

No More Plaudits for Algorithm Audits: A Provocation

Alicia DeVrio¹ & David Gray Widder²

¹Carnegie Mellon University

²University of Texas at Austin

Abstract

Algorithm audits have been increasingly proposed, pursued, and prioritized as a way to improve problematic algorithmic systems. And yet, the same issues continue to plague algorithmic systems, with algorithm audits making little headway toward remediation of algorithmic harm. In this provocation, we argue that algorithm audits are fundamentally insufficient mechanisms by which to ameliorate the violent impacts of algorithmic systems because they give the appearance of addressing algorithmic harms but do not meaningfully challenge the profitable systems they claim to scrutinize. To show this, we examine how algorithm audits sit within structures of power, bringing attention to three interrelated pitfalls of algorithm auditing: that algorithm audits too often *further* the concentration of power, that algorithm audits *limit* how both problems and solutions are envisioned, and that algorithm audits *preserve* a facade of positive change. We conclude with discussion of some necessary commitments for future work to start to grapple with and avoid these pitfalls and of how the insufficiencies of algorithm audits highlight broken dynamics in the broader AI accountability ecosystem. As such, we contend that fixing problematic algorithmic behaviors requires deprioritizing algorithm audits and instead shifting scrutiny to the systems of power in which audits and other accountability practices sit, which in turn mediate whether these practices and the theory of change they espouse are effective.

Introduction

Algorithmic systems have repeatedly displayed problematic behavior in ways that create, further, and amplify wrongful and structural harms as a result (e.g., (Shelby et al. 2023; Metcalf et al. 2021; Diberardino, Baleshta, and Stark 2024)). Recent egregious cases include wrongful arrests following police use of facial recognition technology (MacMillan, Ovalle, and Schaffer 2025), algorithmic amplification of Nazi hate speech on TikTok (Doctor, Fiennes, and O’Connor 2024), and algorithmic welfare payment discrimination against disabled people and single mothers (Meaker 2024), among many others. In response to harms like these, algorithm audits have been proposed and pursued in attempts to remediate algorithmic system issues. Significant academic, industry, and government attention has focused on and prioritized algorithm auditing, leading the method to

rise to prominence as a way to surface and counteract harmful algorithmic systems. For example, Safiya Umoja Noble’s algorithm audit highlighted search algorithm discrimination, specifically that Google searches for “black girls” yielded overly sexualized results (Noble 2018, 2013; Bandy 2021).

But are algorithm audits meaningfully fixing anything? Google has repeatedly claimed to address the issues in its search algorithm, but issues continue to plague the system. Noble herself describes how the initial searches that attuned her to issues within Google’s search results “launched a years-long research process, following offenses and ‘fixes’” (scare quotes original) (Noble 2018). In 2011, Noble identified the problematic search results for “black girls” in Google’s search algorithm (Noble 2012, 2018). While Google fixed this specific issue, in 2020 searching “Black girls” in Google’s advertising platform resulted in algorithmically suggested adult content ads (Asher-Schapiro 2020; Yin and Sankin 2020). And this came after Latanya Sweeney’s 2012 audit showing how Google’s advertising algorithm associated Black-sounding names with arrest record ads (Sweeney 2013)—an issue Google had also already applied fixes to (Goodman 2015).

Cycles like this, from audit to issue to fix and back again, are not difficult to find. Google’s translation service has gone in similar circles related to algorithmic issues of sexism (e.g., (Prates, Avelar, and Lamb 2019; Kuczmarski 2018; Bond 2020)). Auditing found racism in algorithmic predictive policing software PredPol in 2016 (Smith IV 2016), but PredPol in 2019 was still used by many US cities (Haskins 2019)—and then in 2021 another audit found the same issues in the software (Sankin et al. 2021). In practice algorithm audits often surface and provide evidence for iterations of the same issues, issues that are already known, over and over.

In line with recent AI ethics scholarship interrogating the theories of change that guide the community’s work (Gansky and McDonald 2022; Wilkinson et al. 2023; Olteanu et al. 2023), we are concerned about the theory of change upon which algorithm audits rely. The auditing landscape is going around in circles, repeatedly reinventing the wheel, inuring society to issues in algorithmic systems, both when they are brought up and when they are not sufficiently addressed. When society becomes accustomed to this circle—when the issues that audits find become normal and thus less shock-

ing—audits lose their ability to spark change, as they usually lack direct means of intervention (Friedler et al. 2023). This, in turn, degrades the utility of algorithm audits as a tool to use to accomplish real, positive change.

Even when those in control have taken action to address algorithmic system issues raised by audits, the actions taken are often middling at best. Fixes are enacted in which little about the algorithmic system as a whole is changed, such as in the examples involving Google above, or those in power disregard the results of audits by arguing the audit’s methodology is incorrect. And this latter response has come even in the case of very prominent audits. For example, after Virginia Eubanks detailed issues with the Allegheny Family Screening Tool (AFST) in the book *Automating Inequality* (Eubanks 2018b,a), the Allegheny County Department of Human Services released a statement calling the audit inaccurate (Fitzgerald 2018; Eubanks 2018c). Similarly, after ProPublica surfaced issues in Northpointe’s COMPAS algorithm used to predict recidivism (Angwin et al. 2016), Northpointe released a report that claims the audit’s methods have errors (Dieterich, Mendoza, and Brennan 2016). Round and around the issues go.

