). The act of avoiding seemingly subjective judgments is one way participants **maintain the facade of objectivity**. However, this approach de-emphasizes the ways in which data scientists already make subjective judgments and assumptions (Mothilal et al. 2025; Silberzahn et al. 2018). Ultimately, those participants seemed to dismiss the complexity of making judgments about fairness and rejected methods that they interpreted as changing reality.

About half of participants sought to conduct robust statistical evaluation in order to feel comfortable proceeding. This preference appeared to address concerns about the perception that DMMs were not rigorous, or that the participant would be **overstepping their role as a data scientist**. These participants sought statistical confirmation of their hypotheses about

how unfairness manifested in the data to "*see if it's a data issue or if that's really reflective of reality*" (P6⋆). For example, P14⋆ wanted "*to check basically if there were heterogeneous impacts on different types of populations*." Participants described moving to "*the next steps of the [process]*" to show "*that there is in fact discrimination that can't be accounted for in any other way*" (P13) rather than further consideration of modifying data. While finding statistical confirmation would help to justify the use of DMMs or to provide a ritualistic assessment of the "ground truth," for most participants it seemed to stall ideation on ways to address fairness in favor of more familiar implementation concerns. In the event that their analysis found that racial disparities (Figure 2b) were just reflective of reality, participants like P6 felt they would be unable to dispute that truth.

When ideating about what they would need to use a DMM in practice, most participants felt they would need an expert to endorse their approach. Within the constraints of an organization, participants viewed themselves more as instruments that distill recommendations and context into results than decision makers. For example, P2⋆ asked, "*What do other people say to do... in the industry domain?*" P14⋆ discussed their practice of finding "*research papers that have proved that it works... have tested that they are ethical and are fair*."

In apparent contradiction to the highly iterative and exploratory nature of data science work (Zhang, Muller, and Wang 2020; Alspaugh et al. 2019; Muller et al. 2019), we observed that participants were reluctant and uncomfortable experimenting with novel methods without prior or external validation. Participants discussed **shifting this responsibility** onto researchers in "*sociology journal[s]*" (P7) and "*experts that have all this domain knowledge... versus a data scientist that might have been at a company for only two or three years*" (P12⋆). In this way, we observe that **complex** fairness objectives may be deprioritized to accommodate the status quo: "*I would want to know [how] actuaries at all the best places do stuff*" and to include as an addendum "*whatever sort of quotas on the back end that I actually thought were consistent with a societally good outcome*" (P8). Participants return to the status quo because those methods are in alignment with **normative practices**, while DMMs are not.

By demarcating fairness as a secondary objective, participants **maintain objectivity** and **shift responsibility** onto others. P19 thought modifying the data was **dishonest** and would be like getting "*the answer the client wants to hear,*" whereas they were "*more interested in getting the answer that the data is telling me*." With DMMs in opposition to their responsibilities, it makes sense that a majority of participants believed their primary duty was to assess "*first and foremost... accuracy*" (P9). P9⋆ further said, "*My main job is not to... equal the foreclosure rate [between demographics],*" but disclaimed that "*if there's some legal or some ethical issues there, I'm going to point them out. I'm going to **let somebody else** make that decision*." Reinforcing the primacy of post-hoc evaluation, P16 believed model evaluation was "*the only way that you can truly tell if [a DMM] is useful or not*." They believed data issues were limited to "*data drifting*" and "*data quality*" (P16). Other participants described similar processes where they would look "*at the outcomes and see which one produce[s] the least amount of bias*" (P5). When participants emphasized their role, they tended to emphasize the importance of **neutrality** and **objectivity** over addressing fairness concerns, even when their actions would replicate previous unfair decision-making.

Several participants expressed fear and uncertainty about meddling in consequential decisions. Consider P17, who was cognizant of the harms in accepting previous decision-making, yet cautious of the strategies presented. She said, "*I think a lot of care and thought would need to be put into how those labels are changed... I could see it working, potentially, but, that's a big leap to make*." Participants implicitly described how the gravity of the mortgage context demands a distinct frame of analysis to match, which does not align with principles around **simplicity**. One participant felt that modeling a "*serious topic*" like mortgage lending would require more thorough analysis than would be needed to model lower stakes outcomes that "*maybe [don't] need that sort of scrutiny*" (P14). Similarly, P4⋆ compared the mortgage context to a previous experience where they experimented with synthetic data generation to assist with an image classification problem: "*An artwork is very important, but an artwork is not somebody's house. Like there's a part of me that's like if you're working with data that will affect people's lives to this matter like I'm less inclined to consider generating synthetic data*" (P4⋆). While P17 and P14 felt that analyzing the effects of DMMs could introduce significant uncertainty, P4 was reluctant to even experiment with using a method that did not feel appropriate to the situation.

4.3 Familiarity and Assessment of DMMs

Participants had only limited familiarity with the DMMs we discussed in the study; none had encountered them as fairness interventions. Their perception of the DMM, and whether it fit with the participant's self-conception as a data scientist, played an important role in determining whether the participant viewed that DMM as a valid technique. Participants were open to using DMMs when they were perceived to address tacitly **objective** values like dataset imbalance.

A majority of participants were familiar with synthetic data and with re-weighting. Some participants were familiar with the transformation method, and only a few were familiar with changing target labels. Participants commonly compared DMMs to techniques used in traditional data science and ML. For example, P5⋆ described their prior experience with synthetic data as follows: "*It sounds like SMOTE. I've used it... but I didn't like it*." When asked about re-weighting, participants noted its ability to address dataset imbalance concerns. For example, P12⋆ explained, "*Yeah, I've used weights before. It's super common to do... For example... you would want to have a higher weight for fraud versus non-fraud just because there's a lot less fraud out there*." The transformation method was compared to methods like PCA—"*I know... the principal component analysis where you can like collapse multiple columns into one factor*" (P6⋆)—or to normalization methods—"*I've used scaling features, which kind of looks like that's what income is*" (P9⋆).

