Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies

Briana Vecchione
Cornell University
Ithaca, USA
bv225@cornell.edu

Solon Barocas
Microsoft Research and Cornell University
New York, USA
solon@microsoft.com

Karen Levy
Cornell University
Ithaca, USA
karen.levy@cornell.edu

ABSTRACT

"Algorithmic audits" have been embraced as tools to investigate the functioning and consequences of sociotechnical systems. Though the term is used somewhat loosely in the algorithmic context and encompasses a variety of methods, it maintains a close connection to audit studies in the social sciences—which have, for decades, used experimental methods to measure the prevalence of discrimination across domains like housing and employment. In the social sciences, audit studies originated in a strong tradition of social justice and participatory action, often involving collaboration between researchers and communities; but scholars have argued that, over time, social science audits have become somewhat distanced from these original goals and priorities. We draw from this history in order to highlight difficult tensions that have shaped the development of social science audits, and to assess their implications in the context of algorithmic auditing. In doing so, we put forth considerations to assist in the development of robust and engaged assessments of sociotechnical systems that draw from auditing's roots in racial equity and social justice.

ACM Reference Format:
Briana Vecchione, Solon Barocas, and Karen Levy. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '21), October 5–9, 2021, –, NY, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3465416.3483294

1 INTRODUCTION

The past decade has been marked by an increasing push to subject algorithms to audit [9, 46, 54, 88]. The call for such audits reflects a growing sense that algorithms play an important, yet opaque, role in the decisions that shape people's life chances—as well as a recognition that audits have been uniquely helpful in advancing our understanding of the concrete consequences of algorithms in the wild and in assessing their likely impacts.

As audits have proliferated, however, the meaning of the term has become ambiguous, making it hard to pin down what audits actually entail and what they aim to deliver [18, 56, 75]. It is common for *any* empirical investigation of an algorithm to be deemed an audit, despite the fact that such studies may involve very different measurement techniques and may focus on very different research questions. This confusion is understandable: colloquial use of the term audit can refer to a broad range of activities—including legal investigations by government tax authorities, compliance-oriented inspections by accounting firms, and the like—and the term is taken up in different ways in various disciplines (e.g., anthropology [81] and public management [65]).

Recent work has attempted to give a more precise description of what algorithmic audits could—and should—entail, offering important methodological recommendations and raising challenging questions [63, 64, 76]. This paper aims to complement these efforts by drawing lessons from the history of audit studies, which originate in the social sciences.

Starting in the 1970s, social scientists, advocates, and community organizers developed empirical methods to detect and measure the degree of discrimination in domains like housing. These initial efforts were driven by explicit concerns with racial equity and social justice, developed with direct participation of the affected communities and oriented around accountability and reform. Audit studies also represented an important methodological innovation, as they involved having real people go through the actual process of seeking housing, with the aim of measuring the degree to which characteristics like race influenced their treatment. Audit studies in the social sciences are a paradigmatic example of a "field experiment"—that is, a controlled experiment conducted not inside a lab, but instead out in the real world, with the goal of observing how actual decision makers behave. Such methods have been heralded as particularly effective techniques for uncovering the degree of discrimination that different groups face across various domains, especially as more overt signals of racial prejudice have declined [57]. A particular version of the audit study—the "correspondence study," in which researchers submit, for example, resumes to employers that systematically vary signals of gender or race—is now commonly viewed as the gold standard for empirical investigation of discrimination [13].

While the methodological rigor of audit studies is one of their primary advantages, the merits of the technique have also been subject to debate within social science. As Cherry and Bendick point out in their excellent historical overview of audit studies, from which we will draw extensively: the "single-minded pursuit of rigor [in latter-day audit studies] risks sacrificing other considerations historically associated with auditing's unique contributions to both society and science" [21]. In particular, scholars observe that field experiments to detect discrimination have become disembedded

EAAMO '21, October 5–9, 2021, –, NY, USA

ACM ISBN 978-1-4503-8553-4/21/10.
https://doi.org/10.1145/3465416.3483294

from communities subject to discrimination, and thus disconnected from their original social justice aims.

These points of debate within the social sciences are particularly instructive for those conducting algorithmic audits. Our paper thus proceeds accordingly. After providing a brief history of audit studies in the social sciences, we describe the many ways in which algorithmic auditing is indebted to this prior work, while also frequently departing from it. We then discuss how the history of audits in the social sciences highlights a number of difficult tensions that those conducting algorithmic audits would do well to consider—and how algorithmic audits might be designed and conducted in a way that accounts for the racial equity and social justice roots of auditing in social science.

2 THE SOCIAL JUSTICE ORIGINS OF SOCIAL SCIENCE AUDITING

Social science auditing originated in activist research. In the 1940s and 1950s, civic organizations partnered with researchers in order to better assess the state of race relations in the United States [21, 33]. Since civil rights were not yet statutorily protected, these groups had to rely on "persuasion and cooperation" to effect change; the currency of this strategy was *information* about the incidence of discrimination, which audit studies could provide. Community groups marshalled audit studies to raise awareness of problems—including, in some cases, community groups meeting with audited entities after-the-fact to discuss what the study had revealed about their specific behavior [21].

For researchers, these collaborations were a key example of the participatory action research (PAR) paradigm. PAR methods involve collaboratively and reflexively engaging with research participants, and treating them as more than merely informants in a researcher-led study. Instead, researchers and practitioners co-design and co-execute projects from the ground up, with dual goals of both "advancing scientific knowledge [and] achieving practical objectives" [89]. Participatory action emphasizes the importance of conducting research *with* participants, not on or for them [51].

Methods that arose during this time included techniques like community self-surveys, in which community members themselves collected data via in-person and/or telephone interviews. Communities themselves organized the studies, with support and training from researchers. A key priority of community-led studies was the experience of participation in research itself. Engaging in the research process was understood to support community autonomy, increase trust in the results, and spur reform. As the foreword to the 1951 manual *How to Conduct a Community Self-Survey of Civil Rights* explains: "Knowledge self-obtained seems more authentic than second-hand knowledge. It reflects both an inner sense of urgency and a dependable view of social reality. Thus it is pointed toward effective social action. Since the self-survey leads to full-bodied participation on the part of the citizen it is a valuable tool of modern democracy" [91].