At their core, algorithm audits are supposed to address issues in algorithmic systems—but what constitutes an algorithmic issue depends on perspective (Widder and Goues 2024). For those who interact with algorithmic systems or are otherwise impacted by algorithmic systems, the issues are the harmful, violent enactments of systems that were not designed or built for them to thrive in (Alkhatib 2021)—that is, for impacted people, the issues are the negative impacts they are forced to experience: structural and wrongful harms. However, for those in control of algorithmic systems, such as tech companies, the issues are the complaints they receive about harmful algorithmic behaviors—complaints that bring the perception that their algorithmic systems perform poorly, the suggestion that they do not care about adverse impacts from their systems, and broadly bad PR. That is, for those in control of algorithmic systems, the issues are not the problematic algorithmic behaviors themselves but the *barriers those problematic behaviors create to their business incentives under capitalism*.

These differing perspectives of what constitutes an algorithmic system issue affect understandings of what algorithm audits are supposed to do, leading to a fundamental tension within algorithm audits. On their face, algorithm audits are designed and used to help surface, provide evidence for, and remediate algorithmic system issues. However, algorithm audits have been significantly prioritized by those already in control of algorithmic systems, and returning to their perspective of algorithmic system issues above, algorithm audits can then be understood as ways to reduce complaints about algorithmic systems and limit the resource investment they must spend on algorithmic system issues, all aligned with profit incentives. That is, algorithm audits serve to further the capitalist pursuits of those in control of algorithmic systems by convincing those affected by algorithmic systems that enough is being done to address the harms they encounter. Thus, audits are used to give the appearance of addressing harms, but do so in a way that does not mean-

ingfully challenge the profitable systems they scrutinize. Or: algorithm audits support those with power in claiming to prioritize care while actually enacting prioritization of profit at the expense of those who use, interact with, and are affected by algorithmic systems.

Because of this, we argue that algorithm audits are insufficient, ineffective, and even harmful mechanisms to address the structural and wrongful harms of algorithmic systems. In this provocation we examine where algorithm audits sit within systems of power, bringing attention to three central pitfalls of algorithm audits manifested by this double bind: that algorithm audits further the concentration of power, that algorithm audits limit how both problems and solutions are envisioned, and that algorithm audits preserve a facade of positive change. We then outline some necessary commitments for future work as starting points for the AI ethics community to begin to confront and avoid these pitfalls of algorithm audits. These include suggestions that AI ethics community members orient themselves and their work toward tackling systems of harm, expanding notions of what is considered valuable evidence, acting directly and helping others act directly, and working to shift power to those affected by algorithmic systems. We conclude by discussing how the insufficiencies of algorithm audits highlight how the broader AI accountability ecosystem is broken, and call for the AI ethics community to thus deprioritize algorithm audits and instead shift scrutiny to the systems of power in which audits sit.

Our primary contributions are as follows:

- Three pitfalls developed from synthesizing and building on existing critiques that elevate these critiques into an urgent provocation that centers how broader power dynamics inhibit algorithm audits—and indeed the larger AI accountability ecosystem—from achieving lasting positive impact.
- A critique of and push to develop a better theory of change that drives how the AI ethics community handles problematic algorithmic systems impacts, along with some starting directions for how we as a community can begin to rethink the existing AI accountability ecosystem such that its structuring better challenges power and holds those in control of algorithmic systems accountable for their systems’ impacts.

What Are Algorithm Audits?

In this work, we discuss algorithm audits and the theory of change they espouse. We align our work with Gansky and McDonald’s argument that FAccT theories of change are ineffective, especially that they cannot be effective due to the dominance of profit and corporate power (Gansky and McDonald 2022). Following Birhane et al.’s critiques of algorithm audits as suffering from power asymmetries as well as limitations in scope, methodology, and practice (Birhane et al. 2024) and following calls to focus more on power dynamics in AI ethics work (Birhane et al. 2022b; Birhane 2021), our work questions whether addressing these issues is currently or ever possible under the banner of algorithm auditing due to audits’ positioning within larger structures

of power. In this section we explain our understanding of algorithm audits used in this piece.

Significant work has investigated and studied algorithm audits, which have been defined in a wide variety of ways (Ojewale et al. 2025; Vecchione, Levy, and Barocas 2021). To name a few prominent examples: Sandvig et al. propose algorithm auditing as “a starting point for how researchers might look inside the black box of the algorithm to pursue knowledge about pressing public problems” (Sandvig et al. 2014); Raji and Buolamwini describe algorithm auditing as involving “the collection and analysis of outcomes from a fixed algorithm” in order to “uncover problematic patterns in models of interest” (Raji and Buolamwini 2019); Jack Bandy understands algorithm auditing as “an empirical study investigating a public algorithmic system for potential problematic behavior” (Bandy 2021); Metaxa et al. describe algorithm auditing as “a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences about its opaque inner workings” (Metaxa et al. 2021); Birhane et al. and Ojewale et al. call for AI accountability to be a significant goal of algorithm audits (Birhane et al. 2024; Ojewale et al. 2025).

