Making Algorithms Public: Reimagining Auditing from Matters of Fact to Matters of Concern

R. STUART GEIGER♦
UDAYAN TANDON
ANOOLIA GAKHOKIDZE
LIAN SONG
LILLY IRANI[1]
University of California, San Diego

Stakeholders concerned with bias, discrimination, and fairness in algorithmic systems are increasingly turning to audits, which typically apply generalizable methods and formal standards to investigate opaque systems. We discuss four attempts to audit algorithmic systems with varying levels of success—depending on the scope of both the system to be audited and the audit's success criteria. Such scoping is contestable, negotiable, and political, linked to dominant institutions and movements to change them. Algorithmic auditing is typically envisioned as settling "matters-of-fact" about how opaque algorithmic systems behave: definitive declarations that (de)certify a system. However, there is little consensus about the decisions to be automated or about the institutions automating them. We reposition algorithmic auditing as an ongoing and ever-changing practice around "matters-of-concern." This involves building infrastructures for the public to engage in open-ended democratic understanding, contestation, and problem solving—not just about algorithms in themselves, but the institutions and power structures deploying them. Auditors must recognize their privilege in scoping to "relevant" institutional standards and concerns, especially when stakeholders seek to reform or reimagine them.

Keywords: algorithms, artificial intelligence, auditing, transparency, discrimination, activism

R. Stuart Geiger: sgeiger@ucsd.edu
Udayan Tandon: utandon@eng.ucsd.edu
Anoolia Gakhokidze: annygakhokidze@gmail.com
Lian Song: lhsong@ucsd.edu
Lilly Irani: lirani@ucsd.edu
Date submitted: 11-19-2022

[1] All authors thank Kathryne Metcalf, Magdalena Donea, Julia Kott, Jennifer Chien, Emma Jablonski, and anonymous peer reviewers for feedback on this manuscript, and Christian Sandvig, Karrie Karrahalios, Meg Young, Peaks Krafft, Mike Katell, Michael Ekstrand, Katie Shilton, Chris Kelty, Kristen Vaccaro, Mary Anne Smart, Aaron Broukhim, and Tyus Liu for conversations that informed this scholarship.

Algorithmic systems—from simple rule-based scripts to complex machine learning models—are growing in capacity and delegated key decisions within sectors including employment, education, finance, criminal justice, social media, journalism, and more. While their developers and allies often present these opaque systems as more "objective" than humans, they come with debatably unintended consequences, including reproducing inequalities (Eubanks, 2018; Pasquale, 2015). In response, developers, researchers, impacted communities, policymakers, nonprofits, journalists, and activists have turned to *audit studies* to interrogate algorithmic systems for bias, fairness, and related issues.

While stakeholders often turn to audits for definitive answers, we argue that such methods can fail to deliver that certainty, yet they can simultaneously deepen a collective understanding. We examine the scope, range, and limits of audits of algorithmic systems. What constitutes an audit—of an algorithm or an organization's use of an algorithm? What are the goals of audits, and how are they intended to work? Furthermore, do they work? A dominant assumption is that auditors adjudicate questions about how an opaque algorithmic system behaves, analogous to those who investigate medicines, consumer products, environmental impacts, or embezzlement. While some algorithmic audits have led to meaningful changes, others have failed to settle the issue, generated controversy, and raised more questions than they answered.

We discuss four attempts to audit algorithmic systems with varying levels of success—depending on what constitutes success. Using concepts from science and technology studies, we examine audits as knowledge production efforts and demonstrate how audits differ in terms of the scope of the inquiry and the system to be audited. This scoping is necessarily contestable, negotiable, local, and political.

We use a sociotechnical systems approach (Alkhatib & Bernstein, 2019; Bijker, 1995; Burrell, 2016; Elish, 2019; Radiya-Dixit & Neff, 2023; Rakova & Dobbe, 2023; Seaver, 2019) to discuss the scope of the system: Is the audit of an abstract algorithm or an organization deploying that algorithm in context? We use Bruno Latour's (2004) distinction between "matters-of-fact" and "matters-of-concern" (p. 1) to discuss the scope and role of the inquiry. A matter-of-fact approach to a problem seeks to definitively settle a specific empirical question, whereas a matter-of-concern approach embraces competing views of what the question is and how to investigate it.

The dominant approach to algorithmic auditing positions audits as settling matters-of-fact, focusing on generalizable methods and formal standards—typically mathematical and statistical—that can be applied to any system to determine if it is problematic or not. These standards and methods are often presented as "guarantees" (Corbett-Davies & Goel, 2018, p. 1) of fairness or related concepts: If followed, they ensure systems are good or at least benign. Such methods can be used in a narrower "matters-of-fact" mode, in which all potential issues must be reducible to formal generalizable standards of fairness, discrimination, etc. In this mode, the goal is usually to arrive at a system that is certified as passing; if so, success is declared, and stakeholders are expected to drop objections. All of our cases failed to resolve a stakeholder issue, yet within a "matters-of-concern" lens, they could be seen as successes in how they opened up and made space for broader concerns and negotiations around the institutional use of algorithmic systems.

Understanding audits as matters-of-concern is crucial given that auditing is framed as producing information that will lead to accountability. However, transparency is a resource, not a guarantee of accountability (Ananny & Crawford, 2018; Donia, 2022; Eyert & Lopez, 2023; Irani & Marx, 2021; Metcalf, Singh, Moss, Tafesse, & Watkins, 2023). The matters-of-concern lens reframes concerns like fairness or discrimination as irresolvable through universal methods that provide definitive guarantees. Instead, it emphasizes how standards are locally negotiated in specific institutional contexts of use and are never guaranteed (Lampland & Star, 2009). We do not argue that it is necessarily unwise or impossible to empirically investigate whether an algorithm's behavior violates a certain formal definition of an un/desirable concept like fairness or discrimination. Instead, we show the limits of scoping this kind of work too narrowly, especially when more attention is placed "passing" an audit rather than its substantive outcomes. We argue for a matter-of-concern approach that emphasizes the kind of ongoing local oversight and continual renegotiation that is core to participatory democratic governance.

Background: What is Auditing?

Audit studies emerged alongside the 1960s–70s civil rights movements. New laws prohibited racial discrimination in housing and employment, but adherence and enforcement varied. Audit studies were controlled experiments to reveal discrimination. For example, auditors sent actors who differed in race to apply for housing, using scripts to inquire about the same kinds of units and giving similar incomes and backgrounds. On average, Black applicants were rejected more often and quoted higher rents (Saltman, 1975).

Many contemporary algorithmic auditors cite this lineage. The influential "Auditing Algorithms" (Sandvig, Hamilton, Karahalios, & Langbort, 2014) introduce such audits as "the most prevalent social scientific method for the detection of discrimination" (p. 6). Algorithmic auditors often compare the "black boxes" of opaque AIs, humans, and organizations. When one lacks internal access to decision-making systems (algorithmic, mental, or organizational), such experiments can reveal intentional or unintentional discrimination. Standard approaches compare statistics such as false positive rates by categories like race and gender or compare outcomes of similar cases that differ only by such categories. There is a proliferation of formal methods and definitions in the algorithmic fairness literature (Narayanan, 2018).