Later, audit studies became an important tool for enforcement of civil rights statutes. This was most notable in housing, where large-scale in-person audits in the 1970s—sponsored by the Department of Housing and Urban Development, and executed in partnership with community groups and local researchers—demonstrated the prevalence of discrimination against Black prospective tenants and homebuyers across the country, and influenced federal policy [21, 94].

Over time, however, enthusiasm for community/researcher partnerships to study discrimination waned. Participatory action perspectives in general became less common within the academy, as partnerships were resource-intensive and poorly incentivized. Concomitantly, more sophisticated quantitative modeling gained traction across the social sciences, creating expectations for greater statistical and methodological rigor in academic work [21]. In in-person audits, even when testers are carefully matched—when the researcher takes care to control for factors like age, physical appearance, and the like (that is, everything but the variable being tested)—unobservable differences and experimenter effects may still confound the study and are difficult to measure [57]. And by the 2000s, many employment and housing application processes occurred online rather than in-person, making them less amenable to in-person audits [33].

These dynamics have contributed to a trend away from in-person audits and toward correspondence studies. Correspondence studies afford researchers great control over the variables presented, and can be conducted at much broader scale than in-person audits, including across many geographic contexts [15, 60]. In these studies, researchers rely on fictitious correspondence by hypothetical individuals online or by mail, systematically varying the characteristics of "applicants" to test for discrimination. In one well-known correspondence audit by Bertrand and Mullainathan [14], researchers applied to over 1,300 newspaper job advertisements via fax and mail, sending four systematically varied resumes to each employer. The study's findings were striking: all else equal, Black applicants were about 50% less likely than white applicants to receive a callback. The study was also remarkable due to its unprecedented scale, demonstrating that a large-scale correspondence audit could be done by only a few researchers. Correspondence studies, of course, face their own methodological limitations—a notable difficulty is how effectively researchers can signal the characteristic of interest (and only the characteristic of interest) using applicant names or other indicators [23, 32]—but their comparative advantages have led them to dominate audit studies in the social sciences, and to be considered a "significant methodological advance" over in-person tests [13, 38]. Correspondence studies in the social sciences have been used to examine discrimination involving a host of personal characteristics other than race and gender, including employment history [59], LGBT status [53], immigration status [35], and many other characteristics [33].

3 ALGORITHMIC AUDITING

In some ways, algorithmic audits have much in common with audit studies in the social sciences, suggesting that lessons from the latter should easily carry over to the former. Yet recent work has shown that the goals motivating activities deemed algorithmic audits—and the methods used to conduct them—can be surprisingly capacious [9] and often differ from those in the social sciences. In the developing academic and policy discourse—and increasingly in colloquial language—the term algorithmic audit has come to refer to almost any kind of empirical study of algorithms, even though audit studies

within the social sciences refer specifically to field experiments designed to measure the extent of discrimination within the domain. This degree of variation in algorithmic audits has created some confusion about what the term algorithmic auditing means and what qualifies an activity as an algorithmic audit. Understanding variations in both the goals and methods of algorithmic audits can help to ensure that the lessons we draw from the social sciences are well-matched to the particular type of algorithmic audits for which they are relevant.

While some algorithmic audits are focused on establishing whether an algorithm considers legally proscribed factors or whether an algorithmic process is sensitive to changes in these factors [3, 31, 39, 43, 50, 90], other algorithmic audits have different goals. For example, some algorithmic audits attempt to uncover violations of procedural regularity—that is, when an algorithm fails to function as promised and in a consistent manner, but without specific emphasis on discrimination [25, 46, 69, 72]. Others are motivated by a concern with transparency, aiming to uncovering how an algorithm works—with the term auditing often meaning something like reverse engineering [1, 2, 26]. Even traditional evaluations of algorithms' performance in terms of accuracy, precision, and recall [37] are sometimes described as audits; more recent efforts have begun to disaggregate these results by group, revealing when algorithms perform less well for some groups than others [4, 10, 20]—a problem often described as bias, but different in nature than the discriminatory treatment examined by social science audits. Still others focus on completely different notions of bias, like the balance of political views in algorithmically curated content [68].

While audit studies in the social sciences are by definition experiments (in which the researcher directly manipulates the variable of interest, and then measures outcomes), the wide range of studies described as algorithmic audits can involve purely observational studies of algorithmic outcomes (in which the researcher measures outcomes without manipulation) as well as direct inspection of the algorithms themselves. Algorithmic audits that take an experimental approach often rely on so-called "sock puppets," where computer programs are used to impersonate users of a platform [5, 71]. This is similar to the approach taken in correspondence studies in the social sciences, but with a greater degree of automation. Like correspondence studies, sock-puppet-based algorithmic audits allow researchers to carefully vary inputs, furnishing greater statistical rigor. Researchers can also enlist volunteers to participate in an audit—a practice Sandvig et al. [71] describe as a "noninvasive user audit." In this case, real users disclose information about their interactions with a platform (e.g., search queries and results), with the goal of allowing researchers to learn something about the platform's algorithm, despite the lack of experimental control. A recent example is The Markup's "citizen browser project," where a panel of individuals installed a custom browser on their computers that tracks data about their Facebook and YouTube accounts, hopefully leading to insights about how each platform's algorithms operate [49].

In what follows, we aim to draw lessons from the social sciences that are most relevant to those who look to audits as a way to uncover and address discrimination in algorithms specifically. In narrowing our focus to audits concerned with discrimination, we do not intend to suggest that audits cannot or should not play a role in evaluating algorithmic systems with other normative concerns in mind. Nor do we mean to suggest that insights from the history of audit studies in the social sciences are irrelevant to algorithmic audits with a different focus. However, it is difficult to discuss the relative merits of different approaches to audits when the purpose of such audits is left under-specified. We therefore focus the rest of our discussion on audits that aim to uncover and root out discrimination.

4 THE ROOTS OF SOCIAL SCIENCE AUDITING AND THE FUTURE OF ALGORITHMIC AUDITING

The history of auditing in the social sciences can help us chart a path forward for algorithmic auditing. Below, we describe four key considerations, each of which draws from tensions within social science regarding the aims and capabilities of audit studies, and describe their implications in the context of algorithmic audits. Our goal is to draw from the past in order to help us better reflect on the future: we should learn from the lessons of social science to intentionally consider what role we want algorithmic audits to play in advancing social justice and how to go about designing and conducting them. In so doing, we again note the debt that our analysis owes to Cherry and Bendick's insightful critique of audits in the social sciences, which emphasizes their roots in the activist scholarship tradition [21].