Guided by these dominant definitions in academic literature, we understand algorithm audits as systematic investigations of algorithmic systems to surface problematic issues, with the aim of achieving some form of meaningful accountability.

This begs the question: what is an algorithmic system? Algorithms and AI lack clear, universal definitions¹ (Lum and Chowdhury 2021; Suchman 2023); as such, we understand algorithmic systems based on their political implications. Ali Alkhatib has aimed to capture the broader project of AI by defining AI as “an ideological project to shift authority and autonomy away from individuals, towards centralized structures of power” (Alkhatib 2024a). We adopt this framing, gesturing toward the interplay between imaginaries and realities of algorithmic systems that can make powerful that which is seen as powerful (Cave, Dihal, and Dillon 2020).

Returning, then, to algorithm audits with this understanding of algorithmic systems in hand clarifies the purpose of algorithm auditing. Broadly speaking, interventions against problematic algorithmic behavior that aim to be effective should ultimately counteract reductions of authority and autonomy, which we understand as problematic issues that, as per the above definitions, are currents embedded in the ideology of AI. As algorithm audits aim to surface and achieve meaningful accountability for algorithmic issues, algorithm audits as a project should ultimately halt or lessen the congealing of power inherent in algorithmic systems and shift power away from domineering institutions toward systemically disempowered individuals. In the following section we examine how algorithm audits are ineffective at accomplishing this.

¹Based on this, in addition to the frequent conflation of the terms and what they represent, we use “algorithms” and “AI” interchangeably throughout this work.

Pitfalls of Algorithm Audits

In this section, we highlight three interrelated pitfalls of algorithm audits that surface upon examining how audits sit within larger structures of power. In doing this, we trouble conceptualizations of audits’ abilities to address problematic, harmful issues in algorithmic systems. Our broad scope of inquiry examines algorithm audits within larger dynamics of power. The power-related pitfalls we draw out apply to external audits—arguably the only truly transparent and trustworthy forms of algorithm audits (Costanza-Chock, Raji, and Buolamwini 2022; Friedler et al. 2023)—as well as, in similar if not worse ways, internal audits. We note that distinctions between categories of internal and external algorithm audits can be complex and uncertain (Terzis, Veale, and Gaumann 2024). Despite work finding internal audits insufficient and calling for more external audits (Longpre et al. 2025), in our examination external audits do not escape the core pitfalls we identify because they too are situated within and fail to significantly change larger dynamics of power in the AI accountability ecosystem (e.g., (Gerchick et al. 2025; Wright et al. 2024; Groves et al. 2024)).

This section provides insight into the following three pitfalls of algorithm audits: First, we describe how algorithm audits must be reckoned with as mechanisms that support the powerful in steering the narratives around negative impacts from algorithmic systems, in externalizing the labor required to address those impacts, and in retaining control throughout. Next, we highlight how algorithm audits frame problems too narrowly, constrain and exclude forms of knowledge, and are limited in their visions of solutions; that is, algorithm audits lack the expansiveness, on multiple fronts, required to sufficiently address issues arising from, perpetuated by, and intensified with algorithmic systems. Finally, we discuss how algorithm audits performatively reify and maintain the appearance of doing good, a facade that distracts from the insufficient capacity that audits have to actually drive change, which in turn leads to frequent reiteration of roughly known issues.

Algorithm Audits Further the Concentration of Power

Algorithm audits serve to further solidify existing power structures, concentrating additional power in the hands of those who already have significant control over algorithmic systems (Widder, Whittaker, and West 2024), even as they purport to shift power to those with less. Using an algorithm audit is itself a choice that points in some directions and forecloses others, regardless of whether an algorithm audit can be characterized as more or less internal or external, and regardless of whether that choice is reducible to a single person, group, or institution. While internal audits inherently suffer from auditors coming from the bed of power, external audits are often beholden in other ways to power asymmetries and the interests of those in control of the systems they target (Birhane et al. 2024), such as through nondisclosure agreements that restrict how much they can share about their findings (Costanza-Chock, Raji, and Buolamwini 2022; Ojewale et al. 2025). The prioritization of

algorithm audits—any kinds of algorithm audits—thus furthers the concentration of power by instantiating a commitment to ways of knowing and ways of taking action that have been deemed acceptable and desirable within the larger currents of systemic power, often via people and institutions that hold considerable sway over those currents. We draw from DeVrio et al.’s framing of “algorithmlords” as “those holding greater power to envision and construct algorithmic systems, occupying positions ‘above’ those subjugated by their decisions” (DeVrio, Eslami, and Holstein 2024). For example, designers and developers could be algorithmlords while creating an algorithmic system, as could company management or city officials while deploying an algorithmic system, as could legislative members or other policy experts while requiring certain behaviors from an algorithmic system. We understand the designation of algorithmlord to apply in a way that varies depending on context and does not necessarily signal individuals’ endorsement of the systems they are part of, as even people in relatively powerful positions can lack the ability or knowledge to change these systems and their incentives (Gururaja et al. 2023; Widder and Nafus 2023).