Of the four DMMs in any context (i.e., not just fairness), participants were the least familiar with changing labels. Only

a few participants stated that they were familiar with the idea, and only outside the context of fairness. Some participants expressed that "*for some reason, [changing labels] seems like the most naive approach*" (P2), that it "*seems like the least sophisticated way to do it*" (P9), and that "*it kind of feels like you're trying to build a model using a model that you just created... It's a bit confusing*" (P13). On the other hand, transformation was perceived as unnecessarily complex. Around half of participants anticipated issues with explaining this method to other stakeholders in their organizations.

Methods like re-weighting and synthetic data were acceptable to most participants because they felt they could address ostensibly value-neutral concerns like data imbalance or dataset drift. Unlike transformation and changing labels, these two DMMs were generally accepted by participants. About half of participants thought re-weighting and synthetic data could address some issues in the data: "*This method feels better than [changing labels]... I'm definitely more familiar with it and it just feels more robust in a way*" (P5). Generally, participants were familiar and immediately comfortable using re-weighting or synthetic data to handle data imbalance or distribution drift problems. Referencing re-weighting, P1 explained, "*If there's a data imbalance, then yes, it's fair to try to learn more from the data you have less of.*" Some participants discussed re-weighting and synthetic data as emphasizing particularly salient or rare data points, like this participant who spoke on the basis of prior experiences with synthetic data: "*I can certainly understand theoretical justifications... we've defined a portion of this high dimensional space where we can interpolate and... get something new out of the same class that occupies a niche in that space*" (P8⋆). In other contexts, P8 had expressed that "changing reality" was wrong, yet here described the utility in synthetically generating new data. This apparent contradiction demonstrates how perspectives on DMMs can shift across contexts and practitioners.

When considering whether to use DMMs, a few participants considered whether it fit with their self-conception as a data scientist or statistician. This facet was particularly discernible when participants reacted negatively to a DMM: "*The statistician in me doesn't like weighting based on the outcome... I just have a gut reaction to that... So I'm skeptical*" (P17⋆). P17 and others' remarks put DMMs in direct opposition with their professional self-conception, highlighting misalignment between implementation and **normative practices**. In Section 5, we discuss how this reaction is indicative of a pattern of tensions between DMMs and canonical values.

4.4 Participants' Alternatives to Address Fairness

Alternative approaches participants proposed rarely involved any direct modification of existing data, instead favoring the collection of more and additional types of data. Other participants suggested non-technical approaches like community banking or outreach programs. Illuminatingly, data scientists would typically not be responsible for those approaches.

Many of participants' proposed approaches circumvented the need to use a DMM or contend with the validity of the existing data. A majority of participants said they "*would obviously try to get some newer data and be less reliant on the older data side of things*" (P12) or add more contextual information like "*different deprivation indexes and income data and all sorts of things based on location*" (P6). Indeed, participants saw collecting more data as a solution to data imbalance just as well as socio-political issues. However, a few participants contested its utility. For instance, P9 highlighted financial limitations, saying that collecting data would help "*very little, but I'd be open to trying, especially if it's on somebody else's dime.*" P6⋆ specified that the challenge is acquiring the "*right pieces of data*" without "*[wasting] all your time looking for a million different data fields and then end up finding out that none of them are relevant anyway.*" In a precarious balance of competing time and financial priorities, collecting more data may or may not pan out to meaningful benefits, especially considering that collection of more data from the same or similar data-generating processes will contain similar issues as the existing data.

Some participants proposed non-technical interventions that employed a "*policy lens... to get in deeper... when the numbers are really not telling a story*" (P14⋆). Participants described re-purposing hypotheses for DMM intervention to initiate policies like "*programmer outreach to the underrepresented communities to say 'like you should apply for this loan'*" (P2). Participants also emphasized a "*qualitative component... like narratives from the field*" (P14⋆) to identify alternative solutions for mortgage disparities. As P19 described, "*Maybe it's a low-income area, so like introducing more jobs in the area. Trying to bolster the economy of the area in some way.*" P11⋆ staunchly felt that "*data transformation doesn't impact that social transformation,*" thus a "*system-wide shift*" was needed. They specifically wanted "*options like cooperatives and funding that [are] based on collective ownership of properties.*"

While these reactions recognized the difficulty of meaningfully addressing issues like historical, systematic discrimination, they also can be understood in light of participants' reluctance to take personal responsibility for addressing fairness issues. Most non-technical solutions would inherently not be solutions implemented by data science practitioners. Furthermore, several non-technical solutions would require significant investments of time and money to carry out. While many fairness interventions and tools aim to 'solutionize' social issues, practitioners and interventions alone may not have the power or ability to effect "*social transformation.*"

5 Fairness and Data Science Norms

We find a fundamental contradiction embedded within data science-practice. Participants were unwilling to engage in overt intervention in the form of DMMs, but were more than willing to engage in other (similar) forms when provided a suitable story around **objectivity**. This story is reflective of deep-rooted and widely accepted values within the data science community. We argue that for certain contexts, like mortgage lending, some fairness goals cannot be achieved without challenging or considering the impact of these values.