4.1 Beyond discrete moments of decision making

Social science audit studies can make claims only about discrete *moments* at which bias can emerge—and be readily measured—in a process [11]. For example, many correspondence studies in the hiring context zero in on the decision to invite an applicant for an interview based on an evaluation of their resume; earlier and later stages of hiring processes, and the biases that may emerge in them, are not captured. In an analysis by Quillian et al. [61] of over 100 audit studies of racial discrimination in the employment process, only 13 considered the ultimate employment outcome (a job offer), choosing instead to treat the "callback"—an invitation to interview—as a proxy for the outcome of interest. Despite this, they show that substantial additional discrimination occurs *after* the callback stage, and is missed by most studies.

There are good reasons for such narrow focus in social science auditing. As Cherry and Bendick describe, social scientists are drawn to correspondence studies over in-person audits for many reasons, including their amenability to making rigorous statistical claims, high degree of researcher control, and ability to scale. These methods naturally lend themselves to examining only certain stages of social processes in which people can be evaluated based on indicators that can be ascertained through documents. Doing so can also help to eliminate noise in the data, and can focus findings on specific mechanisms through which bias proliferates. Finally, research ethics may require that researchers restrict field experiments to parts of a process that are unlikely to cause significant harm to research participants and others (particularly in light of the deception that audit studies require); the cost to an employer of

evaluating an extra resume is much lower than that of interviewing a live tester, who may be displacing a genuine job candidate.

Therefore, while correspondence audits can be rigorously controlled and demonstrate statistically significant effects, they can offer only a "thin slice" view of where in a process discrimination can occur. Other social scientific methods can complement audits to shed light on other stages in a process. For example, Rivera's ethnographic examination of elite firm hiring practices delineates nine stages of hiring, from recruitment to deliberation [67]. Rivera demonstrates that privileged applicants are advantaged across the board—not only by resume indicators (e.g., university attended) but by class markers that emerge elsewhere in the process, like in in-person interviews (e.g., cultural homogeneity and interaction style). Because some of these stages are less amenable to audit methods, audits that focus only on resume screening necessarily underestimate the true degree of discrimination job applicants face in a hiring process.

Similar concerns arise in the context of algorithmic auditing. Selbst et al. [73] caution researchers who study fairness in machine learning against the *framing trap*—the "failure to model the entire system over which a social criterion, such as fairness, will be enforced." They encourage researchers to analyze not only how bias may arise in a machine learning model in isolation, but how it may arise *sociotechnically*, when humans and institutions interact with the model in the social world. For example, a judge receiving recommendations from a risk assessment tool [79] or a hiring manager obtaining a "fair" list of screened job applicants may make decisions in concert with the technical tool that simply relocates bias to a new, opaque stage of the process. At the very least, algorithmic audits that focus solely on the technical components of a sociotechnical system should be understood to be subject to this limitation. But often, the distinction is blurred; a process may be trumpeted as "fair" with respect to its technical components, without due attention to the caveat that this assessment applies at best to only particular stages of a process [16, 62].

We might also attempt more "end-to-end" audits that encompass humans and institutions in interaction with algorithmic systems, accounting for operational and social aspects of use alongside technical considerations [44]. Ample evidence suggests that best-case assumptions about how an algorithm is used—which might be the basis for audits proclaiming they have met some fairness or quality criterion—may not hold empirically. A clear example emerged when the ACLU demonstrated that Amazon's face recognition product mistakenly matched 28 Members of Congress to mugshots: Amazon countered that the ACLU had intentionally misrepresented the product by failing to use its recommended 99% confidence threshold for public safety applications [28, 78]. The ACLU countered that it had used an 80% confidence threshold which was the *default setting* for the product, and therefore likely to reflect actual use, no matter what Amazon's manual recommended. (Research demonstrates that police departments *do* routinely engage in all kinds of less-than-ideal practices around face recognition algorithms—including using both composite sketches and celebrity lookalike photos as input data [34]). On the other "end," Stevenson and Doleac demonstrate how human judges inconsistently *follow* the sentencing recommendations proffered by risk assessment algorithms in the criminal justice context, potentially reintroducing biases and further skewing outcomes by race and age [80]. The key point is that audits might be most meaningfully deployed to assess bias in *actual, not optimal*, conditions of end-to-end use.

More broadly, both social scientific and algorithmic auditing necessarily underestimate both the incidence and impact of discrimination in the social world. In addition to the "thin-slice" problem we have detailed, it is important to recognize that many discriminatory processes cannot be examined via audits at *any* stage. For example, jobs that are filled through social networks rather than through job ads cannot be readily audited [8], despite being both frequent routes to employment [36, 85] and likely to propagate discriminatory outcomes due to homophily [17, 52]. Blue-collar jobs more often require in-person application than white-collar jobs, and thus may be less readily auditable using correspondence studies [57]. Perhaps most importantly, audits typically assess *episodic* incidents of discrimination—one job search, one credit application—not discrimination that accrues *cumulatively*. But as Small and Pager note, discrimination causes harm "not just at critical junctures but also over the slow, lifelong buildup of its everyday sting" [77]. While we should appreciate what can be learned from audit studies, we should also appreciate what can't be.

4.2 The dangers of adopting and abandoning experimental controls

Audit studies aim to measure the degree to which differences in the perceived gender or race of job applicants, for example, affect the way that employers treat candidates who are otherwise identical. In this respect, audit studies are an attempt to detect what the law deems "disparate treatment": decision making that takes legally proscribed variables like gender or race into account. The fundamental premise of audit studies is simple, but powerful: if decision makers treat candidates differently who are the same except for their gender or race, then the decision must be taking this feature into account, as nothing else could explain the difference. Tightly controlling for other factors that might influence the decision thus allows researchers to conclude that the decision maker's awareness of applicants' gender or race must be influencing the outcome. This explains the enormous effort that researchers invest in training in-person testers, aiming to ensure that testers present equivalent background stories, answer questions in the same way, and otherwise comport themselves in a manner that makes them as similar as possible except for their gender or race (more on this challenge in a moment) [57]. It likewise explains the allure of correspondence studies, which allow researchers to avoid this challenge altogether, as the testing takes place via paper resumes over which researchers have complete control.