Controversies Over the Sociotechnical Systems, of Which Algorithms Are but One Part

The statistical methods used to (dis)claim bias with certain degrees of confidence are generalizable and have been used for decades in fair housing and employment audits. A canonical classroom example is to determine if a coin is biased after observing several coin flips. In a humorous reflection, statisticians Gelman and Nolan (2002) discuss the difficulty of constructing an inherently biased coin for students to test. Unlike loaded dice, biased coin tosses are typically because of how someone throws a coin (e.g., using a sleight of hand), not the materiality of the coin.

Their quip is an instructive expansion of what "the system" is to audit. One audit may only examine the coin when flipped fairly, while another may examine the system of how someone flips a coin. A sociotechnical systems lens leads us to ask if the audit includes the humans who oversee whether a coin is properly flipped or the legitimacy of deciding by a presumably fair coin flip in the first place. While a statistical audit of coin flipping can expand scope socio-technically to settle matters-of-fact when allegations of a biased coin and/or flipper arise, the issue of whether it is "fair" to decide an issue by flipping a "fair" coin is outside the statistical paradigm and not reducible to matters-of-fact—although it can become central as a matter-of-concern.

Scholars advocate not just studying algorithms in themselves, but also unpacking the heterogeneous "algorithmic systems" (Seaver, 2019, p. 412) that must exist for the algorithms to work as intended in practice (Burrell, 2016; Dourish, 2016; Raji, Kumar, Horowitz, & Selbst, 2022). Many systems have "messy boundaries" (Keyes & Austin, 2022, p. 8), are dynamic, updated based on new data, and reliant on human labor and discretion, including entering, cleaning, and labeling data (Gray & Suri, 2019; Irani & Silberman, 2013).

From a matter-of-fact perspective, an audit can be criticized for how auditors bound their inquiry into the system; a coin's physical behavior is irrelevant if the real concern is slight-of-hand. Metaxa et al. (2022) discussed how candidate screening company HireVue hired an independent auditing firm, which found no bias. HireVue did not allow the audit to cover its most controversial features, such as facial analysis and employee performance predictions (Engler, 2021). However, from a matter-of-concern perspective, the audit can be seen as the first of many efforts to build interest and capacity for accountability efforts. The matter-of-concern perspective reframes dissatisfaction with the audit's restricted scope as elaborating concerns about algorithmic job-screening systems.

Controversies Over the Scope of the Inquiry

To observe differential outcomes, auditors must formally define categories of difference. Identity categories are contextual, complex, contested, and deployed in contradictory ways (Abdu, Pasquetto, & Jacobs, 2023; Hanna, Denton, Smart, & Smith-Loud, 2020; Keyes, 2018; Sandvig, Hamilton, Karahalios, & Langbort, 2016). For example, what does it mean to be of a certain race? Is race a self-identified racial identity in which the "correct" result is the subject's self-reported race? Or is race skin tone/color, which is easier to standardize but does not cleanly map to racial identity? Or is race observed by a data labeler, meaning that the "correct" result depends on the labeler's cultural context?

From a matter-of-fact perspective, an audit can be dismissed for inappropriate formalization of demographic categories. Yet, from a matter-of-concern perspective, disagreement over demographic categories is an opportunity to expand a shared understanding of not just the algorithmic system, but also how people relate differently to the institutions that design, operate, and are supported by the system.

Another tension is over which categories to audit. An expanded scope can be seen in Twitter's algorithmic audit competition around its image-cropping algorithm, launched after a viral tweet showed it disproportionately cropped out Black faces. "Everyday auditors" (Shen, DeVos, Eslami, & Holstein, 2021, p. 1) responded by testing for and finding similar biases against images of people who are elderly, disabled, heavy, or included Arabic script.

Actually Existing Audits: From Matters-of-Fact to Matters-of-Concern

To illustrate the difference between audits as matters-of-fact versus matters-of-concern, we discuss four attempts to audit algorithmic decision making in high-risk contexts: Facial recognition, criminal recidivism prediction, job candidate screening, and gunshot detection. We chose these cases because all involved attempts by various stakeholders to better understand how a high-risk opaque algorithmic system operates and to hold its developers and institutional users accountable for its proper operation. All audits were initially organized around settling matters-of-fact, but ultimately raised matters-of-concern, albeit in different ways. The systems audited differ in terms of their underlying technology, general purpose, deployed context of use, and sociotechnical complexity. All audits resulted in significant public engagement around the audit, although they differed in the origin of the audit, auditors' backgrounds and professions, methods and standards applied, the kinds of public engaged, and the degree and criteria of success of the audit.

For each of these cases (Table 1), we describe the algorithm in its sociotechnical context, discuss the audit(s) performed on the system, and then examine the framing and reception of the audit among wider publics. We argue that seeing each audit as a matter-of-concern, rather than as settling matters-of-fact, better accounts for the processes of public knowledge production and contestation that unfolded.

Table 1. Cases and Overview of Findings.

Case	**What was the suspected issue with the system?**	**What fact did the auditors find about the system?**	**What was the broader issue in conflict around the system?**
Gender Shades facial recognition	Accuracy, especially intersectionally by gender and race	High accuracy for "lighter-skinned males," low accuracy for "darker-skin females" (p. 1)	The use of highly accurate facial recognition in surveillance and policing
COMPAS recidivism/machine bias	False positive and negative rates, especially by race; many other issues	Black defendants were twice as likely to be incorrectly predicted to be future criminals; many implementation and validity concerns	Contradicting definitions of fairness; using cases from the biased past to decide current cases; many deployment issues
Pymetrics job candidate screening	Differential acceptance rates beyond EEOC's four-fifths rule	No bias above threshold according to Pymetrics' methods, but no intersectional analysis	Can stakeholders trust audits funded, scoped, and co-authored with auditees? What disclosures should publications require?
Shot Spotter gunshot detection	False positives mistaking city noises for gunshots; privacy; overpolicing	Full audit resisted by company. External audit would be dangerous and infeasible.	How transparent should developers of surveillance be about their products? Is it a privacy violation? Will it lead to overpolicing?

Case 1: Facial Processing Technology and the Gender Shades Audit

Facial processing technologies (FPTs) are widely deployed by law enforcement and employers to detect faces, from social media photo tagging to surveillance. Cloud computing providers such as Amazon and Microsoft offer general-purpose FPT services. Some uniquely identify faces, while others only make demographic classifications. Organizations can use the same FPTs quite differently, impacting the scope of an audit.

In the Gender Shades project, Buolamwini and Gebru (2018), both Black women, investigated three FPTs that classify faces as "female" or "male." They tested accuracy rates across gender presentation and skin tone, with intersectionality in mind. The audit found that all systems were less accurate in assigning the "correct" label for faces that were both darker and more feminine. Buolamwini and Gebru (2018) attributed this bias to developers not training the model using demographically balanced data—a mistake that more diverse developer teams may have avoided.

The audit focused on matters-of-fact, discussing strategies for evaluating and improving the differential accuracy rates. Their stated value was that performance for all subgroups should be as similar and as close to 100% as possible. They only focused on the inputs and outputs of commercial gender classifiers without investigating the use of FPTs. The audit circulated as an academic paper at a major computer science conference on fairness, accountability, and transparency (ACM FAccT) and rapidly spread through videos and social media. It was widely covered by journalists, leading to a documentary (*Coded Bias*). None of the companies contested the matters-of-fact presented in the audit. Microsoft and IBM issued statements accepting the results and pledging to improve their systems. This aligns with their stated vision of accountability: Once evidence of bias is settled as a matter-of-fact, various mechanisms will hold companies accountable.