5.1 Practitioners and Ethical Decision-Making

There is an apparent contradiction in participants' reluctance to engage in data modification and their preferred approach

of collecting more data. In both cases, they would engage in actions with intent to produce more desirable outcomes. However, the action where their agency is discernible is seen as less permissible. Prior work explains this contradiction through the big-data paradigm, which is the belief that large volumes of data provide an assumption-free and thus objective view of the world (van Dijck 2014; boyd and Crawford 2012). In this point of view, collecting more data is objective and simple; other interventions are not.

However, we echo arguments that this viewpoint is inaccurate. Data scientists exert significant agency throughout the data-science process, influencing how the data is understood and interpreted. Participant attitudes towards DMMs closely resemble what feminist scholars have coined the 'god-trick' (Haraway 1988). Participants sought to minimize the perception that they were engaged in actively interpreting and therefore shaping their model and its predictions, presenting it as an objective view from nowhere. As noted by Green, the interpretations practitioners engage in are political in the sense that they—intentionally or not—make allocative decisions about social goods (Green 2021). By renouncing their agency, practitioners attempt to maintain their objectivity, but ultimately constrain the permissible set of interventions. A key challenge is to develop practices, norms, and tools that encourage data science practitioners to acknowledge their interpretive agency.

5.2 Implications for Data Science Practice

We interpret the reluctance to challenge the status quo through data modification as symptomatic of self-imposed limitations of data science practitioners' **canonical values** and practices. While we are not suggesting that data modification alone is a magic shortcut to equitable outcomes, it does give insight into how practitioners may perceive fairness in certain contexts: a series of non-technical judgments that alter reality in a way that is undesirable. Given this perception of data modification, it is little surprise that choices that oppose the status quo are rarely made (Wachter, Mittelstadt, and Russell 2021; Schwöbel and Remmers 2022; Abebe et al. 2020). To address this pattern of thought, future work should deliberately craft norms that support acknowledging the interpretive and partial nature of data-science work (Davis, Williams, and Yang 2021; Green 2022; Kasirzadeh 2022; Jacobs and Wallach 2021). Addressing fairness in practice requires expanding the scope of what "counts" as data science.

Collecting more data, for instance, seemed to be a common suggestion not because it will likely solve or address identified fairness problems—in all likelihood, collecting more data from the same data-generating process will reproduce the same fairness issues. Practitioners may be reluctant to view DMMs as acceptable because modifying data seems to prefigure or predispose a particular outcome. However, as a practitioner with reparative fairness goals, this line of thinking may be unproductive and possibly stifling. With respect to fairness, practitioners seemed to feel like their role was passive. In short, practitioners identify and share findings about bias, rather than address or challenge its presence.

By surfacing the barriers canonical values pose to fairness work, we do not argue that these values should be rejected entirely. Arbitrary algorithmic decisions are a real and consequential form of algorithmic harm (Eubanks 2018; Long et al. 2023). Collaboration with subject-matter experts and governance structures form an important part of defining responsibilities and goals in practice. Our point is that practitioners *already* have a great deal of agency. Data science ought to develop practices and norms that more clearly account for the full scope of fairness decisions. Although fully developing these norms and practices is beyond the scope of this paper, we argue that rethinking the philosophy of data science will be necessary to overcome the structural barriers to adoption of DMMs that we observed in this study.

5.3 Needs for Cross-Discipline Collaboration

Reformulating these norms will require collaboration that goes beyond the offloading of ethical considerations to other actors, a pattern found both in this and other studies (Diaz and Smith 2024; Sloane et al. 2022). In some cases, the collaboration may require data scientists to take steps that may initially appear to them to be wrong, forcing them to engage reflexively about their assumptions and implicit beliefs. In short, rather than fairness being just another constraint or optimization target, fairness requires that data scientists change the way they think about what they do.

Data science should look to engage with the fields of science and technology studies, design, and human-computer interaction when developing these norms. These fields have developed robust methodologies for characterizing and acting on conflicting understandings and needs. Changing how data scientists self-conceptualize will require both interdisciplinary research and integration into professional instruction. While a large undertaking, we believe that these steps will better position data science and data scientists to respond in socially beneficial ways to fairness and other challenges.

6 Supporting a Paradigm Shift in AI and DS

Building upon previous work studying practitioner interactions with algorithmic fairness (Deng et al. 2022; Richardson et al. 2021; Balayn et al. 2023; Harrison et al. 2024; Deng, Barocas, and Wortman Vaughan 2025; Mothilal et al. 2025; Dong et al. 2025), this work contributes a characterization of practitioner perceptions of DMMs that, among other things, reveals conflict and tension between canonical values in data science and data modification. These values contributed to many barriers for the adoption of DMMs when they were viewed as potentially useful, though many participants rejected DMMs outright simply because they modify data. While a full account of how data science must change is outside the scope of this work, we identify two critical design areas supported with evidence from participants.

6.1 Justifying Interventions

Though most participants expressed that a compelling story would be necessary to justify DMM usage to stakeholders, we instead see a need for increased transparency and documentation of all pre-processing decisions. Participants described needing a "*strong narrative for whatever you're choosing and why the data [modification] method*" (P7) and a "*defensible,*

scientifically grounded argument" (P3). Other times, participants described needing to "*understand [the process] enough to explain it to my boss*" (P2⋆). Participants felt they needed a scientifically grounded argument to use DMMs because they may not be perceived as objective. However, working with data in the U.S. mortgage context necessitates relying on biased historical decision-making practices and problematic measurements (Suresh and Guttag 2021; Jacobs and Wallach 2021; Taylor 2019).