Unfortunately, these methods do not always translate well to algorithmic decision making because many algorithms used in high-stakes domains (like hiring) do not include gender or race as direct inputs to the decision.[1] While there are a few instances of

[1] This is not to say that potential—and sometimes quite obvious—proxies for gender and race might not remain part of the process. In the well-cited Reuters story about Amazon's abandoned AI recruitment tool, the tool was found to penalize job candidates whose resumes included language like "women's chess club captain," giving away applicants' gender, even if it was not an explicit feature [24]. To our knowledge, however, no algorithmic audits have tried to test experimentally whether algorithms differ in their behavior when such features are varied while everything else is held constant. Experiments along these lines would be very similar to those that have been

algorithmic audits that have tried to establish whether algorithms are indeed engaged in disparate treatment [25, 83]—and, in fact, find that they are—the vast majority aim to measure what the law calls "disparate impact": differences in the outcomes experienced by difference groups even when the decision making process only relies on seemingly benign variables. This approach abandons the overriding focus on experimental control that is characteristic of the more recent audit studies in the social sciences and instead aims to measure how different groups fare when subject to algorithmic decision making, *given natural differences between groups in the variables that the algorithm takes into account.* In many respects, this is a return to the approach characteristic of earlier audit studies.

Changing attitudes within the social sciences about the relative merits of each approach offer helpful lessons for those undertaking algorithmic audits. Carefully controlled audit studies offer a degree of methodological rigor that makes it much easier to attribute different outcomes directly to the variable of interest. In the absence of careful controls, advocates argue, differences in outcome (e.g., the callback rate) might be explained by differences in other covariates (e.g., formal qualifications, education, work experience, etc.), the distribution of which might differ across groups. It is far more challenging to make an argument that decision makers have discriminated on the basis of gender or race when there are other potential explanations for the disparity in outcomes. But the criticisms of tightly controlled experimental approaches in the social sciences also reveal the danger of treating the process of establishing discrimination as a matter of showing that certain outcomes would have been the same, but for a difference only in some discrete marker of gender or race [74]. As Hu and Kohler-Hausmann argue, gender and race are not isolated attributes disconnected from all other facts about a person—attributes that can be varied experimentally without affecting other factors traditionally subject to tight control [42]. What it *means* to be a woman or be a Black person is not determined by which gender or race a person happens to tick on a form—or have ticked for them. These are complex social constructs that implicate a profound range of details about a person and the way others interact with them [95]. When researchers control for everything but some discrete marker of gender or race (e.g., a name), they overlook how gendered and racialized identity necessarily implicates the many other factors that are subject to careful experimental control [96].

Many algorithmic audits have—perhaps unwittingly—avoided some of this tension by focusing instead on differences in the accuracy of these decisions. This approach seems to rest on an intuitive normative belief that while people from different groups may not be entitled to the same outcomes, given possible differences in the qualities possessed by members of these groups, they are all entitled to equally accurate assessment. This is quite a departure from the focus of audit studies in the social sciences (both the early and more recent variants), which have focused less on evaluating the accuracy of decision making, in part due to methodological limitations (e.g., researchers would need to know how well a job applicant would have done had they been hired) and in part due to the way that they conceptualize discrimination (i.e., as differences in treatment, given differences in overt markers of gender or race.[2] Along the same lines, some algorithmic audits seek to document that a decision-making process produces disparities in outcomes for different groups without controlling for any differences between these groups. Certain audits treat these differences in outcomes as inherently suspect—and are thus vulnerable to criticisms about a lack of appropriate control. Others, however, attempt to identify the factors in the decision-making process that contribute to the disparity in outcome so that their legitimacy as decision criteria might be critically assessed. Rather than testing whether specific pre-determined factors (e.g., gender or race) matter to a decision, as is the approach in audit studies, these algorithmic audits asks which factors—among all those under consideration—explain the difference in outcomes and whether these are problematic in some way.

Each of these approaches to algorithmic audits do not try to manipulate a specific marker of gender or race while holding other features constant; instead, they take the natural differences in the distribution of these features by gender or race as a given and then evaluate how the algorithm handles these cases, looking for systematic differences in accuracy or outcomes. As such, claims about the degree to which an algorithm results in differences in accuracy or outcomes depend entirely on the distribution of features in the populations that have been used to perform the evaluation [10]. In practice, many algorithmic audits are performed with the population that is in the training data, but this population may not reflect the population that will be subject to the algorithm. What appears to be no notable difference in accuracy or outcomes between groups in the training data might not hold when the model encounters future populations from each group with a distribution of features that are quite different. It is not possible to make general claims about a model's differential performance, its potential to cause disparate impact, or the cause of these disparities without making certain assumptions about the likelihood that the distribution of features in the evaluation dataset will match the distribution in the target domain [48]. And yet, these assumptions are rarely disclosed, including in such high-profile cases like the recent audit of HireVue's automated assessment tool conducted by O'Neil Risk Consulting & Algorithmic Auditing (ORCAA) [6]. ORCAA's report, which finds that there was no "bias" in the assessment, does not offer any details about the dataset used to perform the evaluation. Failing to disclose this information might be a carry-over from carefully controlled audit studies, where it was easy to infer what data had been used because the goal was to use the same data (e.g., resume) in all cases while only varying gender or race. This is *not* the case with most algorithmic audits and thus much more detailed disclosure is necessary to be able to interpret the results.

4.3 Forms of knowledge and forms of evidence

An important consideration in the design of an audit study is *what type of knowledge* it seeks to produce. Does an audit study seek to create generalized knowledge about the "state of the world"—for

conducted in the social sciences over the past decade, which, beyond using names as a signal of gender or race, also manipulate other aspects of job applicants' resume like professional associations and personal interests that encourage employers to draw similar inferences [57].