Algorithm audits orient additional power toward algorithmlords by allowing them to decide what is a problem in an algorithmic system. STS scholar Michel Callon has explored and described the power wielded in how an issue is “problematized” (Callon 1984), and likewise, those who identify problems in algorithmic systems using audits have significant power to define the shape of the problem. Instead of allowing people affected by algorithmic harms to raise issues in ways that they choose, emphasizing parts of problematic issues they find most salient or urgent via their own means, algorithm audits place narrative control in the hands of algorithmlords: problem formulation occurs in specific ways due to auditing. To that end, issues are regarded as unknown or alleged until an algorithm audit documents them in a formal way, thereby relegating to the sidelines other efforts to identify and remedy issues that may occur without an audit.

Additionally, defining problems via algorithm audits empowers the definers to determine the solutions to those problems, as the shape of a problem lends itself more to certain directions forward. Narrative control of both algorithmic problems and their solutions by those already in power is further facilitated by power imbalances between relevant parties as well as by a society that has largely bought into engineering logic and systematic reasoning as more legitimate than other ways of knowing. Zong and Matias have pointed out in the realm of data collection that “it can appear to be a good thing when data collectors offer well-defined processes for people to opt out or voice dissent. However, data collectors often use these pathways of dissent to limit refusers’ agency by directing them away from other unspoken possibilities” (Zong and Matias 2024). Similarly, algorithm audits can appear to offer clear ways to raise harmful issues to the attention of algorithmlords, but they permit little agency for those affected and instead grant algorithmlords the power to direct when and how algorithmic systems are investigated, as well as in what ways issues are attended to.

More participatory forms of algorithm auditing can return some agency to affected people, but they contribute to the concentration of power in their own ways. While many algorithm audits do not even pretend to share or shift control of the process to those affected by algorithmic harm, forms of algorithm auditing have been proposed and pursued that include and empower communities and affected people in participation and some control of auditing processes (e.g., (Shen et al. 2021; DeVos et al. 2022; Deng et al. 2023; Katell et al. 2020; Krafft et al. 2021)). Algorithm audit processes that not only involve but are driven by on-the-ground communities have the potential to address some of the pitfalls we describe here, as evidenced by much work on public co-design of technologies and related technospaces (e.g., (Hope et al. 2019; Suresh et al. 2022)). However, algorithm audits that grant affected people and communities significant control to direct the process are far from the rule, and mere inclusion or participation in processes like algorithm audits should not be straightforwardly equated with shifts of power to those people (Arnstein 1969; Birhane et al. 2022a; Sloane et al. 2022; Costanza-Chock, Raji, and Buolamwini 2022). An especially pernicious aspect of participation like this in algorithm auditing, as described by the organizers of *Queer in AI*, is that it may “deny space for alternatives, thereby serving as a justification of the systems in question” (of *Queer in AI* et al. 2023).

Algorithm auditing can also prop up existing structures of power by supporting algorithmlords in externalizing the work required to vet and improve their algorithmic systems. Participatory forms of auditing can exploit the labor of people in a system reliant upon trickle-down positive effects that neglects to ensure individuals are actually better off themselves as a result of their participation (Dalal, Hall, and Johnson 2024; Birhane et al. 2022a). Regardless of the level of involvement of affected people in an algorithm audit, the process often allows and supports relocating the burden of ensuring that algorithmic systems are safe, fit for purpose, ethical, fair, or otherwise good away from companies and others in control of these systems who should be responsible for their impacts and toward external groups who do the work of auditing.

Algorithm Audits Limit How Both Problems & Solutions Are Envisioned

That algorithm audits allow those in power to control the framing of both issues within algorithmic systems and potential ways to address those issues, as argued in the previous section, implies that algorithm audits severely limit the framing of both problems and solutions. While above we focus on how this consolidates power, in this section we discuss in what ways this shapes the problems and solutions themselves. Algorithm audits are mechanistically limited in their remit (Birhane et al. 2024), a fact of design that confines their capacity to adequately identify and handle algorithmic system issues. That is, due to their formulation, algorithm audits are inherently small scope, individualizing, and reductionist, obscuring systems-level critique by restricting what is understood as acceptable evidence and focusing primarily on tractable problems and solutions.

The mechanistic limitations of algorithm audits delegitimize types of issues that fall beyond the bounds of their inclusion as well as subjugate many potential futures—and their visionaries—through terms that make alternative possibilities incomprehensible. These pitfalls of overly repressed, deficient imagination for framing both problems in and solutions for algorithmic systems occurs across the spectrum of internal to external algorithm audits.

Algorithm audits constrain evidence by only making use of certain types of evidence. Algorithm audit evidence typically must allow for quantitative ways of viewing a potential problem, as many audits seek to statistically prove whether issues exist or are significant enough to warrant pursuing further. David Widder details how casting AI ethics work as objective, rational, and quantitative has legitimized it but has come at the cost of dismissing other kinds of knowledge that are challenging to fit within neat engineering-like processes such as audits (Widder 2024). Fundamentally this restricts what sorts of evidence is and is not acceptable for an algorithm audit. Qualitative forms of evidence such as personal anecdotes about experiences with an algorithmic system must be able to be categorized, measured, and compared. Similar to Sara Ahmed’s work describing how certain forms of complaints are legitimized by the infrastructures that exist for complaining (Ahmed 2021), algorithm audits’ epistemic constraints on evidence further notions of what sorts of evidence is not only acceptable for audits but legitimate writ large.