Data pre-processing itself involves a number of often invisible judgments that affect models (Lucchesi et al. 2022; Harrison et al. 2024; Kasica, Berret, and Munzner 2023; Simson, Fabris, and Kern 2024; Schelter et al. 2019; Guha et al. 2024; Paullada et al. 2021). Thus, we see a need for tools and procedures to make data decisions visible and understandable. Though some tools for visualizing pre-processing steps exist, their capabilities for explaining *why* the particular steps were taken have been explored less well (Lucchesi et al. 2022). Moreover, the normative tendency to carelessly handle data challenges the notion that these practices will be readily adopted, suggesting that a broader push is required to foreground more intentional and transparent data practices (Simson, Fabris, and Kern 2024). Participants did suggest that practitioners would be amenable to this new part of their practice: "*If the tool could kind of explain why a certain transformation would kind of change the data… this is what the transformation does, this is how it changes the data, then I guess that would be useful to decide… if a transformation is sort of valid or not*" (P20).

At the same time, more open discussion of overt intervention in the data-science process may require communicative work outside of any specific tool or process. P17 surmised that "*people might object to… the idea of down-weighting certain people,*" but thought that this could be overcome "*if we spin it as… a bunch of [people] in Chicago being up-weighted as… examples that we want to correct for*." Their comments highlight the necessity of human judgment to avoid the trap of ill-placed **objectivity**. By making the entire pre-processing pipeline transparent, practitioners could communicate the effects of *all* judgments and assumptions in their work, not just those related to fairness goals.

6.2 Identifying Evaluation Criteria

Many participants expressed concern that changing the data would deprive them of the means to evaluate whether an intervention worked or not. We see a need to develop theory and tools for identifying evaluation criteria for interventions when the "ground truth" should not be trusted. Participants expressed the sentiment that the specific method employed—DMM or not—mattered less than the end result. In other words, the end justified the means: "*I would be okay using something like this, if it's giving me the outputs that I would want… You know, sometimes if it's working and you're trying to meet a deadline? Then you have to just let go of it. But ideally, you would want to know what is going on*" (P19⋆). Though this perspective may initially seem contradictory, it demonstrates the discrepancy between practitioner ideals and practitioner realities. Under ideal conditions, participants would have strong justification for all decisions. Under real constraints or pressures, however, they may settle for those that satisfy other values (e.g., simplicity).

Several participants identified this tension and struggled to conceptualize evaluation under certain DMMs: "*I just don't see a way to prove that this modified data set is accurate… we don't have any baseline comparison, we don't have any predictions that we're trying to generate or to compare to*" (P13). Changing target labels most evidently challenges a reliance on accuracy; if goals are oriented towards improving outcomes for certain people, then replicating all previous decisions is a sign of failure, not success. Complicating this perspective is the insufficiency of means to measure and evaluate data absent a model (Rismani et al. 2025). P7 echoed this sentiment, saying, "*This whole process is tricky in that the data doesn't have necessarily all the answers*." The lack of infrastructure with which to evaluate models trained on *modified* data presents an open challenge for practitioners, yet one that could be addressed through frameworks of recognizing harms and hazards (Ojewale et al. 2024; Rismani et al. 2023; Shelby et al. 2023). As it stands, familiar techniques are rendered ineffective. Consequently, participants are uncertain how to proceed. Reflecting this belief, participants' alternative approaches typically avoided direct modification of "ground truth" (Section 4.4).

6.3 Limitations

This study has a number of limitations typical of qualitative research. While participants in our study represented a variety of backgrounds and employers, we ultimately report on a convenience sample that is not necessarily representative of data-science practitioners at large or responsible data-science practitioners in particular. Interviews were only around one hour long, limiting the time participants had to ideate and restricting the level of detail given about DMMs and the data. Nevertheless, our approach enabled rich engagement with ethical and practical factors that would challenge making fairness judgments about data in practice. Our study was also limited to only one scenario and context to consider DMM usage. Though a public health scenario in a different country, for instance, may have generated different participant reactions, we see our results as capturing an important worst-case baseline for understanding the interactions of canonical data science values and fairness.

7 Conclusion

In this work, we conducted the first empirical study of how data-science practitioners approach data modification as a response to pressing and complicated fairness issues. We found that participants' shared canonical values constrain the interventions they deem valid and ethical, ultimately preventing many from considering DMMs. Beyond those tensions, participants anticipated a number of communication- and infrastructure-based barriers arising from DMMs. Though we construct concrete recommendations for these barriers, we hope our findings demonstrate a need for alternative approaches to fairness interventions in practice.

Positionality Statement

Our team is based in North America and is composed of computer scientists who work in specific sub-fields including AI and HCI. Informed by our collective HCI background, our approach was shaped by a focus on understanding the factors that influence practitioner perceptions of data modification. Similarly, as outsiders to industry practices, there was a gap in our understanding of normative practices within organizations. We acknowledge that our position as HCI researchers in this work lends a subjective focus on sociotechnical and interpersonal barriers, rather than legal or algorithmic barriers. To engage with themes of broader impact to fairness research, we aimed to critically investigate existing barriers.

Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 2140001. We thank the practitioners who took the time to participate in these interviews and the anonymous reviewers for their feedback and valuable suggestions.

References

Aaronson, D.; Hartley, D.; and Mazumder, B. 2021. The Effects of the 1930s HOLC "Redlining" Maps. *American Economic Journal: Economic Policy*, 13(4): 355–92.

Abebe, R.; Barocas, S.; Kleinberg, J.; Levy, K.; Raghavan, M.; and Robinson, D. G. 2020. Roles for Computing in Social Change. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 252–260.

Agarwal, A.; Beygelzimer, A.; Dudík, M.; Langford, J.; and Wallach, H. 2018. A Reductions Approach to Fair Classification. In *Proceedings of the International Conference on Machine Learning*, 60–69.