[2]Whether the difference in accuracy is attributable to differences in gender or race or to other features that happen to correlate with gender or race is irrelevant to this analysis, as the goal is to simply show that an algorithm's accuracy varies between groups given natural differences in their covariates.

example, an average or aggregate measure of the degree of discrimination present in a labor market? Or does it seek to create specific knowledge about the practices of a *particular* employer, landlord, or credit provider? Both forms of knowledge can be valuable. Social scientific audit studies have tended to favor the former—unsurprisingly so, given social science's emphasis on describing the social world. But we might imagine that communities affected by discrimination may be less interested in such general knowledge, and more interested in knowledge that supports action targeted at specific discriminating institutions.

These different goals may lead to different audit study designs. For example, many contemporary audit studies are *unpaired*—that is, treatment and control testers are sent to *different* randomized recipients, and collective outcomes compared statistically in order to assess the amount of discrimination experienced in aggregate [21]. In contrast, in a *paired* audit, each recipient evaluates two applicants who are matched on all characteristics except for the variable of interest. Unpaired audits may be more efficient for researchers: in some cases they can be more statistically powerful, and can reduce the risk that entities become suspicious that they are being subjected to an audit [86, 87]. But unpaired audits do not allow for identification of *specific* discriminating decision makers, since no single entity assesses multiple testers. As Cherry and Bendick describe, unpaired audits report "villainy without villains"—they "document an abstract evil attributable only to the overall population from which the audit sample was drawn" [21].[3] But because algorithmic audits *do* often focus on evaluation of a specific, identifiable system, they may be able to avoid the "villainy without villains" problem, and may be put to use for different forms of social activism, as early social scientific audits did.[4]

A final note on this point involves the form and communication of empirical findings from audit studies, and the degree to which such knowledge can support concrete social change. In-person audits have the virtue of narrative depth: live testers' vivid descriptions of their actual experiences being subjected to biased treatment can be extremely powerful for capturing public attention and galvanizing reforms [12, 21, 60]. In-person audits thus can produce "both stories and statistics" [21], each of which has its own persuasive authority to effect change. Paper-based correspondence studies are less likely to produce compelling qualitative accounts since no real person is subjected to biased treatment.

In the algorithmic realm, both stories and statistics have a role to play. Audit researchers have compellingly integrated vivid qualitative accounts—for example, Joy Buolamwini's experience of being invisible to face recognition software unless she wore a white mask [19]—alongside rigorous analysis quantifying the extent of the problem. Other researchers have provided compelling narratives to communicate audit findings, like the aforementioned ACLU demonstration that Amazon's face recognition product erroneously matched 28 Members of Congress (including, notably, civil rights hero Rep. John Lewis) to mugshots [78], or ProPublica's famed investigation of the COMPAS risk assessment algorithm, which paired statistical analysis with the stories of individuals subjected to it [4]. Efforts to integrate quantitative findings alongside compelling narratives of individual cases are promising ways of achieving non-academic policy impacts that can serve community goals more effectively than either approach in isolation.

4.4 Ensuring meaningful community alliances

Beyond the creation of different forms of knowledge and evidence, audit studies may differ in terms of their type and degree of engagement with community groups. As Cherry and Bendick document, in early audits, partnerships with community groups were considered "as integral an objective of the activity as were published reports" [21], drawing from a participatory action research tradition. Over time, such collaboration waned, as social scientists began to focus more on abstract modeling and correspondence studies, which lacked roles for human testers from community groups.

In the algorithmic sphere, scholars have remarked on the degree to which the framing of fairness research may be disconnected from the actual priorities of affected communities, and the tendency to turn the lived experience of discrimination into an abstract intellectual exercise [27, 41]. Therefore, we should ask: what might meaningful community engagement look like in algorithmic auditing? To what degree could researchers collaborate with communities in designing their approaches, and what forms might such collaboration take?

One potentially promising avenue is the development of crowdsourced or "collective" community audits. Crowdsourced audits might involve soliciting participation from willing volunteers who share their data with researchers. A good example is the coordinated use of the GDPR's subject access rights, which allow any individual to obtain a record of their personal data and information about how their data have been used in automated decision making systems. Individuals who trigger data access requests about themselves may then "donate" their data to researchers, who use these data in aggregate to audit systems for discrimination and reverse-engineer their functioning [7]. This collective approach can be understood as a means of rebalancing information asymmetries in the service of social justice [47]. Another approach borrows from the successful model of "bug bounty" programs, which crowdsource security testing by offering researchers financial incentives for identifying security vulnerabilities. Algorithmic bug bounties may function similarly by providing infrastructure for users to report perceived algorithmic biases in the services they use [22, 30, 92].

"Organic" audits are a related development: spontaneous, community-led efforts that arise as a result of perceived bias in a public-facing algorithm, often disseminated through social media. For example, in 2019, entrepreneur David Heinemeier Hansson noted on Twitter [40] that his Apple Card credit limit was 20 times that of his wife—despite them having access to all the same assets, and despite customer service's assurances that the algorithm was fair. The Twitter thread went viral with many others (including Steve Wozniak!) noting that they'd encountered similar issues, and

[3] That said, even paired audits may not be able to demonstrate statistically significant discrimination if each recipient only receives (say) one pair of resumes, but they at least provide one pair of comparative data points for each audited entity.

[4] Of course, there may also be virtues to using algorithmic audits to create generalized knowledge, perhaps by auditing a *set* of systems in comparison to one another. For example, a recent examination by Rieke et al. [66] of hiring platforms used by 15 large hourly employers is able to make general claims about common obstacles facing job-seekers by virtue of its comparative research design. One could easily imagine an algorithmic audit study that similarly assesses discrimination across multiple platforms.

communicating the broad strokes of their own application characteristics and outputs. Of course, this "audit" lacked systematic sampling, experimental control, or statistical rigor, and as such lacked the hallmarks of a scientific study, but it drew enough attention to the problem that the issue was covered in prominent media and sparked an investigation from New York's State Department of Financial Services [84] (which eventually cleared Apple of any wrongdoing [55]). A similar dynamic unfolded when a student posted on Twitter that Zoom routinely "removed the head" of a Black colleague when that colleague used a virtual background, suggesting that Zoom's algorithms failed to adequately detect darker skin tones [45]. When the student posted various images to Twitter to illustrate the problem, Twitter's image cropping algorithms repeatedly previewed only those parts of images with white faces—suggesting that Twitter's algorithms *also* suffered from algorithmic bias in determining what part of an image was most salient to preview (despite the fact that Twitter had already conducted an internal bias audit on its system). This was quickly followed by many others posting their own sets of images with similar results. The event prompted Twitter to announce it was revisiting its image cropping algorithms and changing its interface to allow users more control over image previews [58, 93].