In addition to constraining and problematically reshaping evidence, algorithm audits exclude other forms of evidence. For example, algorithm audits follow standardized forms of reasoning and logic, non-emotional forms including but not limited to seeking statistical significance as a vehicle for establishing proof. As such, they sideline emotional, embodied, and experiential forms of evidence that can also significantly highlight to what extent and in what ways issues can manifest in and around algorithmic systems (Keyes and Austin 2022). We relate this to the concept of respectability politics in which the presentation of issues and desires for change can be more successful when they adhere to notions of what is socially acceptable (e.g., (Harris 2003; Higginbotham 1993; Pitcan, Marwick, and boyd 2018)); similarly, algorithm audits advance specific notions of what is acceptable—that which is or can be systematically proven—and are designed to promote the surfacing and addressing of issues that work with and within this.

In part due to the limited nature of evidence they deem acceptable, algorithm audits frame algorithmic system problems narrowly. Methodologically algorithm audits focus on specific examples of algorithmic system failures that are straightforward to isolate, prove, and fix. This means that the audits are set up to show problems that have more tractable solutions, such as racial bias or disparate impact in an algorithmic system which can be addressed by balancing the discrepancy. In turn, this may contribute to why industry and others in charge of algorithmic systems have shown so much interest in audits, investing significant time and money into them in recent years. Indeed, audits provide natural solutions—but only ones constrained to the scope of the in-

quiry. By focusing on specific point examples of harm, audits are not a natural tool for documenting systemic problems emanating from algorithmic systems, of which there are many (Alkhatib 2021). As Birhane et al. put it, reductions in problem and subsequently solution scopes can “risk equating processes and outcomes that are *fair* with those that are *equitable* and *just*” (Birhane et al. 2022b).

Overly narrow problem identification limits the reform made possible by audits to the logics of an already existing system. Algorithm audits are framed under the assumption that normally algorithmic systems are correct, innocent until proven guilty, and that issues revealed by audits are aberrations which can be individually removed. This makes the issues highlighted by people affected by algorithmic systems guilty until proven innocent: that is, the issues rather than the system bear the burden of proof and legitimacy. This move naturalizes algorithmic systems, forcing potential solutions to algorithmic issues to work within the assumption that the algorithmic system at hand will always exist, to take as predetermined the existence of the algorithmic system and its overarching logic. As such, algorithm audits cannot be broadly imaginative in their visions of alternative futures, of better worlds (see, e.g., (Hamilton 2021; Benjamin 2024)).

Algorithm Audits Preserve a Facade of Positive Change

Finally, algorithm audits performatively purport to address problematic algorithmic behaviors but in reality frequently fail to accomplish this. This makes algorithm audits not only deceptive but even actively harmful themselves, as the prioritization of audits consumes valuable resources that could be used otherwise for positive change. Contributing to the appearance of positive change occurs regardless of whether algorithm audits are internal or external. While external algorithm audits avoid conflicts of interest and internal algorithm audits have potential to surface issues before deployment, both can provide reputation management for those in control of an algorithmic system (Raji et al. 2020b). That is, algorithmlords can gesture at any type of algorithm audit to signal that they are working to address issues from their systems, even when those audits fail to achieve lasting positive outcomes.

It follows then that algorithm auditing can be virtue signalling, when co-opted by powerful institutions for their own advantage. This can be considered a form of “ethics washing” whereby institutions like tech companies attempt to keep up the appearance of being ethical without actually substantiating with depth of action (Bietti 2020). This veneer of self-critique furthers the perception that positive change is occurring regarding harmful algorithmic systems while in reality enabling stasis (Goodman and Trehu 2022). Similarly, Lee Vinsel has used the term “criti-hype” to describe how some criticism of technologies can simultaneously generate interest in that technology and, more insidiously, distract from ongoing harms actively affecting people (Vinsel 2021). As Ahmed wrote about a complaint submitted to academic institutions: “In effect, the report was a means by which the university was able to not deal with the situation by appearing to have dealt with it” (Ahmed 2021).

Beyond or perhaps because of this, algorithm audits at this point tend to reiterate issues generally known to exist. Racial and gender bias, for example, seem to repeatedly be “discovered” through audits of algorithmic systems. Seeking to prove harms rarely leads to adequate addressing of practices that cause those harms in practice (Barabas 2022; Birhane et al. 2022b; Gonen and Goldberg 2019), which can contribute to attempts to prove the same issues over and over again. Algorithm audits themselves often have little structural power to enforce changes based on their results; as such, their prioritization can result in focusing time, effort, and other valuable resource on surfacing or providing more evidence for known issues rather than working toward ways to address those issues, wider issues, and less commonly thought of issues.