Alghamdi, W.; Hsu, H.; Jeong, H.; Wang, H.; Michalak, P.; Asoodeh, S.; and Calmon, F. 2022. Beyond Adult and COMPAS: Fair Multi-Class Prediction via Information Projection. *Advances in Neural Information Processing Systems*, 35: 38747–38760.

Alspaugh, S.; Zokaei, N.; Liu, A.; Jin, C.; and Hearst, M. A. 2019. Futzing and Moseying: Interviews with Professional Data Analysts on Exploration Practices. *IEEE Transactions on Visualization and Computer Graphics*, 25(1): 22–31.

Amnesty International. 2024. Denmark: AI-powered Welfare System Fuels Mass Surveillance and Risks Discriminating Against Marginalized Groups. Https://www.amnesty.org/en/latest/news/2024/11/denmark-ai-powered-welfare-system-fuels-mass-surveillance-and-risks-discriminating-against-marginalized-groups-report/.

Angwin, J.; Larson, J.; Kirchner, L.; and Mattu, S. 2016. Machine Bias. ProPublica. Https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Balayn, A.; Lofi, C.; and Houben, G.-J. 2021. Managing Bias and Unfairness in Data for Decision Support: A Survey of Machine Learning and Data Engineering Approaches to Identify and Mitigate Bias and Unfairness Within Data Management and Analytics Systems. *The VLDB Journal*, 30(5): 739–768.

Balayn, A.; Yurrita, M.; Yang, J.; and Gadiraju, U. 2023. "Fairness Toolkits, A Checkbox Culture?" On the Factors that Fragment Developer Practices in Handling Algorithmic Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 482–495.

Barabas, C.; Doyle, C.; Rubinovitz, J.; and Dinakar, K. 2020. Studying up: Reorienting the Study of Algorithmic Fairness Around Issues of Power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 167–176.

Barocas, S.; Hardt, M.; and Narayanan, A. 2023. *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Barocas, S.; and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review*, 104: 671.

Bellamy, R. K. E.; Dey, K.; Hind, M.; Hoffman, S. C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; Nagar, S.; Ramamurthy, K. N.; Richards, J.; Saha, D.; Sattigeri, P.; Singh, M.; Varshney, K. R.; and Zhang, Y. 2018. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. arXiv:1810.01943.

Besbris, M.; Kuk, J.; Owens, A.; and Schachter, A. 2022. Predatory Inclusion in the Market for Rental Housing: A Multicity Empirical Test. *Socius*, 8.

Bird, S.; Dudík, M.; Edgar, R.; Horn, B.; Lutz, R.; Milan, V.; Sameki, M.; Wallach, H.; and Walker, K. 2020. Fairlearn: A Toolkit for Assessing and Improving Fairness in AI. *Microsoft, Tech. Rep. MSR-TR-2020-32*.

Black, E.; Naidu, R.; Ghani, R.; Rodolfa, K.; Ho, D.; and Heidari, H. 2023. Toward Operationalizing Pipeline-aware ML Fairness: A Research Agenda for Developing Practical Guidelines and Tools. In *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Blandford, A. E. 2013. Semi-Structured Qualitative Studies. *Interaction Design Foundation*.

boyd, d.; and Crawford, K. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society*, 15(5): 662–679.

Braun, V.; and Clarke, V. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology*, 3(2): 77–101.

Braun, V.; and Clarke, V. 2019. Reflecting on Reflexive Thematic Analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4): 589–597.

Burgess, M. 2023. This Algorithm Could Ruin Your Life. Wired. Https://www.wired.com/story/welfare-algorithms-discrimination/.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. *Advances in Neural Information Processing Systems*, 30.

Carroll, J. M.; Jo, J.; Kim, J.; Lin, Y.-F.; and Chen, E. 2026. What Happened to Scenario-Based Design in HCI?: A Scoping Review. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, 1–29.

Chai, J.; and Wang, X. 2022. Fairness with Adaptive Weights. In *Proceedings of the 39th International Conference on Machine Learning*, 2853–2866.

Chawla, N. V.; Bowyer, K. W.; Hall, L. O.; and Kegelmeyer, W. P. 2002. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, 16: 321–357.

Chen, S.; and Zhu, S. 2025. Counterfactual Fairness Through Transforming Data Orthogonal to Bias. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 239–249.

Chouldechova, A. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big data*, 5(2): 153–163.

Consumer Financial Protection Bureau. 2019. HMDA - Home Mortgage Disclosure Act. https://ffiec.cfpb.gov/.

Cunningham, J. L.; Shao, K. Z.; Pang, R. Y.; and Mengist, N. E. 2025. Advancing NLP Data Equity: Practitioner Responsibility and Accountability in NLP Data Practices. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 679–691.

Davis, J. L.; Williams, A.; and Yang, M. W. 2021. Algorithmic Reparation. *Big Data & Society*, 8(2): 20539517211044808.

De Paolis Kaluza, M. C.; Tholeti, T.; Chen, Y.; Baeza-Yates, R.; Radivojac, P.; and Jain, S. 2025. Are Your Fairness Metrics Accurate? A Semi-Supervised Approach to Improving Fairness Estimates Under Sample Selection Bias. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, 439–450.

Delaney, E.; Fu, Z.; Wachter, S.; Mittelstadt, B.; and Russell, C. 2024. OxonFair: A Flexible Toolkit for Algorithmic Fairness. *Advances in Neural Information Processing Systems*, 37: 94209–94245.

Deng, W. H.; Barocas, S.; and Wortman Vaughan, J. 2025. Supporting Industry Computing Researchers in Assessing, Articulating, and Addressing the Potential Negative Societal Impact of Their Work. *Proceedings of the ACM on Human-Computer Interaction*, 9(2): 1–37.