Although the developers accepted the audit as a settled matter-of-fact, others noted how FPTs were implemented in policing and other surveilling institutions. Critics who sought reform or abolition of surveillance and/or policing circulated the results as matters-of-fact that supported their critical position. However, they raised concerns about improving the accuracy of FPTs. Hamid (2020) argued that such systems are disproportionately deployed in minority neighborhoods. This leads to a self-fulfilling cycle in which more observations of suspicious activity send more police into such neighborhoods, increasing official observed crime rates, which have socioeconomic impacts and legitimize more policing (Barabas, Beard, Dryer, Semel, & Solomun, 2020; Scannell, 2019; Shapiro, 2019). Rather than debiasing general-purpose FPTs, critics argued that focus should be directed at the deployed context of use around issues like overpolicing and toward defunding/dismantling carceral surveillance technologies altogether.

A Gender Shades Frequently Asked Questions published shortly after reflects tensions between the audit's role around settling matters-of-fact versus raising matters-of-concern. Their response to "Is your goal to improve facial analysis technology?" states "even flawless facial analysis technology . . . can still be abused" (Gender Shades, n.d., p. 15). Next, they recommend not using FPTs that have not been audited but noting that "citizens should be given an opportunity to decide if this kind of technology should be used" (Gender Shades, n.d., p. 15).

Two years later, Gebru and an expanded team performed a second project to audit FPTs, Saving Face (Raji, Gebru et al., 2020). The authors found no significant differences in accuracy by gender presentation and/or skin color for the systems they audited two years earlier—a major success in one sense. However, Saving Face also went beyond matters-of-fact and raised concerns around FPTs, including design considerations and ethical tensions. They noted that just because a system passes an audit does not mean it is good; its statistical audit of accuracy should be a "low bar not to be caught tripping over" (Raji, Gebru et al., 2020, p. 150). They argued that auditors should expand the scope of inquiry by interrogating the system's deployment in context. They argued that auditors should publish information on the "limit of the audit's scope, and the context in which results should be interpreted and appropriately acted upon" (Raji, Gebru et al., 2020, p. 150).

Most importantly, the work holding FPT developers and users accountable did not end with audits. Buolamwini gave a popular TED talk about bias in FPTs (Buolamwini, 2016) and testified before the U.S. Congress (Buolamwini, 2019). The Gender Shades audit helped build awareness about the existence of FPTs and bias in algorithmic systems of all kinds. This lent credibility to later calls for other government agencies to stop using similar technologies (Buolamwini, 2022). The momentum that was built up around these findings and issues expanded the team's legitimacy as experts in FPTs and many other algorithmic technologies. Both have founded organizations dedicated to critically investigating harms around algorithmic systems of all kinds, especially toward marginalized groups (the Algorithmic Justice League and the Distributed Artificial Intelligence Research Institute).

Case 2: COMPAS Recidivism Prediction and Machine Bias Audit

COMPAS was designed for criminal courts to predict recidivism: The likelihood that a defendant, if released, will be charged with another crime. COMPAS produces "risk scores" shown to judges when determining if defendants should be released or held in jail until trial. To use COMPAS, an interviewer asks the defendant up to 137 survey questions, including questions about their childhood, education, housing, how many friends have been arrested, boredom, how they feel others see them, or if the law helps average people. These responses are combined with their criminal records, such that defendants will have higher risk scores if their answers are closer to defendants who reoffended, but a lower risk score if their answers are closer to defendants who did not reoffend. The defendants in the comparison groups are historical cases curated by the company, which offers comparisons to updated and local cases for additional fees (Equivant, 2019).

ProPublica's Machine Bias project (Angwin, Larson, Mattu, & Kirchner, 2016) made headlines alleging racial biases in COMPAS. Their core claim was based on false positive and negative rates: They examined defendants who scored with COMPAS in 2013–2014 and then reviewed criminal records after two years to see if the prediction was correct. When comparing defendants with similar records and backgrounds, Black defendants were almost twice as likely to be incorrectly classified as "high risk" future criminals and unnecessarily held in jail until trial.

While Gender Shades audited a general-purpose system independent of its deployed context, the Machine Bias auditors designed their audit specifically around how COMPAS was deployed in Broward

County, Florida. Auditors' claims about unequal error rates have been extensively debated as an unresolved matter-of-fact within the statistical auditing paradigm. We focus on their expanded matters-of-concern approach, with an open-ended investigation that holistically examines COMPAS's deployment. First, they critique how "reoffending" is defined as rearrested and charged, not convicted, with biases in who is arrested and/or charged. They critique intake questions, which means that defendants' freedom depends on whether their personal beliefs or childhoods are too similar to those incarcerated. They show how judges selectively interpret scores in different contexts, undermining claims that COMPAS leads to more "objective" decisions. Finally, they found errors in imported criminal records. These show the importance of how auditors scope the system and their audits.

COMPAS's developer, Northpointe (now Equiviant), quickly challenged the audit as a matter-of-fact. Rebuttals focused on false positive/negative rates, which Northpointe said was the wrong statistical metric, as their algorithm was designed for predictive parity. This means that COMPAS produces similar risk scores based on defendants' likelihood of being rearrested and charged, regardless of racial grouping. When, as in Broward County, Black defendants on bail were almost twice as likely to be rearrested, should a model "correctly" anticipate a higher risk of rearrests for Black defendants? Given the rising attention to racial bias in policing, critics asked: To what extent is the system predicting the criminal behavior of the defendant versus predicting who police are more likely to arrest? (Mulligan, Kroll, Kohli, & Wong, 2019; See Green, 2020).

Computer scientists debated the conflict as a mathematical paradox or impossibility theorem in reconciling different ways of measuring whether a system is fair when there are unequal base rates between demographic groups (Chouldechova, 2017; Kleinberg, Mullainathan, & Raghavan, 2016). Computer science venues became inundated with papers about COMPAS, arguing for different formal definitions of various concepts and how to measure them. As Green (2020) argues, what computer scientists called a paradox or impossibility theorem is better understood as a conflict between different understandings of the criminal justice system, e.g., whether one believes that historical crime data is "colorblind" and objectively reflects criminal activity or if it reflects biased policing practices. This is outside the scope of formal methods and requires specific contextual expertise.

In contrast, many outside of computer science responded in ways that included the matter-of-fact claim about error rates, but also focused on the many other concerns that are not reducible to a single metric. The expose began a wave of academic research, legal challenges, investigative journalism, and political organizing about COMPAS and similar commercial software tools. These systems were targeted by civil rights organizations, digital rights advocacy organizations, and AI ethics organizations. From critics' perspective, there is no mathematical metric or technical fix that will make it a good idea to model the future of criminal justice in its past or to use psychometric survey questions to decide who should be incarcerated. Several lawsuits contested the constitutionality of COMPAS, but these challenges failed, as the system was found to make "recommendations" rather than determinations (Brenner et al., 2020).

Neither ProPublica's audit nor Northpointe's response settled matters-of-fact, but the COMPAS audit certainly generated matters-of-concern. While COMPAS had been in use since 2001, the existence of such algorithms in criminal justice came as a surprise to many, whose first introduction to COMPAS was through the audit. The 2016 audit was launched during the Black Lives Matter movement, which focused on race in policing and criminal justice, as well as public attitudes toward technology shifting from utopian to dystopian. Teachers and researchers included the ProPublica audit in reading lists about both racial justice and algorithmic auditing, expanding the audience for debates beyond those seeking to make an algorithm a site of veridiction.