Community-based auditing techniques are advantageous in that they can surface potential biases and test cases that internal auditors might overlook. They also have the important advantage of fostering community awareness and mobilization much more than a researcher-led or internal audit likely would—and concomitantly, greater public accountability. But these techniques also may be less rigorously controlled and systematically sampled than a centralized, researcher-led audit would be, and subject to greater data quality concerns from less-well-trained volunteers. These are important trade-offs—and quite similar to those confronted by other distributed data collection and "citizen science" efforts, which have developed a number of best practices to address sampling biases and achieve data quality [70, 82]. Another approach is to combine community-and researcher-led approaches collaboratively [71] by using community results to identify areas for researchers to subsequently probe systematically (akin to qualitative theory-building and quantitative theory-testing in social science research [29]). More centralized researcher control—like the collection of data from a representative panel—can address concerns about unsystematic selection of test cases, but seems less likely to result in the meaningful community engagement and capacity-building that characterized early participatory audits.

5 CONCLUSION

In this work, we have highlighted aspects of the history of audit studies in the social sciences—and ongoing debates about their design and implementation—that are useful for the development of algorithmic auditing. The history suggests that there is significant valuable knowledge that can be generated by carefully controlled experiments conducted in contexts marked by discrimination. But statistical rigor comes at a price. Earlier orientations to auditing had different desirable features that have been diminished over time—particularly with respect to community engagement.

As auditing develops as a research method in the algorithmic context, we would do well to heed this history, and to acknowledge the trade-offs that different approaches entail. How expansive should we be in defining discrimination, and what *can't* we measure using audit methods? What are the advantages of documenting the experiences of actual people, even if doing so diminishes our ability to make rigorous causal claims? What forms of knowledge and evidence should we be trying to create to support affected groups, and how should we engage people meaningfully in research about discrimination? The answers to these questions are not straightforward—but considering lessons from the past can help us reflect critically on the role of algorithmic auditing in fostering accountability and change.

ACKNOWLEDGMENTS

We thank the John D. and Catherine T. MacArthur Foundation for support and Cornell University's Artificial Intelligence, Policy, and Practice (AIPP) Initiative and David Pedulla for valuable feedback.

REFERENCES

[1] Ada Lovelace Institute. 2020. Examining the Black Box. https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/.
[2] Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2018. Auditing Black-Box Models for Indirect Influence. *Knowledge and Information Systems* 54, 1 (2018), 95–122.
[3] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination through Optimization: How Facebook's Ad Delivery can Lead to Biased Outcomes. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–30.
[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
[5] Joshua Asplund, Motahhare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. AAAI Press, Menlo Park, CA, 24–35.
[6] O'Neil Risk Consulting & Algorithmic Auditing. 2020. *Description of Algorithmic Audit: Pre-built Assessments.* Technical Report. O'Neil Risk Consulting & Algorithmic Auditing.
[7] Jef Ausloos and Michael Veale. 2021. Researching with Data Rights. , 136–157 pages.
[8] Delia Baldassarri and Maria Abascal. 2017. Field Experiments across the Social Sciences. *Annual Review of Sociology* 43 (2017), 41–73.
[9] Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–34.
[10] Solon Barocas, Anhong Guo, Ece Kamar, Jacquelyn Krones, Meredith Ringel Morris, Jennifer Wortman Vaughan, Duncan Wadsworth, and Hanna Wallach. 2021. Designing Disaggregated Evaluations of AI Systems: Choices, Considerations, and Tradeoffs. arXiv:arXiv:2103.06076
[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2017. Fairness in Machine Learning. *Nips Tutorial* 1 (2017), 2017.
[12] Marc Bendick Jr and Ana P Nunes. 2012. Developing the Research Basis for Controlling Bias in Hiring. *Journal of Social Issues* 68, 2 (2012), 238–262.
[13] Marianne Bertrand and Esther Duflo. 2017. Field Experiments on Discrimination. *Handbook of Economic Field Experiments* 1 (2017), 309–393.
[14] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review* 94, 4 (2004), 991–1013.
[15] Max Besbris, Jacob William Faber, Peter Rich, and Patrick Sharkey. 2018. The Geography of Stigma: Experimental Methods to Identify the Penalty of Place. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer, Berlin, Germany, 159–177.
[16] Miranda Bogen and Aaron Rieke. 2018. Help Wanted: An Examination of Hiring Algorithms, Equity, and Bias. https://www.upturn.org/reports/2018/hiring-algorithms/.
[17] Danah Boyd, Karen Levy, and Alice Marwick. 2014. The Networked Nature of Algorithmic Discrimination. , 53–57 pages.

[18] Shea Brown, Jovana Davidovic, and Ali Hasan. 2021. The Algorithm Audit: Scoring the Algorithms that Score us. *Big Data & Society* 8, 1 (2021), 1–12.

[19] Joy Buolamwini. 2018. When the Robot Doesn't See Dark Skin. https://www.nytimes.com/2018/06/21/opinion/facial-analysis-technology-bias.html.

[20] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. ACM, New York, NY, 77–91.

[21] Frances Cherry and Marc Bendick. 2018. Making it count: Discrimination Auditing and the Activist Scholar Tradition. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer, Berlin, Germany, 45–62.

[22] Rumman Chowdhury and Jutta Williams. 2021. Introducing Twitter's First Algorithmic Bias Bounty Challenge. https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

[23] Charles Crabtree and Volha Chykina. 2018. Last Name Selection in Audit Studies. *Sociological Science* 5 (2018), 21–28.

[24] Jeffrey Dastin. 2018. Amazon Scraps Secret AI Recruiting Tool that Showed Bias against Women. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

[25] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (2015), 92–112.

[26] Nicholas Diakopoulos. 2015. Algorithmic Accountability: Journalistic Investigation of Computational Power Structures. *Digital journalism* 3, 3 (2015), 398–415.

[27] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism.* MIT Press, Cambridge, MA.

[28] Melanie Ehrenkranz. 2019. Amazon's Face Recognition Tech Once Again Pegs Politicians as Criminals. https://gizmodo.com/amazons-face-recognition-tech-once-again-pegs-politicia-1837215790.