Deng, W. H.; Nagireddy, M.; Lee, M. S. A.; Singh, J.; Wu, Z. S.; Holstein, K.; and Zhu, H. 2022. Exploring How Machine Learning Practitioners (Try To) Use Fairness Toolkits. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 473–484.

Deng, W. H.; Yildirim, N.; Chang, M.; Eslami, M.; Holstein, K.; and Madaio, M. 2023. Investigating Practices and Opportunities for Cross-functional Collaboration around AI Fairness in Industry Practice. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 705–716.

Derenoncourt, E.; Kim, C. H.; Kuhn, M.; and Schularick, M. 2024. Wealth of two nations: The US racial wealth gap, 1860–2020. *The Quarterly Journal of Economics*, 139(2): 693–750.

Diaz, M.; and Smith, A. D. R. 2024. What Makes An Expert? Reviewing How ML Researchers Define "Expert". In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 7, 358–370.

Dong, Z.; Barrett, T.; Patil, A. B.; Shoda, Y.; Battle, L.; and Wall, E. 2025. A Design Space of Behavior Change Interventions for Responsible Data Science. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, 37–53.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226.

Edwards, R. 2019. African Americans and the Southern Homestead Act. *Great Plains Quarterly*, 39(2): 103–129.

Eubanks, V. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. Macmillan+ ORM.

Feathers, T. 2023. False Alarm: How Wisconsin Uses Race and Income to Label Students "High Risk" – The Markup. The Markup. Https://themarkup.org/machine-learning/2023/04/27/false-alarm-how-wisconsin-uses-race-and-income-to-label-students-high-risk.

Feldman, M.; Friedler, S. A.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and Removing Disparate Impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268.

Gilad, A.; Deutch, D.; and Roy, S. 2020. On Multiple Semantics for Declarative Database Repairs. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data*, 817–831.

Green, B. 2021. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*, 2(3): 249–265.

Green, B. 2022. Escaping the Impossibility of Fairness: From Formal to Substantive Algorithmic Fairness. *Philosophy & Technology*, 35(4): 90.

Grgić-Hlača, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Guha, S.; Khan, F. A.; Stoyanovich, J.; and Schelter, S. 2024. Automated Data Cleaning Can Hurt Fairness in Machine Learning-Based Decision Making. *IEEE Transactions on Knowledge and Data Engineering*, 36(12): 7368–7379.

Hansen, D.; Devic, S.; Nakkiran, P.; and Sharan, V. 2024. When is Multicalibration Post-Processing Necessary? *Advances in Neural Information Processing Systems*, 37: 38383–38455.

Haraway, D. 1988. Situated Knowledges: The Science Question in Feminism and the Privilege of Partial Perspective. *Feminist Studies*, 14(3): 575–599.

Harding, S. 1995. "Strong Objectivity": A Response to the New Objectivity Question. *Synthese*, 104(3): 331–349.

Hardt, M.; Chen, X.; Cheng, X.; Donini, M.; Gelman, J.; Gollaprolu, S.; He, J.; Larroy, P.; Liu, X.; McCarthy, N.; et al. 2021. Amazon Sagemaker Clarify: Machine Learning Bias Detection and Explainability in the Cloud. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, 2974–2983.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *Advances in Neural Information Processing Systems 29*, 3315–3323.

Harrison, G.; Bryson, K.; Bamba, A. E. B.; Dovichi, L.; Binion, A. H.; Borem, A.; and Ur, B. 2024. JupyterLab in Retrograde: Contextual Notifications That Highlight Fairness and Bias Issues for Data Scientists. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*.

Holstein, K.; Wortman Vaughan, J.; Daumé III, H.; Dudik, M.; and Wallach, H. 2019. Improving Fairness in Machine Learning Systems: What Do Industry Practitioners Need? In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Hort, M.; Chen, Z.; Zhang, J. M.; Harman, M.; and Sarro, F. 2024. Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey. *ACM Journal on Responsible Computing*, 1(2): 1–52.

Jacobs, A. Z.; and Wallach, H. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 375–385.

Jain, S.; Hamidieh, K.; Georgiev, K.; Ilyas, A.; Ghassemi, M.; and Mądry, A. 2024. Improving Subgroup Robustness via Data Selection. *Advances in Neural Information Processing Systems*, 37: 94490–94511.

Kamiran, F.; and Calders, T. 2012. Data Preprocessing Techniques for Classification Without Discrimination. *Knowledge and Information Systems*, 33(1): 1–33.

Kasica, S.; Berret, C.; and Munzner, T. 2023. Dirty Data in the Newsroom: Comparing Data Preparation in Journalism and Data Science. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Kasirzadeh, A. 2022. Algorithmic Fairness and Structural Injustice: Insights From Feminist Political Philosophy. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 349–356.

Krasanakis, E.; Spyromitros-Xioufis, E.; Papadopoulos, S.; and Kompatsiaris, Y. 2018. Adaptive Sensitive Reweighting to Mitigate Bias in Fairness-aware Classification. In *Proceedings of the 2018 World Wide Web Conference*, 853–862.

Krishnan, S.; Franklin, M. J.; Goldberg, K.; and Wu, E. 2017. Boostclean: Automated Error Detection and Repair for Machine Learning. *arXiv:1711.01299*.

Lakkaraju, H.; Kleinberg, J.; Leskovec, J.; Ludwig, J.; and Mullainathan, S. 2017. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 275–284.