In addition to resource seepage, algorithm audits can mask harms that come with the method. For example, some algorithm audits can lead to incremental, often quantitative improvements, such as the decrease in bias following the famous Gender Shades audit (Raji and Buolamwini 2019); however, some biases still persisted years later, suggesting that larger changes than those offered by algorithm audits are required for justice (Raji and Buolamwini 2022). Additionally, the algorithmic system being audited in that case involved gender classification, an oppressive technology itself (e.g., (Keyes 2018; Scheuerman, Paul, and Brubaker 2019; Hamidi, Scheuerman, and Branham 2018; Raji et al. 2020a)). This highlights how, beneath the surface of algorithm audit fixes, there often lie more fundamental human rights issues with the algorithmic system that algorithm audits are unfit to handle (e.g., (Keyes, Hutson, and Durbin 2019; Hassein 2017)). To that end, Keyes and Austin have described how auditing can produce new harms even while aiming to address existing harms (Keyes and Austin 2022). In garnering evidence of harm, algorithm audits often ask survivors of trauma from algorithmic systems to relive their experiences and translate them to be legible for others who may be more invested in the audit than in advocating for those affected or otherwise working toward justice or harm reduction.

Ways Forward: Grappling with Power Toward a More Effective Theory of Change

In this provocation, we have argued that algorithm audits are insufficient, ineffective, even detrimental ways to address the structural and wrongful harms of algorithmic systems. To do this, above we have examined some significant, power-related pitfalls of algorithm audits. In this section, we provide directions forward for the AI ethics community to reconsider the algorithm auditing theory of change and reorient future work in light of this, all via attending to and grappling with power dynamics. We start by providing a set of commitments the AI ethics community must take on in order to begin to avoid the pitfalls we identify above. Then, we discuss how the pitfalls of algorithm audits highlight broken dynamics in the larger algorithmic accountability ecosystem, and consequently we encourage restructuring this ecosystem such that more radical futures are prioritized

and those with power over algorithmic systems are held accountable for their systems’ impacts.

Embracing New Commitments

In this section, we suggest necessary commitments for the AI ethics community to take on such that they can begin to grapple with and avoid the above pitfalls. These commitments are not intended to be comprehensive fixes and do not perfectly correspond to the pitfalls described above; rather, we see these as fundamental orientations that, when truly pledged and pursued, might help in future avoidance of algorithm audit pitfalls. We seek to encourage pragmatic directions that on the one hand do not accept sociotechnical systems as they are, as unable to be changed or controlled (Postman 1998), while on the other do not ignore that existing structural currents shape the potential impacts of any actions we undertake.

First, we entreat the AI ethics community to commit to *tackling systems of harm* rather than point instances of problematic algorithmic behavior. Algorithm audits tend to focus on individual issues within algorithmic systems; as such, we see opportunities for expanding focus to encompass entire algorithmic systems, including the factors that contribute to their design, development, and deployment in the world as well as their internal working. This has been understood as a commitment to “non-reformist reforms” which, in contrast to reformist reforms that look for ways to improve existing systems, look for “ways to achieve emancipatory social conditions” by starting beyond rather than within existing systems (Green 2021; Gorz 1987; Huber 2025). The incremental changes that come from reformist reforms might be seen as positive on their face but in reality can reinforce oppression long-term by supporting the status quo (Green 2021). As such, we echo other scholars in calling for algorithmic issues to be engaged with as sociotechnical problems that can only have sociotechnical solutions (e.g., (Shelby et al. 2023; Selbst et al. 2019; Forsythe 2001)). This acknowledges how algorithmic systems can be imbued with harms as a feature (Hamilton 2021), requiring that their entire logics and existence be brought under question for adequate mitigation of their issues.

Second, we call for the AI ethics community to commit to *expanding notions of what is considered valuable evidence* beyond the systematic, quantitative nature of evidence that is legitimized by algorithm audits. Scholars have emphasized the importance of translation work (Muller 2004; Sum et al. 2023); we similarly see opportunities for researchers to convey evidence of algorithmic system issues from affected people to those with the power to address those issues in ways that are comprehensible and understood as significant to those in power. We understand that translation work can be conducted via algorithm auditing where necessary. In many cases, however, we see opportunities for the AI ethics community to work to convey the legitimacy and importance of varied forms of evidence to those in power regardless of what the existing norms of legitimacy and importance might dictate. We especially see the need for advocacy for human aspects of algorithmic impacts—often excluded by algorithm audits—as evidence. For example, emotional evi-

dence conveying how algorithmic issues made people feel or acts of anger in response to problematic algorithmic behavior can be understood as legitimate and productive feedback about in what ways and to what extent algorithmic systems have issues that can be addressed. Including and elevating these sorts of evidence can expand legitimacy to more of the ways that people engage with and are affected by algorithmic systems as well as support the redirection of resources toward addressing issues rather than toward seeking to find additional proof.