[29] Kathleen M Eisenhardt and Melissa E Graebner. 2007. Theory Building from Cases: Opportunities and Challenges. *Academy of Management Journal* 50, 1 (2007), 25–32.

[30] Amit Elazari Bar On. 2018. We Need Bug Bounties for Bad Algorithms. https://www.vice.com/en/article/8xkyj3/we-need-bug-bounties-for-bad-algorithms.

[31] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be Careful; Things can be Worse than they Appear": Understanding Biased Algorithms and Users' Behavior around Them in Rating Platforms. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 11. AAAI Press, Palo Alto, California, 62–71.

[32] S Michael Gaddis. 2017. How Black are Lakisha and Jamal? Racial Perceptions from Names used in Correspondence Audit Studies. *Sociological Science* 4 (2017), 469–489.

[33] S Michael Gaddis. 2018. *An Introduction to Audit Studies in the Social Sciences.* Springer, Berlin, Germany. 3–44 pages.

[34] Clare Garvie. 2019. Garbage In, Garbage Out. https://www.flawedfacedata.com/.

[35] Micah Gell-Redman, Neil Visalvanich, Charles Crabtree, and Christopher J Fariss. 2018. It's all About Race: How State Legislators Respond to Immigrant Constituents. *Political Research Quarterly* 71, 3 (2018), 517–531.

[36] Mark Granovetter. 1995. *Getting a Job: A Study of Contacts and Careers.* University of Chicago Press, Chicago, IL.

[37] Asela Gunawardana and Guy Shani. 2009. A Survey of Accuracy Evaluation Metrics of Recommendation Tasks. *Journal of Machine Learning Research* 10, 12 (2009), 2936–2962.

[38] Jonathan Guryan and Kerwin Kofi Charles. 2013. Taste-Based or Statistical Discrimination: the Economics of Discrimination Returns to its Roots. *The Economic Journal* 123, 572 (2013), F417–F432.

[39] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from Taskrabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, New York, NY, 1914–1933.

[40] David Heinemeier Hansson. 2019. The AppleCard is such a fucking sexist program. My wife and I filed joint tax returns, live in a community-property state, and have been married for a long time. Yet Apple's black box algorithm thinks I deserve 20x the credit limit she does. No appeals work. https://twitter.com/dhh/status/1192540900393705474.

[41] Anna Lauren Hoffmann. 2019. Where Fairness Fails: Data, Algorithms, and the Limits of Antidiscrimination Discourse. *Information, Communication & Society* 22, 7 (2019), 900–915.

[42] Lily Hu and Issa Kohler-Hausmann. 2020. What's Sex Got To Do With Fair Machine Learning. arXiv:arXiv:2006.01770

[43] Basileal Imana, Aleksandra Korolova, and John Heidemann. 2021. Auditing for Discrimination in Algorithms Delivering Job Ads. In *Proceedings of the Web Conference 2021*. ACM, New York, NY, 3767–3778.

[44] Lucas Introna and Helen Nissenbaum. 2010. Facial Recognition Technology a Survey of Policy and Implementation Issues. , 8–55 pages.

[45] Jason Slotkin. 2020. Twitter Announces Changes To Image Cropping Amid Bias Concern. https://www.npr.org/sections/live-updates-protests-for-racial-justice/2020/10/02/919638417/twitter-announces-changes-to-image-cropping-amid-bias-concern.

[46] Joshua Kroll, Joanna Huey, Solon Barocas, Edward Felten, Joel Reidenberg, David Robinson, and Harlan Yu. 2017. Accountable Algorithms. *University of Pennsylvania Law Review* 165, 3 (2017), 633–705.

[47] René Mahieu and Jef Ausloos. 2020. Recognising and Enabling the Collective Dimension of the GDPR and the Right of Access.

[48] Manish Raghavan and Solon Barocas. 2019. Challenges for Mitigating Bias in Algorithmic Hiring. https://www.brookings.edu/research/challenges-for-mitigating-bias-in-algorithmic-hiring/.

[49] The Markup. 2020. The Citizen Browser Project: Auditing the Algorithms of Disinformation. https://themarkup.org/citizen-browser.

[50] Emmanuel Martinez and Lauren Kirchner. 2021. The Secret Bias Hidden in Mortgage-Approval Algorithms. https://themarkup.org/denied/2021/08/25/the-secret-bias-hidden-in-mortgage-approval-algorithms.

[51] Alice McIntyre. 2007. *Participatory Action Research.* Sage Publications, New York, NY.

[52] Miller McPherson, Lynn Smith-Lovin, and James M Cook. 2001. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* 27, 1 (2001), 415–444.

[53] Emma Mishel. 2016. Discrimination Against Queer Women in the US Workforce: A Résumé Audit Study. *Socius* 2 (2016), 1–13.

[54] Brent Mittelstadt. 2016. Auditing for Transparency in Content Personalization Systems. *International Journal of Communication* 10 (2016), 4991–5002.

[55] Shahien Nasiripour and Greg Farrell. 2021. Goldman Cleared of Bias in New York Review of Apple Card. https://www.bloomberg.com/news/articles/2021-03-23/goldman-didn-t-discriminate-with-apple-card-n-y-regulator-says.

[56] Alfred Ng. 2021. Can Auditing Eliminate Bias from Algorithms? https://themarkup.org/ask-the-markup/2021/02/23/can-auditing-eliminate-bias-from-algorithms

[57] Devah Pager. 2007. The Use of Field Experiments for Studies of Employment discrimination: Contributions, Critiques, and Directions for the Future. *The Annals of the American Academy of Political and Social Science* 609, 1 (2007), 104–133.

[58] Parag Agrawal and Dantley Davis. 2020. Transparency around Image Cropping and Changes to Come. https://blog.twitter.com/official/en_us/topics/product/2020/transparency-image-cropping.html.

[59] David S Pedulla. 2016. Penalized or Protected? Gender and the Consequences of Nonstandard and Mismatched Employment Histories. *American Sociological Review* 81, 2 (2016), 262–289.

[60] David S Pedulla. 2018. Emerging Frontiers in Audit Study Research: Mechanisms, Variation, and Representativeness. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer, New York, NY, 179–195.