Case 3: Pymetrics Job Candidate Screening System and Independent-but-Cooperative Audit

Pymetrics provides job candidate screening services to companies. While most hiring is based on resumes, Pymetrics uses proprietary psychometric games that they claim are based on psychological studies. They advertise that these games allegedly measure qualities like extroversion or generosity without bias. Pymetrics makes a client company's current ideal employees play the games, then Pymetrics identifies candidates whose gameplay-based data match current employees.

Pymetrics claimed that their "algorithms constantly test for and remove ethnic or gender biases that arise, leading to more women and minority hires" (Ryan, 2018, p. 8). They claim to internally audit models for bias, deploying only the best-performing models that also meet a fairness metric called the four-fifth rule, thus guaranteeing fairness. In the United States, Title VII of the 1964 Civil Rights Act prohibited hiring practices with "disparate or adverse impact" (Equal Employment Opportunity Commission [EEOC], 1979, p. 11996) on certain protected classes. Congress delegated the definition of this term to the Equal Employment Opportunity Commission (EEOC), which created the four-fifths rule.

This rule states that if employers use a test or procedure that does not measure bona fide occupational qualifications, a greater than 20% difference in pass rates violates the rule. If 10% of men but only 5% of women pass, this is a 50% difference and unacceptable; if 10% of men but only 8% of women pass, this is a 20% gap and at the cutoff. However, the EEOC states that this "rule of thumb is not intended as a legal definition" (EEOC, 1979, p. 11998) of discrimination. It is mandatory to self-report to regulators, who then decide whether a company's hiring practices should be further investigated. Pymetrics only audits for discrimination between single-identity categories, not intersectional or subgroup biases, as in the Gender Shades project, which we discuss later.

Computer scientist Christo Wilson and his lab contracted a "collaborative audit" with Pymetrics (Wilson et al., 2021, p. 1), which they introduced by referencing reflections from the Saving Face (Raji, Gebru et al., 2020) follow-up audit to Gender Shades. While Raji, Gebru et al. (2020) stressed the importance of external audits, Wilson et al. (2021) argued for a quasi-independent approach. Pymetrics granted the team access to source code, documentation, and internal data. Companies usually guard such information as intellectual property. Wilson's team signed nondisclosure agreements, but was allowed to publish all contracts, the negotiated scope of work, budgets, and other documents (Wilson, 2022).

Like ProPublica's COMPAS audit, the Pymetrics audit did not just seek to establish a single statistical fairness metric as a matter-of-fact. The audit included checking the candidate screening process and source code for correctness and vulnerabilities. This included validating that the internal audit program's source code implemented the proper statistical tests, that models did not directly input demographic data, and various quality control checks for human error and sabotage. Such an expanded scope was possible because of the collaboration, although Wilson et al. (2021) noted that Pymetrics could have changed its source code or procedures since its audit.

The final article—co-authored by Wilson's team and the Pymetrics team—declared that the system passed the audit. However, specific criteria were excluded from the audit at the beginning. One issue is the psychometric games that allegedly predict fit for a job; the auditors take this as a given and "do not comment on the rationality and ethics of using these measures to evaluate a candidate's suitability for employment" (Wilson et al., 2021, p. 670). The auditors also did not examine intersectional or subgroup biases, because they claimed enforcing the four-fifth rule intersectionally could lead to "selecting a less performant model" (i.e., less accurate) and "intersectionality is not recognized by the relevant regulatory agencies" (Wilson et al., 2021, p. 675). However, many local U.S. and non U.S. jurisdictions explicitly prohibit intersectional discrimination, including New York City, San Francisco, Canada, and across Europe (Davis, 2022). Second, there is legal ambiguity over Title VII, as federal courts have disagreed over intersectional discrimination claims (Beck, 2022). The EEOC's guidance states that while there "is no obligation to make comparisons for subgroups . . . any apparent exclusion of a subgroup may suggest the presence of discrimination" (EEOC, 1979, q. 17).

After the audit's publication, Pymetrics publicly advertised itself as independently audited, referencing Wilson et al.'s (2021) paper and the ACM FAccT conference that published it. Despite the paper stating that the audit was collaborative and scoped by Pymetrics, Pymetrics described Wilson and his team as third parties who were given the freedom to audit as they saw fit: "They had access to the codebase, data, and a representative set of models, and ran their own statistical tests" (Pymetrics, 2021, p. 3). This was subsequently critiqued by researchers in the algorithmic auditing field, particularly within FAccT.

Young, Katell, and Krafft (2022) critiqued it as a symbolic exercise in labeling Pymetrics as audited, with the conference complicit. They argued that auditors' and auditees' interests are in conflict: Auditees favor narrower audits they will pass, whereas auditors want to examine deeper and broader issues. They call the integrity of the audit into question, given the inclusion of Pymetrics' staff as coauthors and taking Pymetrics' internal standards and assumptions as given. This case instigated organizing within FAccT to institute financial and conflict-of-interest disclosures for paper authors and leadership. The FAccT organizers instituted a disclaimer that "Any product or service evaluated in any of these articles, or any claim therein, is not guaranteed or endorsed by FAccT, which is not an auditing organization" (ACM FAccT Conference, 2022, p. 1).

From a matter-of-fact perspective, the audit has not settled the status of Pymetrics, but from a matter-of-concern perspective, it catalyzed a reevaluation of the standards, functions, and trustworthiness of audits, auditors, and auditees. Critics who organized against the audit were ambivalently invested in the idea of an independent audit that could establish matters-of-fact, even as they unsettled the limited scope of this particular audit. The case also inspired other auditors to develop new approaches to better ensure their independence (Costanza-Chock, Raji, & Buolamwini, 2022). These recommendations include accreditation procedures for auditors, harm incident reporting mechanisms, increased involvement from other stakeholders, and mandatory disclosure.

Case 4: ShotSpotter Gunshot Detection System and Communities' Capacity for Oversight

ShotSpotter embeds neighborhoods with always-on microphones that triangulate locations of gunshot-like sounds. This information purports to enable improved police responses to gunshot events. The company portrays its product as a neutral, race-blind technical solution for the benefit of terror-stricken communities that do not reliably report gunfire in their own neighborhoods (Clark, 2017). However, it has been widely criticized by local and national advocacy groups over issues of privacy and discriminatory deployment (Guariglia, 2021; Winkley, 2016). Beyond the privacy issue, a major concern is that ShotSpotter's acoustic technology is error-prone.

The company includes disclaimers in contracts with police departments that the system can mistake, say, the sound of a car backfiring, fireworks, or a recording of gunshots for gunshots (ShotSpotter, 2021). Publicly, they claim 97% accuracy, but one city contract guarantees only 80% accuracy (Columbia Police Department, 2019). Several high-profile cases involved prosecutors using false positives to hold people in jail for extended periods of time, before dropping charges against them (Kang & Hudson, 2022).

Similar to the COMPAS recidivism case, there was also concern that ShotSpotter may contribute to a vicious cycle of data criminalization. ShotSpotter is sold as a tool for "solving" crime in poor and minority communities; It is rarely installed city-wide. Marginalized neighborhoods are disproportionately represented in policing data, thus justifying future overpolicing, producing still more data, and so on (Barabas et al., 2020; Scannell, 2019; Shapiro, 2019).