To that end, we encourage AI ethics community members to commit to *acting directly and helping others act directly* rather than working to convince others to act to improve conditions created and exacerbated by harmful algorithmic systems. The anarchist principle of direct action provides a way to make change without waiting for the approval or buy-in of powerful institutions or people and has been proposed as useful for the realm of AI ethics (Agnew 2024). Affected people already frequently respond to algorithmic harm they encounter using what power they have in a variety of ways (DeVrio, Eslami, and Holstein 2024; Kulynych et al. 2020; Whitney et al. 2021). For example, destruction and sabotage of algorithmic systems has been observed and even promoted as a way to more comprehensively address harmful algorithmic systems (Alkhatib 2024b; DeVrio, Eslami, and Holstein 2024; Kulynych et al. 2020; Vincent et al. 2021). Relatedly, we highlight that anger can be productive and a form of power (Cherry 2020; Ahmed 2014; Lorde 1981; Bailey 2018). Though algorithm audits relegate emotions to the sidelines of what justifies evidence—as discussed above—as well as what is useful for building power, we see anger as having the potential to bring groups together for collective action, to motivate people to have their voices heard, and to encourage those in charge of algorithmic systems to listen and make tangible change in response. For example, anger can spur significant negative PR, which can be a strong incentive for tech companies to address issues (e.g., (Shen et al. 2021; DeVrio, Eslami, and Holstein 2024)). Relational ethics also offers pathways toward more just outcomes in and around algorithmic systems (Birhane 2021). For the AI ethics community, this can mean prioritizing work and research that supports people affected by algorithmic systems in what they are already feeling and doing to resist harmful impacts of those systems (Huber et al. 2025). This could involve creating networks to connect people for collective action and solidarity, designing pathways for raising issues in more visible ways, supporting the development of more localized sociotechnical systems that do not rely on big tech, or building tools to allow people to more fully opt out of harmful systems or avoid parts of harmful systems they do not want to engage with.

In this vein, we call on the AI ethics community to commit to *shifting power to those affected by algorithmic systems* in ways that cede more control than many forms of participation do (Harrington, Erete, and Piper 2019; Ahmed 2020; Kalluri 2020). Supporting affected people in taking action directly, as above, is a productive way to do this. Auditing has origins as a community-driven method for accomplishing real change (Vecchione, Levy, and Barocas 2021). More

priority from the AI ethics community can and should be given to what people affected by algorithmic systems are themselves prioritizing (Suresh et al. 2022). The AI ethics community can reprioritize, shifting focus from finding and proving issues via audits to understanding and supporting what people are already doing to resist and shape algorithmic systems to create better on-the-ground outcomes (Sum and DeVrio 2024). This means supporting community ownership and control of algorithm audits instead of bringing in community members as merely included consultants, participants, or people who labor to surface issues. Communities need to have the power to redirect and reprioritize algorithm audits, which necessarily requires AI ethics community researchers and others to cede some of their own power to make those decisions in return. This also means allowing community members to define the frame of what should be done with regards to algorithmic systems, such that the possibility space can include the refusal to engage with, the decision to destroy, or the creation of joyful alternatives to harmful algorithmic systems as well as survival in and mitigation of harmful algorithmic systems. We encourage the AI ethics community to look toward the many community collectives that already exist and to offer themselves as partners, collaborators, and assistants to accomplish their aims.

Scrutinizing the Accountability Ecosystem

The existing structures of accountability surrounding algorithmic systems, especially those related to the regulation of tech companies, often necessitates the forms of language and evidence provided by algorithm audits in order to convince those in power to begin to make changes. Though we argue that algorithm audits have pitfalls that make them ineffective at addressing problematic algorithmic behavior, we would be remiss not to acknowledge that they can be valuable tools in some ways. Audre Lord famously wrote, “For the master’s tools will never dismantle the master’s house. They may allow us to temporarily beat him at his own game, but they will never enable us to bring about genuine change” (Lorde 1984). Following David Widder’s work on hierarchies of epistemic power in AI ethics labor (Widder 2024), we identify the master’s tools as not the audits themselves but as the ways in which audits instantiate hegemonic perspectives about worth: what is worth listening to, what is worth taking action on, who is worth supporting. When audits are leveraged to counter these perspectives, such as through using audit reports to pile on evidence to showcase to those in charge of algorithmic systems, they can eventually lead to incremental benefits.

But there is a sense of stasis in finding and proving very similar issues repeatedly, a sense of helplessness in seeing this as the primary vector for changing an issue-strewn technological landscape. To free us from what can feel like and in many regards is a Sisyphean approach, we look toward ways to expand the notion of what algorithm audits are. Existing research has sought to critique algorithm audits and broaden the scope of what we can consider to be an audit beyond more expert- and/or institutionally-led investigations that focus on systematicness and scientific proof above all (e.g., (Shen et al. 2021; Krafft et al. 2021; Calacci

and Pentland 2022; Geiger et al. 2024)). One piece at a time, we can tactically replace the parts of algorithm audits while remaining under the banner of algorithm auditing, trading sanctioned representations of harm with more expansive ones that better match affected people’s experiences, until we have created an audit of Theseus that shifts the onus of scrutiny from affected people providing evidence for the harms they experience onto those with control over the algorithmic systems that instigate those harms, as it should be.