[61] Lincoln Quillian, John J Lee, and Mariana Oliver. 2020. Evidence From Field Experiments in Hiring Shows Substantial Additional Racial Discrimination after the Callback. *Social Forces* 99, 2 (2020), 732–759.

[62] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating Bias in Algorithmic Hiring: Evaluating Claims and Practices. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 469–481.

[63] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joonseok Lee, and Emily Denton. 2020. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM, New York, NY, 145–151.

[64] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 33–44.

[65] Kristin Reichborn-Kjennerud and Signy Irene Vabo. 2017. Performance Audit as a Contributor to Change and Improvement in Public Administration. *Evaluation* 23, 1 (2017), 6–23.

[66] Aaron Rieke, Urmila Janardan, Mingwei Hsu, and Natasha Duarte. 2021. Essential Work: Analyzing the Hiring Technologies of Large Hourly Employers. https://www.upturn.org/reports/2021/essential-work/.

[67] Lauren A Rivera. 2016. *Pedigree: How Elite Students get Elite Jobs.* Princeton University Press, Princeton, NJ.

[68] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing Partisan Audience Bias within Google Search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–22.

[69] Ronald E Robertson, Alexandra Olteanu, Fernando Diaz, Milad Shokouhi, and Peter Bailey. 2021. "I Can't Reply with That": Characterizing Problematic Email Reply Suggestions. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, 1–18.

[70] Matthew J Salganik. 2019. *Bit by Bit: Social Research in the Digital Age.* Princeton University Press, Princeton, NJ.

[71] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. *Data and discrimination: Converting Critical Concerns into Productive Inquiry* 22 (2014), 4349–4357.

[72] Devansh Saxena, Karla Badillo-Urquiola, Pamela Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. arXiv:arXiv:2107.03487

[73] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 59–68.

[74] Maya Sen and Omar Wasow. 2016. Race as a bundle of sticks: Designs that estimate effects of seemingly immutable characteristics. *Annual Review of Political Science* 19 (2016), 499–522.

[75] Mona Sloane. 2021. The Algorithmic Auditing Trap. https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

[76] Mona Sloane, Emanuel Moss, and Rumman Chowdhury. 2021. A Silicon Valley Love Triangle: Hiring Algorithms, Pseudo-Science, and the Quest for Auditability. arXiv:arXiv:2106.12403

[77] Mario L Small and Devah Pager. 2020. Sociological Perspectives on Racial Discrimination. *Journal of Economic Perspectives* 34, 2 (2020), 49–67.

[78] Jacob Snow. 2018. Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots. https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28.

[79] Megan Stevenson. 2018. Assessing Risk Assessment in Action. *Minnesota Law Review* 103 (2018), 303.

[80] Megan K. Stevenson and Jennifer L. Doleac. 2019. *Algorithmic Risk Assessment in the Hands of Humans*. Technical Report. IZA Institute of Labor Economics.

[81] Marilyn Strathern. 2000. *Audit Cultures: Anthropological Studies in Accountability, Ethics, and the Academy*. Psychology Press, Hove, East Sussex, UK.

[82] Brian L Sullivan, Jocelyn L Aycrigg, Jessie H Barry, Rick E Bonney, Nicholas Bruns, Caren B Cooper, Theo Damoulas, André A Dhondt, Tom Dietterich, Andrew Farnsworth, et al. 2014. The eBird Enterprise: an Integrated Approach to Development and Application of Citizen Science. *Biological Conservation* 169 (2014), 31–40.

[83] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *CACM* 56, 5 (2013), 44–54. https://doi.org/10.1145/2447976.2447990

[84] Taylor Telford. 2019. Apple Card Algorithm Sparks Gender Bias Allegations against Goldman Sachs. https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/.

[85] Lindsey B Trimble and Julie A Kmec. 2011. The Role of Social Networks in Getting a Job. *Sociology Compass* 5, 2 (2011), 165–178.

[86] Mike Vuolo, Christopher Uggen, and Sarah Lageson. 2016. Statistical Power in Experimental Audit Studies: Cautions and Calculations for Matched Tests with Nominal Outcomes. *Sociological Methods & Research* 45, 2 (2016), 260–303.

[87] Mike Vuolo, Christopher Uggen, and Sarah Lageson. 2018. To Match or Not to Match? Statistical and Substantive Considerations in Audit Design and Analysis. In *Audit Studies: Behind the Scenes with Theory, Method, and Nuance*. Springer, New York, NY, 119–140.

[88] Ari Ezra Waldman. 2020. Privacy Law's False Promise. *Washington University Law Review* 97, 3 (2020), 773–834.

[89] William F Whyte. 1989. Advancing Scientific Knowledge through Participatory Action Research. *Sociological Forum* 4, 3 (1989), 367–385.

[90] Christo Wilson, Avijit Ghosh, Shan Jiang, Alan Mislove, Lewis Baker, Janelle Szary, Kelly Trindel, and Frida Polli. 2021. Building and Auditing Fair Algorithms: A Case Study in Candidate Screening. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, 666–677.

[91] Margot Haas Wormser. 1950. *How to Conduct a Community Self-Survey of Civil Rights*. Association Press, New York, NY.

[92] Kyra Yee and Irene Font Peradejordi. 2021. Sharing Learnings from the First Algorithmic Bias Bounty Challenge. https://blog.twitter.com/engineering/en_us/topics/insights/2021/learnings-from-the-first-algorithmic-bias-bounty-challenge.

[93] Kyra Yee, Uthaipon Tantipongpipat, and Shubhanshu Mishra. 2021. Image Cropping on Twitter: Fairness Metrics, their Limitations, and the Importance of Representation, Design, and Agency. arXiv:arXiv:2105.08667

[94] John Yinger. 1998. Evidence on Discrimination in Consumer Markets. *Journal of Economic Perspectives* 12, 2 (1998), 23–40.

[95] Tukufu Zuberi and Eduardo Bonilla-Silva. 2008. *White Logic, White Methods: Racism and Methodology*. Rowman & Littlefield Publishers, Lanham, MD, Chapter 1, 3–30.

[96] Tukufu Zuberi and Eduardo Bonilla-Silva. 2008. *White Logic, White Methods: Racism and Methodology*. Rowman & Littlefield Publishers, Lanham, MD.