Lima, G.; Grgić-Hlača, N.; Langer, M.; and Zou, Y. 2025. Lay Perceptions of Algorithmic Discrimination in the Context of Systemic Injustice. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Liu, T.; Fan, J.; Luo, Y.; Tang, N.; Li, G.; and Du, X. 2021. Adaptive Data Augmentation for Supervised Learning Over Missing Data. *Proceedings of the VLDB Endowment*, 14(7): 1202–1214.

Long, C.; Hsu, H.; Alghamdi, W.; and Calmon, F. 2023. Individual Arbitrariness and Group Fairness. *Advances in Neural Information Processing Systems*, 36: 68602–68624.

Loury, A. 2023. While Things Have Improved, Chicago Remains the Most Segregated City in America. WBEZ. https://www.wbez.org/race-class-communities/2023/06/19/chicago-remains-the-most-segregated-big-city-in-america.

Lucchesi, L. R.; Kuhnert, P. M.; Davis, J. L.; and Xie, L. 2022. Smallset Timelines: A Visual Representation of Data Preprocessing Decisions. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 1136–1153.

Luo, Y.; Li, Z.; Liu, Q.; and Zhu, J. 2025. Fairness Without Demographics Through Learning Graph of Gradients. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, 918–926.

Lutton, L.; Fan, A.; and Loury, A. 2020. Home Loans in Chicago: One Dollar To White Neighborhoods, 12 Cents To Black. WBEZ. https://interactive.wbez.org/2020/banking/disparity.

McDonald, N.; Schoenebeck, S.; and Forte, A. 2019. Reliability and Inter-rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).

Miceli, M.; Posada, J.; and Yang, T. 2022. Studying Up Machine Learning Data: Why Talk About Bias When We Mean Power? *Proceedings of the ACM on Human-Computer Interaction*, 6(GROUP): 1–14.

Mothilal, R. K.; Lalani, F. M.; Ahmed, S. I.; Guha, S.; and Sultana, S. 2025. Talking About the Assumption in the Room. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*.

Muller, M.; Lange, I.; Wang, D.; Piorkowski, D.; Tsay, J.; Liao, Q. V.; Dugan, C.; and Erickson, T. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*.

Nelson, R. K.; Winling, L.; Marciano, R.; Connolly, N.; et al. 2021. Mapping Inequality: Redlining in New Deal America. *University of Richmond Digital Scholarship Lab*, 5(39.1): 94–58.

Nightingale, C. 2012. *Segregation: A Global History of Divided Cities*. Historical Studies of Urban America. University of Chicago Press.

Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2024. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. arXiv:2402.17861.

Paullada, A.; Raji, I. D.; Bender, E. M.; Denton, E.; and Hanna, A. 2021. Data and its (Dis)contents: A Survey of Dataset Development and Use in Machine Learning Research. *Patterns*, 2(11).

Perdomo, J. C.; Britton, T.; Hardt, M.; and Abebe, R. 2023. Difficult Lessons on Social Prediction from Wisconsin Public Schools. arXiv:2304.06205.

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for Individual Fairness. *Advances in Neural Information Processing Systems*, 34: 25944–25955.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J.; and Weinberger, K. Q. 2017. On Fairness and Calibration. *Advances in Neural Information Processing Systems*, 30.

Rakova, B.; Yang, J.; Cramer, H.; and Chowdhury, R. 2021. Where Responsible AI Meets Reality: Practitioner Perspectives on Enablers for Shifting Organizational Practices. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

Rekatsinas, T.; Chu, X.; Ilyas, I. F.; and Ré, C. 2017. HoloClean: Holistic Data Repairs with Probabilistic Inference. *Proceedings of the VLDB Endowment*, 10(11): 1190–1201.

Richardson, B.; Garcia-Gathright, J.; Way, S. F.; Thom, J.; and Cramer, H. 2021. Towards Fairness in Practice: A Practitioner-Oriented Rubric for Evaluating Fair ML Toolkits. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Rismani, S.; Shelby, R.; Davis, L.; Rostamzadeh, N.; and Moon, Aj. 2025. Measuring What Matters: Connecting AI Ethics Evaluations to System Attributes, Hazards, and Harms. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, 2199–2213.

Rismani, S.; Shelby, R.; Smart, A.; Jatho, E.; Kroll, J.; Moon, A.; and Rostamzadeh, N. 2023. From Plane Crashes to Algorithmic Harm: Applicability of Safety Engineering Frameworks for Responsible ML. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2023. Improving Fair Training Under Correlation Shifts. In *Proceedings of the International Conference on Machine Learning*, 29179–29209.

Rosson, M. B.; and Carroll, J. M. 2007. Scenario-Based Design. In *The Human-Computer Interaction Handbook*, 1067–1086. CRC Press.

Saleiro, P.; Kuester, B.; Hinkson, L.; London, J.; Stevens, A.; Anisfeld, A.; Rodolfa, K. T.; and Ghani, R. 2018. Aequitas: A Bias and Fairness Audit Toolkit. *arXiv:1811.05577*.

Salimi, B.; Howe, B.; and Suciu, D. 2020. Database Repair Meets Algorithmic Fairness. *SIGMOD Rec.*, 49(1): 34–41.

Salimi, B.; Rodriguez, L.; Howe, B.; and Suciu, D. 2019. Interventional Fairness: Causal Database Repair for Algorithmic Fairness. In *Proceedings of the 2019 International Conference on Management of Data*, 793–810.

Sambasivan, N.; Kapania, S.; Highfill, H.; Akrong, D.; Paritosh, P.; and Aroyo, L. M. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*.

Saunders, B.; Sim, J.; Kingstone, T.; Baker, S.; Waterfield, J.; Bartlam, B.; Burroughs, H.; and Jinks, C. 2018. Saturation in Qualitative Research: Exploring its Conceptualization and Operationalization. *Quality & Quantity*, 52(4): 1893–1907.