ShotSpotter expanded in the same period that the Black Lives Matter movement raised concerns about race and policing. Within communities organized around policing and surveillance, an open question is whether to call for audits of ShotSpotter's accuracy. A cornerstone of audit studies is the analysis of differential outcomes through empirical mechanisms such as matched pairs studies and group-level differences in error rates. However, the isolation of variables for experimental input and output is nearly impossible without cooperation from ShotSpotter, which has resisted auditing efforts. For example, an auditor could travel to specific locations, either fire a gun or play a similar noise, and then measure if police investigate—an incredibly risky approach.

ShotSpotter has resisted auditing efforts and does not make the data it offers police readily available to the communities it surveils, contractually withholding data as proprietary trade secrets (ShotSpotter, 2015). ShotSpotter offered the American Civil Liberties Union (ACLU) a confidential review only of the company's source code, but no data. The ACLU declined, citing both a lack of resources and the fact that examining source code without data would not settle the issue. Instead, the ACLU analyst recommended "a broader systems audit" (Stanley, 2015, p. 10) by an independent firm, which would likely raise similar debates and negotiations over scope as the Pymetrics audit.

Notably, ShotSpotter discussed the audit's goal as "assur[ing] populations of the narrow focus of these microphones" (Stanley, 2015, p. 10), such as their claim that they do not store conversations. ShotSpotter's sought to settle specific matters-of-fact they decided were the core issues, rather than responding to wider concerns about which neighborhoods and people were subject to its surveillance and automated police alerts. Such offers for "collaborative" audits of ShotSpotter's technical capacity do not capture the broader tension between police and communities that ShotSpotter surveils and transforms.

In contrast, a 2021 study examined these broader concerns using an explicitly community-centered approach (MacArthur Justice Center, 2021). Auditors used a similar strategy as the COMPAS audit: using public records laws to obtain information about real-world cases involving ShotSpotter and then examining patterns in outcomes. They reviewed police records for responses to ShotSpotter alerts over the course of 21 months found that 86% of responses concluded with no reported crime whatsoever. These findings are corroborated in other journalistic accounts of ShotSpotter's use in other cities (Grant, 2020).

In response, a critical report commissioned and publicized by ShotSpotter contested the audit as matters-of-fact, claiming that the public records were "an incomplete source of information" (Edgeworth Analytics, 2021, p. 9). They argued that even when ShotSpotter-dispatched officers did not file crime/incident reports, a gun crime could still have occurred. Such *argumentum ad ignorantiam* applies to any such external audit, as auditors would need unimaginably ubiquitous surveillance data to verify that a gun crime truly did not occur. The report asserted that ShotSpotter nevertheless assists police in ways not documented in such records. An earlier ShotSpotter-commissioned report instead defended ShotSpotter by asking the police if they believed it was accurate and useful, who generally reported that they did (Selby, Henderson, & Tayyabkhan, 2011). ShotSpotter's data thus benefits police in ways the public is told they must simply trust—legitimizing law enforcement's own opacity, insulating ShotSpotter from public oversight, and justifying more policing and surveillance.

Neither ShotSpotter-commissioned report directly engages with the audit's concerns about overpolicing and data criminalization. Instead, their concern with false positives is primarily framed as wasted police labor. Police are presented as neutral evaluators of their institution and historical crime data as an objective way to decide where to deploy ShotSpotter. ShotSpotter thus responded to concerns about "accuracy" in a matter-of-fact approach, seeking to discredit specific empirical claims about the performance and utility of its acoustic classifier, using its own proprietary data. In contrast, auditors deployed "accuracy" to raise matters-of-concern around overpolicing and data criminalization, in which past experiences with police violence and discrimination lead them to an incommensurable way of approaching ShotSpotter, the false positive issue, and the sociotechnical system of surveillance-based policing.

Because of how ShotSpotter made itself an exclusive authority about urban gunfire and thus its own product, any existing external audit of ShotSpotter will always be partial and incomplete. Community groups must negotiate complex processes to gain even partial information from public records, which are subject to the discretion of local bureaucratic processes (Irani & Marx, 2021). Yet, the very steps ShotSpotter has taken to resist public oversight have raised its profile among activists, who see such resolute opacity as inherently suspicious. Despite a failure to enact direct change, these auditing attempts can nevertheless serve as an important site for community organizing to raise attention, build counter-knowledge, and press for change through other venues and tactics.

Conclusion: Making Algorithms Public

As our cases show, algorithmic auditing refers to a wide range of efforts to investigate algorithmic systems and hold their developers and users accountable for various standards and expectations. On one end of a spectrum, auditors can approach their work as testing an undeployed algorithm's abstract behavior for a particular formal definition of an issue like fairness or discrimination, which, if met, certifies the algorithm as fair and takes the issue off the table. On the other hand, auditors can approach their work as building resources, capacity, and venues for ongoing democratic understanding, oversight, negotiation, agenda setting, and problem solving about a wide range of concerns around opaque and evolving sociotechnical systems, as they are deployed in specific contexts and institutions.

In the abstract, there is broad consensus that important algorithms should be audited, but there is little consensus on what an audit should entail. Our four cases all began with similar intentions to investigate potential issues related to the behavior of an opaque decision-making system. They began with similar goals to hold the system's developers and users accountable for any behaviors that deviated either from how the developers' represented the system's behavior or how stakeholders believed such a system ought to behave. However, they raised concerns that could not be reconciled as matters-of-fact, given how the audits and/or systems were initially scoped.

What Can Audits as Matters-of-Fact Achieve?

Some potential issues are easier to answer than others with the kinds of formal generalizable methods and standards that dominate the algorithmic fairness literature and characterize matter-of-fact audits, but crucially, this differs with institutional contexts of use. Some institutions are more stable and consented to than others, and all are subject to negotiation, contestation, and change. Audits as matters-of-fact can potentially verify whether systems align with hegemonic values, where there is strong consensus among all stakeholders on the issues, categories, standards, methods, inputs, outputs, goals, and use of the entire sociotechnical system that deploys the algorithm. When there is not such a stable consensus—as in all our cases—then matter-of-fact audits still proceed as if there is one. This can be a form of institutional erasure, coercion, and violence, particularly if the auditors are positioned such that they are the main gatekeeper determining whether a controversial system is deployed.

The COMPAS case shows how auditors raised concerns in the context of use beyond the scope of formal fairness guarantees, such as intake questions used to correlate risk, selective interpretation by judges, erroneous data, and discrimination in historical arrest rates. The developers insist that COMPAS is necessarily fair because it meets predictive parity metrics. This is incommensurable with ProPublica auditors' multifaceted approach to showing various problems of COMPAS as deployed, then asking the public if this is how a criminal justice system should operate. ProPublica's hyperlocal audit would be labor intensive to scale to all COMPAS deployments, unlike Pymetrics' interpretation and implementation of the four-fifths' rule in their internal audits. However, an expanded matter-of-fact audit would be open to identifying quite different operational concerns in other jurisdictions.

With Pymetrics, the intersectionality issue is inseparable from the complex and inconsistent decades-long history of U.S. employment law. Law can operationalize vague concepts, such as "reasonable" or "disparate," through standards and precedents that lack bright lines and quantitative thresholds demanded by formal metrics and guarantees. Yet, given that intersectional discrimination is expressly prohibited in other jurisdictions (Davis, 2022), critics disagreed with how Pymetrics and Wilson et al. (2021) declared some regulatory concerns to be irrelevant.