Schelter, S.; He, Y.; Khilnani, J.; and Stoyanovich, J. 2019. FairPrep: Promoting Data to a First-Class Citizen in Studies on Fairness-Enhancing Interventions. arXiv:1911.12587.

Schwöbel, P.; and Remmers, P. 2022. The Long Arc of Fairness: Formalisations and Ethical Discourse. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2179–2188.

Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Rostamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; et al. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 723–741.

Silberzahn, R.; Uhlmann, E. L.; Martin, D. P.; Anselmi, P.; Aust, F.; Awtrey, E.; Bahník, ŝ.; Bai, F.; Bannard, C.; Bonnier, E.; Carlsson, R.; Cheung, F.; Christensen, G.; Clay, R.; Craig, M. A.; Dalla Rosa, A.; Dam, L.; Evans, M. H.; Flores Cervantes, I.; Fong, N.; Gamez-Djokic, M.; Glenz, A.; Gordon-McKeon, S.; Heaton, T. J.; Hederos, K.; Heene, M.; Hofelich Mohr, A. J.; Högden, F.; Hui, K.; Johannesson, M.; Kalodimos, J.; Kaszubowski, E.; Kennedy, D. M.; Lei, R.; Lindsay, T. A.; Liverani, S.; Madan, C. R.; Molden, D.; Molleman, E.; Morey, R. D.; Mulder, L. B.; Nijstad, B. R.; Pope, N. G.; Pope, B.; Prenoveau, J. M.; Rink, F.; Robusto, E.; Roderique, H.; Sandberg, A.; Schlüter, E.; Schönbrodt, F. D.; Sherman, M. F.; Sommer, S. A.; Sotak, K.; Spain, S.; Spörlein, C.; Stafford, T.; Stefanutti, L.; Tauber, S.; Ullrich, J.; Vianello, M.; Wagenmakers, E.-J.; Witkowiak, M.; Yoon, S.; and Nosek, B. A. 2018. Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Advances in Methods and Practices in Psychological Science*, 1(3): 337–356.

Simson, J.; Fabris, A.; and Kern, C. 2024. Lazy Data Practices Harm Fairness Research. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 642–659.

Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 2022. Participation Is Not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*.

Smith, J. J.; Madaio, M.; Burke, R.; and Fiesler, C. 2025. Pragmatic Fairness: Evaluating ML Fairness Within the Constraints of Industry. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, 628–638.

So, W.; Lohia, P.; Pimplikar, R.; Hosoi, A.; and D'Ignazio, C. 2022. Beyond Fairness: Reparative Algorithms to Address Historical Injustices of Housing Discrimination in the US. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 988–1004.

Suresh, H.; and Guttag, J. 2021. A Framework for Understanding Sources of Harm throughout the Machine Learning

Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*.

Tae, K. H.; Roh, Y.; Oh, Y. H.; Kim, H.; and Whang, S. E. 2019. Data Cleaning for Accurate, Fair, and Robust Models: A Big Data - AI Integration Approach. In *Proceedings of the 3rd International Workshop on Data Management for End-to-End Machine Learning*.

Taylor, K.-Y. 2019. *Race for Profit: How Banks and the Real Estate Industry Undermined Black Homeownership*. Justice, Power, and Politics. UNC Press Books.

van Breugel, B.; Kyono, T.; Berrevoets, J.; and van der Schaar, M. 2021. DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks. In *Advances in Neural Information Processing Systems*, volume 34, 22221–22233.

van Dijck, J. 2014. Datafication, Dataism and Dataveillance: Big Data between Scientific Paradigm and Ideology. *Surveillance & Society*, 12(2): 197–208.

Verma, S.; and Rubin, J. 2018. Fairness Definitions Explained. In *Proceedings of the International Workshop on Software Fairness*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2021. Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law. *West Virginia Law Review*, 123(3).

Wang, Z.; Yin, Z.; Zhang, Y.; Yang, L.; Zhang, T.; Pissinou, N.; Cai, Y.; Hu, S.; Li, Y.; Zhao, L.; et al. 2025. Fg-smote: Towards Fair Node Classification with Graph Neural Network. *ACM SIGKDD Explorations Newsletter*, 26(2): 99–108.

Wang, Z. J.; Kale, A.; Nori, H.; Stella, P.; Nunnally, M. E.; Chau, D. H.; Vorvoreanu, M.; Wortman Vaughan, J.; and Caruana, R. 2022. Interpretability, Then What? Editing Machine Learning Models to Reflect Human Knowledge and Values. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 4132–4142.

Wexler, J.; Pushkarna, M.; Bolukbasi, T.; Wattenberg, M.; Viégas, F.; and Wilson, J. 2020. The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1): 56–65.

Widder, D. G.; Zhen, D.; Dabbish, L.; and Herbsleb, J. 2023. It's About Power: What Ethical Concerns Do Software Engineers Have, and What Do They (Feel They Can) Do about Them? In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 467–479.

Yan, B.; Seto, S.; and Apostoloff, N. 2022. FORML: Learning to Reweight Data for Fairness. arXiv:2202.01719.

Zafar, M. B.; Valera, I.; Gomez-Rodriguez, M.; and Gummadi, K. P. 2019. Fairness Constraints: A Flexible Approach for Fair Classification. *Journal of Machine Learning Research*, 20(75).

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, 325–333.

Zhang, A. X.; Muller, M.; and Wang, D. 2020. How Do Data Science Workers Collaborate? Roles, Workflows, and Tools. *Proc. ACM Hum.-Comput. Interact.*, 4(CSCW1): 22:1–22:23.