Narrowly scoped audits can be deployed to shut down stakeholders' engagement, if auditing is more of a symbolic exercise where it matters more that a system or algorithm is audited than what the audit examined. Algorithms are one entry point in how people seek to transform institutions. As StopLAPDSpying illustrates (Figure 1), algorithms are part of complex sociotechnical systems or ecologies. When stakeholders question "the algorithm," they often call attention to algorithms as deployed in broad ecologies and second-order effects. For example, StopLAPDSpying identifies housing developers and property values as key elements in LAPD's predictive policing ecology. Auditing around matters-of-concern can reveal these broader factors and issues, while auditing around matters-of-fact typically seeks to isolate "the algorithm" from them.

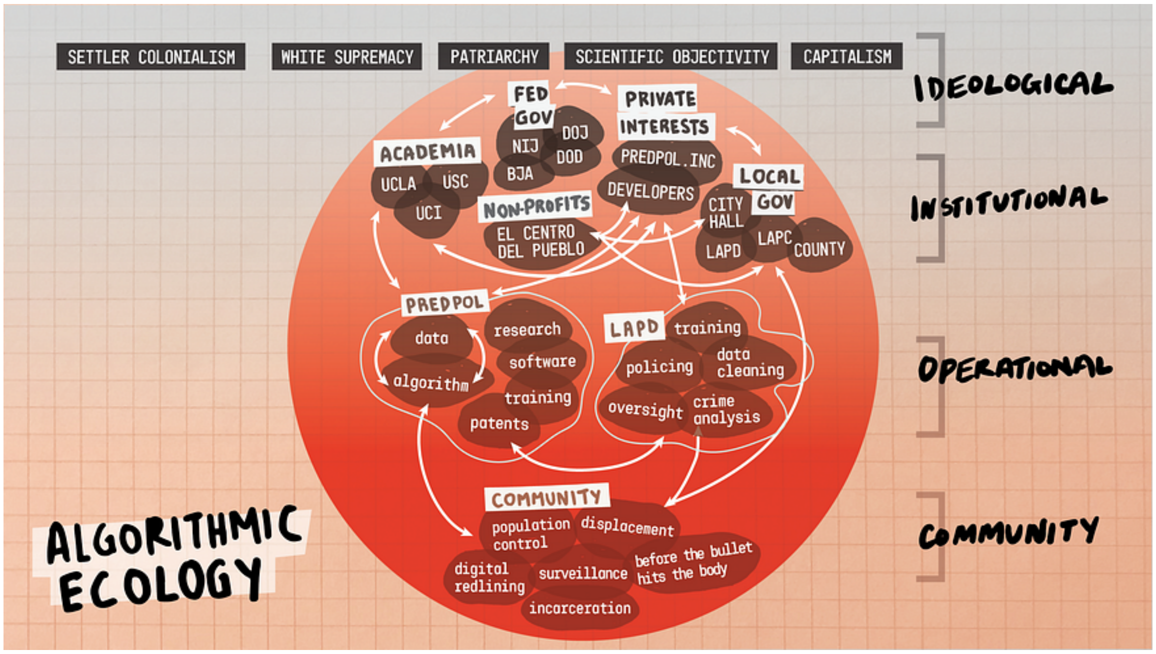

Figure 1. Algorithmic ecology visualization of predictive policing, including ideological, institutional, operational, and community entities (StopLAPDSpying Coalition, 2020).

Auditors who uncritically rely on dominant interpretations of current institutional standards can be complicit in silencing more transformative questions about those institutions and standards and shutting down those whose survival or liberation depends on them (Crooks, 2022). To this end, we emphasize the quite different roles audits can play within broader social, political, economic, and institutional struggles, and call on auditors to recognize their privilege in scoping audits to "relevant" concerns and criteria.

We evaluated the success of our cases in terms of building the capacity for public accountability over opaque systems and institutions. In Table 2, we contrast whether an audit "worked" in a more traditional sense versus whether it built the capacity for public accountability.

Table 2. Comparison of Case Outcomes.

Case	**Did the audit "work"? Did it satisfy stakeholders' concerns by either finding no issues or finding issues that were then fixed?**	**Did the audit build capacity for public accountability over tech and institutions?**
Gender Shades	Yes. The audit found bias in three systems, which was not present in any system after two years.	Yes, the audit gave the authors a platform for a broader justice-oriented movement and widely publicized the concept of algorithmic bias.
COMPAS	No. Auditors claimed to find bias, but developers disagreed. COMPAS is still in widespread use, but company rebranded. Opponents lost legal challenges.	Partial. The audit publicized COMPAS, but inspired far more computer science research about fairness metrics than accountability efforts around COMPAS.
Pymetrics	No. Audit claimed to find no concerns, but generated controversy over its independence and acceptance of auditees' assumptions and standards.	Yes. The audit-catalyzed policy changes within the conference that published it about corporate influence and conflicts of interest.
ShotSpotter	No. Audit is effectively blocked by lack of access to data.	Yes, efforts were conducted within broader social justice movements concerned with policing and surveillance; lack of transparency motivates activists.

Broader Contributions and Connections

Our findings echo similar debates within action research (AR) and participatory design (PD), which involves stakeholders in shaping technology and research. Practitioners debate whether AR or PD methods ensure alignment with stakeholders' needs and perspectives. Like audits, the actual implementation of AR and PD methods varies, from a single workshop led by experts with a convenience sample of participants to long-term collaborations where communities and local organizations set the agenda (Bødker & Kyng, 2018; Brown, 2017; Costanza-Chock, 2020).

Public calls for algorithmic accountability have led to different visions of how institutions should regulate and oversee algorithms. The Algorithmic Accountability Act proposed in the U.S. Congress ("Algorithmic Accountability Act of 2022," 2022) would require companies to conduct and report internal assessments of critical decision-making algorithms to the Federal Trade Commission, which would oversee and summarize the results. Such an Act could steer algorithmic auditing toward matters-of-fact, leading regulators to standardize generalizable formal assessments that certify systems as problematic or benign. The Act could also prompt the FTC to make algorithms public in various ways that can be flexible to new concerns and do not require the public to simply trust that the regulators have found the right matters-of-fact.

For example, the PERVADE project (Shilton et al., 2021) developed a broad triage tool for identifying and documenting dozens of potential concerns in data-intensive systems. Falco et al. (2021) proposed that algorithm developers work under "audit trails" (p. 4), akin to a Flight Data Recorder that automatically collects data for investigators. The FTC could collect, steward, and synthesize many resources for stakeholders, including code and data, system documentation, internal debates, everyday audits (Shen et al., 2021), or testimonies documenting the lived experiences of those impacted. The kinds of accounts made public about algorithms can provoke inquiry, public sensemaking, and accountability. These efforts can also build into wider movements to transform not only algorithms but also the social and institutional practices in which they are deployed (Pasquale, 2020).

Following calls for data agonism (Crooks & Currie, 2021; Young et al., 2022) , we reframe algorithmic accountability as infrastructure for publics to keep assembling around various concerns with algorithmic systems and the institutions using them. With a matter-of-concern approach, there is room for communities to raise, make sense of, and negotiate many potential issues, such as those in our cases, which can be overlooked in a narrowly scoped matter-of-fact approach. Some might object to the inefficiencies of this mode versus more scalable, generalizable methods—capital accumulation seeks to "move fast and break things" (Irani, 2019, p. 16). To that, we respond: precisely.

References

Abdu, A., Pasquetto, I., & Jacobs, A. (2023). An empirical analysis of racial categories in the algorithmic fairness literature. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 1324–1333. doi:10.1145/3593013.3594083

ACM FAccT Conference. (2022). *2022 accepted papers. FAccT 2022*. Retrieved from https://facctconference.org/2022/acceptedpapers.html

Algorithmic Accountability Act of 2022, H.R. 6580, *117th Congress*. (2022). Retrieved from https://www.congress.gov/bill/117th-congress/house-bill/6580/text

Alkhatib, A., & Bernstein, M. (2019). Street-level algorithms. *Proceedings of the 2019 ACM Conference on Human Factors in Computing Systems,* 1–13. doi:10.1145/3290605.3300760

Ananny, M., & Crawford, K. (2018). Seeing without knowing. *New Media & Society, 20*(3), 973–989. doi:10.1177/1461444816676645

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). *Machine bias*. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Barabas, C., Beard, A., Dryer, T., Semel, B., & Solomun, S. (2020). *Abolish the #TechToPrisonPipeline*. Coalition for Critical Technology. Retrieved from https://medium.com/@CoalitionForCriticalTechnology/abolish-the-techtoprisonpipeline-9b5b14366b16

Beck, S. (2022). "We do not live single-issue lives": Bostock v. Clayton County mainstreaming Title VII intersectional discrimination claims. *Minnesota Journal of Law & Inequality, 40*(2), 463.

Bijker, W. (1995). *Of bicycles, bakelites, and bulbs*. Cambridge, MA: MIT Press.

Bødker, S., & Kyng, M. (2018). Participatory design that matters—Facing the big issue. *ACM Transactions on Computer-Human Interaction, 25*(1), 1–31. doi:10.1145/3152421

Brenner, M., Gersen, J., Haley, M., Lin, M., Merchant, A., Millett, R., … Wegner, D. (2020). Constitutional dimensions of predictive algorithms in criminal justice. *Harvard Civil Rights-Civil Liberties Law Review, 55*(1), 267–310.

Brown, A. (2017). *Emergent strategy*. Chico, CA: AK Press.

Buolamwini, J. (2016). *How I'm fighting bias in algorithms*. TEDxBeaconStreet. Retrieved from https://www.ted.com/talks/joy_buolamwini_how_i_m_fighting_bias_in_algorithms

Buolamwini, J. (2019, May 22). *Hearing on facial recognition technology*. 116th Congress. Retrieved from https://www.congress.gov/116/meeting/house/109521/witnesses/HHRG-116-GO00-Wstate-BuolamwiniJ-20190522.pdf

Buolamwini, J. (2022). The IRS should stop using facial recognition. *The Atlantic*. Retrieved from https://www.theatlantic.com/ideas/archive/2022/01/irs-should-stop-using-facial-recognition/621386/

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Proceedings of Machine Learning Research, 81*, 1–15. Retrieved from http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

Burrell, J. (2016). How the machine "thinks." *Big Data & Society, 3*(1). doi:10.1177/2053951715622512

Chouldechova, A. (2017). Fair prediction with disparate impact. *Big Data, 5*(2), 153–163. doi:10.1089/big.2016.0047

Clark, R. (2017). *Let's money ball gun violence reduction*. Quality Policing. Retrieved from https://qualitypolicing.com/violencereduction/clark/

Columbia Police Department. (2019). *Shotspotter gunshot detection and alert system*. Columbiapd.net. Retrieved from https://columbiapd.net/wp-content/uploads/2020/07/General-Order-02.07-ShotSpotter-Gunshot-Detection-and-Alert-System.pdf

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness*. arXiv. doi:10.48550/arXiv.1808.00023

Costanza-Chock, S. (2020). *Design justice*. Cambridge, MA: The MIT Press.

Costanza-Chock, S., Raji, I., & Buolamwini, J. (2022). Who audits the auditors? *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 1571–1583. doi:10.1145/3531146.3533213

Crooks, R. (2022). Seeking liberation: Surveillance, datafication, and race. *Surveillance & Society*, *20*(4), 413–419. doi:10.24908/ss.v20i4.15983

Crooks, R., & Currie, M. (2021). Numbers will not save us. *The Information Society*, *37*(4), 201–213. doi:10.1080/01972243.2021.1920081

Davis, M. (2022). (G)local intersectionality. *Washington and Lee Law Review, 79*(3), 1021–1044. Retrieved from https://scholarlycommons.law.wlu.edu/wlulr/vol79/iss3/6

Donia, J. (2022). Normative logics of algorithmic accountability. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 598. doi:10.1145/3531146.3533123

Dourish, P. (2016). Algorithms and their others: Algorithmic culture in context. *Big Data & Society, 3*(2). doi:10.1177/2053951716665128

Edgeworth Analytics. (2021). *Independent analysis of the macarthur justice center study on ShotSpotter in Chicago*. Retrieved from https://edgeworthanalytics.com/independent-analysis-of-the-mjc-study-on-shotspotter-in-chicago/

Elish, M. C. (2019). Moral crumple zones. *Engaging Science, Technology, and Society, 5*(2019), 40–60. doi:10.17351/ests2019.260

Engler, A. (2021). Independent auditors are struggling to hold AI companies accountable. *Fast Company*. Retrieved from https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue

Equal Employment Opportunity Commission. (1979). Uniform guidelines on employee selection procedures. *Federal Register, 44*(43), 11996–12009.

Equivant. (2019). *Practitioner's guide to COMPAS core*. Equivant. Retrieved from https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf

Eubanks, V. (2018). *Automating inequality*. Boston, MA: St. Martin’s Publishing Group.

Eyert, F., & Lopez, P. (2023). Rethinking transparency as a communicative constellation. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 444–454. doi:10.1145/3593013.3594010

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., … Yeong, Z. (2021). Governing AI safety through independent audits. *Nature Machine Intelligence, 3*(7), 566–571. doi:10.1038/s42256-021-00370-7

Gelman, A., & Nolan, D. (2002). You can load a die, but you can’t bias a coin. *The American Statistician, 56*(4), 308–311. doi:10.1198/000313002605

Gender Shades. (n.d.). *Gender Shades FAQ*. MIT Media Lab. Retrieved from https://www.media.mit.edu/projects/gender-shades/faq/

Grant, K. (2020). Shotspotter sensors send SDPD officers to false alarms more often than advertised. *Voice of San Diego*. Retrieved from https://voiceofsandiego.org/2020/09/22/shotspotter-sensors-send-sdpd-officers-to-false-alarms-more-often-than-advertised/

Gray, M., & Suri, S. (2019). *Ghost work*. Boston, MA: Houghton Mifflin Harcourt.

Green, B. (2020). The false promise of risk assessments. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency,* 594–606. doi:10.1145/3351095.3372869

Guariglia, M. (2021). *It’s time for police to stop using shotspotter*. Electronic Frontier Foundation. Retrieved from https://www.eff.org/deeplinks/2021/07/its-time-police-stop-using-shotspotter

Hamid, S. (2020). Community defense: Sarah T. Hamid on abolishing carceral technologies. *Logic Magazine*, 11. Retrieved from https://logicmag.io/care/community-defense-sarah-t-hamid-on-abolishing-carceral-technologies/

Hanna, A., Denton, E., Smart, A., & Smith-Loud, J. (2020). Towards a critical race methodology in algorithmic fairness. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency,* 501–512. doi:10.1145/3351095.3372826

Irani, L. (2019). *Chasing innovation*. Princeton, NJ: Princeton University Press.

Irani, L., & Marx, J. (2021). *Redacted*. San Diego, CA: Taller California.

Irani, L., & Silberman, M. S. (2013). Turkopticon: Interrupting worker invisibility in Amazon Mechanical Turk. *Proceedings of the 2013 SIGCHI Conference on Human Factors in Computing Systems,* 611–620. doi:10.1145/2470654.2470742

Kang, E., & Hudson, S. (2022). Audible crime scenes: ShotSpotter as diagnostic, policing, and space-making infrastructure. *Science, Technology, and Human Values*. Advance online publication. doi:10.1177/01622439221143217

Keyes, O. (2018). The misgendering machines. *Proceedings of the ACM on Human-Computer Interaction, 2*(CSCW), 1–22. doi:10.1145/3274357

Keyes, O., & Austin, J. (2022). Feeling fixes. *Big Data & Society, 9*(2). doi:10.1177/20539517221113772

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). *Inherent trade-offs in the fair determination of risk scores*. arXiv preprints. doi:10.48550/arXiv.1609.05807

Lampland, M., & Star, S. L. (2009). *Standards and their stories*. Cornell, NY: Cornell University Press.

Latour, B. (2004). Why has critique run out of steam? *Critical Inquiry, 30*(2), 225–248. doi:10.1086/421123

MacArthur Justice Center. (2021). *End police surveillance*. Retrieved from https://endpolicesurveillance.com/research-findings/

Metaxa, D., Park, J. S., Robertson, R. E., Karahalios, K., Wilson, C., Hancock, J., & Sandvig, C. (2021). Auditing algorithms: Understanding algorithmic systems from the outside in. *Foundations and Trends in Human–Computer Interaction, 14*(4), 272–344. doi:10.1561/1100000083

Metcalf, J., Singh, R., Moss, E., Tafesse, E., & Watkins, E. (2023). Taking algorithms to courts. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 1450–1462. doi:10.1145/3593013.3594092

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–36. doi:10.1145/3359221

Narayanan, A. [Arvind Narayanan]. (2018). *21 fairness definitions and their politics* [Video file]. YouTube. Retrieved from https://www.youtube.com/watch?v=jIXIuYdnyyk

Pasquale, F. (2015). *The black box society*. Cambridge, MA: Harvard University Press.

Pasquale, F. (2020). *New laws of robotics*. Cambridge, MA: Harvard University Press.

Pymetrics. (2021). Audited & ethical AI. *Pymetrics.ai*. Retrieved from https://web.archive.org/web/20211017030746/https://www.pymetrics.ai/audited-ethical-ai

Radiya-Dixit, E., & Neff, G. (2023). A sociotechnical audit: Assessing police use of facial recognition. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 1334–1346. doi:10.1145/3593013.3594084

Raji, I., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society,* 145–151. doi:10.1145/3375627.3375820

Raji, I., Kumar, I., Horowitz, A., & Selbst, A. (2022). The fallacy of AI functionality. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 959–972. doi:10.1145/3531146.3533158

Raji, I., Smart, A., White, R., Mitchell, M., Gebru, T., Hutchinson, B., … Barnes, P. (2020). Closing the AI accountability gap. *Proceedings of the 2020 ACM Conference on Fairness, Accountability, and Transparency,* 33–44. doi:10.1145/3351095.3372873

Rakova, B., & Dobbe, R. (2023). Algorithms as social-ecological-technological systems. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency,* 491–509. doi:10.1145/3593013.3594014

Ryan, K. (2018). How Tesla and LinkedIn use neuroscience-based games (instead of resumes) to hire the best talent. *Inc*. Retrieved from https://www.inc.com/kevin-j-ryan/Pymetrics-replacing-resumes-with-brain-games.html

Saltman, J. (1975). Implementing open housing laws through social action. *The Journal of Applied Behavioral Science, 11*(1), 39–61.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). *Auditing algorithms*. Paper presented at Data and Discrimination, Seattle, WA. Retrieved from https://www-personal.umich.edu/~csandvig/research/Auditing%20Algorithms%20--%20Sandvig%20--%20ICA%202014%20Data%20and%20Discrimination%20Preconference.pdf

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2016). When the algorithm itself is a racist. *International Journal of Communication, 10*, 4972–4990.

Scannell, R. J. (2019). This is not minority report. In R. Benjamin (Ed.), *Captivating technology* (pp. 107–129). Durham, NC: Duke University Press.

Seaver, N. (2019). Knowing algorithms. In J. Vertesi & D. Ribes, *DigitalSTS* (pp. 412–422). Princeton, NJ: Princeton University Press.

Selby, N., Henderson, D., & Tayyabkhan, T. (2011). *ShotSpotter gunshot location system efficacy study*. CSG Analysis. Retrieved from https://www.shotspotter.com/wp-content/uploads/2020/12/ShotSpotter_Efficacy_Study_062311_FPV.pdf

Shapiro, A. (2019). Predictive policing for reform? *Surveillance & Society, 17*(3/4), 456–472. doi:10.24908/ss.v17i3/4.10410

Shen, H., DeVos, A., Eslami, M., & Holstein, K. (2021). Everyday algorithm auditing. *Proceedings of the ACM on Human-Computer Interaction, 5*(CSCW2), 1–29. doi:10.1145/3479577

Shilton, K., Moss, E., Gilbert, S. A., Bietz, M. J., Fiesler, C., Metcalf, J., … Zimmer, M. (2021). Excavating awareness and power in data science. *Big Data & Society, 8*(2). doi:10.1177/20539517211040759

ShotSpotter. (2015). *SST customer success training bulletin*. DocumentCloud. Retrieved from https://www.documentcloud.org/documents/3221020-ShotSpotter-nationwide-memo-July-2015

ShotSpotter. (2021). *ShotSpotter Respond™ Q&A*. ShotSpotter. Retrieved from https://www.shotspotter.com/wp-content/uploads/2021/07/ShotSpotter-Respond-FAQ-Jul-2021.pdf

Stanley, J. (2015). *Shotspotter CEO answers questions on gunshot detectors in cities*. ACLU News & Commentary. Retrieved from https://www.aclu.org/news/privacy-technology/shotspotter-ceo-answers-questions-gunshot

StopLAPDspying Coalition. (2020). *The algorithmic ecology*. Medium. Retrieved from https://stoplapdspying.medium.com/the-algorithmic-ecology-an-abolitionist-tool-for-organizing-against-algorithms-14fcbd0e64d0

Wilson, C. (2022). *Algorithm audits*. Christo Wilson. Retrieved from https://cbw.sh/audits.html

Wilson, C., Ghosh, A., Jiang, S., Mislove, A., Baker, L., Szary, J., … Polli, F. (2021). Building and auditing fair algorithms: A case study in candidate screening. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency,* 666–677. doi:10.1145/3442188.3445928

Winkley, L. (2016). *Gunshot detection system goes live in San Diego*. Government Technology. Retrieved from https://www.govtech.com/public-safety/gunshot-detection-system-goes-live-in-san-diego.html

Young, M., Katell, M., & Krafft, P. (2022). Confronting power and corporate capture at the FAccT conference. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency,* 1375–1386. doi:10.1145/3531146.3533194