Ultimately, we view algorithm audits as “tactics of soft resistance” (Wong 2021; Nafus and Sherman 2014), holding some possibilities but reliant upon the discourses and logics of the systems they try to make change from within. But what power might we gain from freeing our imaginations from dominant ways of thinking (Benjamin 2024), beyond “soft” audits? We advocate for the abolition of auditing—or the expansion and repositioning of the notion of auditing within the larger ecosystem of accountability, such that audits become “hard”—in favor of more radical tactics that shift power more wholesale to those who are affected by algorithmic systems’ negative impacts, allowing them to set the terms of what is required for their own survival in the landscape of algorithmic violence and rejecting the notions of legitimacy that current structures of power insist upon. Building on calls to strengthen AI accountability infrastructure (e.g., (Raji et al. 2022; Ojewale et al. 2025; Costanza-Chock, Raji, and Buolamwini 2022; Raji et al. 2020b; Attard-Frost 2025)), we contend that we should direct much of our attention as a community of research and practice toward finding ways to reshape this ecosystem’s power dynamics and finding, reckoning with, and moving in directions that prioritize futures with more imaginative possibility that are unconstrained by or actively refuse the logics and requirements of algorithmlords today. This will be a significant undertaking that requires collective input and effort, interrogation and reworking of stagnant theories of change, and embracing and taking action with the new commitments above in hand to better challenge power.

Locating responsibility with the powerful. The starting commitments we point toward and indeed our argument critiquing the effectiveness of algorithm audits largely focus on the AI ethics community and those affected by algorithmic systems. We have called upon the AI ethics community to reorient its stance toward algorithm audits due to fundamental flaws in the theory of change they pursue. In doing so, we advocate for increased engagement with and power ceded to communities affected by algorithmic systems. These directions are given assuming that those in control of algorithmic systems cannot be relied upon to adequately address the harmful impacts of their own systems.

However, by focusing on affected people and their determinations of what are reasonable ways to remedy algorithmic issues, we have implicitly but unintentionally put some burden of responsibility on them to determine those pathways forward. For instance, expanding the scope of auditing likely still requires affected people to identify issues they encounter, which can burden and even involve harm to those people in the process (Upadhyay et al. 2026). A tension exists here, as affected people attempt to drive change because

the status quo is that their experiences and lives are worsened by harmful algorithmic systems. Only when people affected by algorithmic systems also have power and control over those systems themselves can this tension start to be resolved—that is, when they are in positions such that they engage not only because their survival necessitates it but also to create technospaces of radical joy and thriving for themselves.

To reach this future, we emphasize that responsibility for algorithmic system impacts, and especially for algorithmic system issues, must remain with those with power over algorithmic systems: the organizations and institutions with control of the design, development, and deployment of these systems. We echo Lucy Suchman’s framing of “located accountabilities”, or how those involved in the production of technology should identify, understand, and take responsibility for their positions within larger structures of power (Suchman 2002). At the end of the day, while we look to the potential of systemic change starting now, those with power over algorithmic systems should remain accountable for the outcomes of those systems. Let “soft” audits not placate and distract us from organizing for refusal, demanding hard regulation, and raising hell until change occurs.

Conclusion

In this provocation, we have examined three power-related pitfalls of algorithm auditing that illustrate the ineffectiveness of algorithm audits’ espoused theory of change. Looking forward, we call for a reorientation of the AI ethics community that involves commitments to grappling with power dynamics and reprioritizing such that algorithm audits as they are now are no longer pursued as a primary way to address harmful algorithmic systems. We emphasize that the AI ethics community should rethink the structure of the AI accountability ecosystem and direct more focus, attention, and material resource toward supporting and themselves participating in the work that affected people are already doing to address algorithmic system issues.

Acknowledgments

We thank Ali Alkhatib, Blair Attard-Frost, Ken Holstein, Nari Johnson, Sachita Nishal, and our anonymous reviewers for helping us improve our work.

References

- Agnew, W. 2024. What Can AI Ethics Learn from Anarchism? *XRDS*, 30(4): 22–25.
- Ahmed, A. 2020. We Will Not Be Pacified Through Participation. In *Proceedings of the CSCW 2020 Workshop on Collective Organizing and Social Responsibility*. Tech Otherwise.
- Ahmed, S. 2014. *Feminist Attachments*, 168–190. Edinburgh University Press, ned - new edition, 2 edition. ISBN 9780748691135.
- Ahmed, S. 2021. *Complaint!* New York, USA: Duke University Press. ISBN 9781478022336.

- Alkhatib, A. 2021. To Live in Their Utopia: Why Algorithmic Systems Create Absurd Outcomes. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.
- Alkhatib, A. 2024a. Defining AI.
- Alkhatib, A. 2024b. Destroy AI.
- Angwin, J.; Larson, J.; Mattu, S.; and Kirchner, L. 2016. Machine Bias.
- Arnstein, S. R. 1969. A Ladder Of Citizen Participation. *Journal of the American Institute of Planners*, 35(4): 216–224.
- Asher-Schapiro, A. 2020. INTERVIEW - 'Algorithms of oppression': Big tech urged to combat discrimination.
- Attard-Frost, B. 2025. Transfeminist AI governance. *First Monday*, 30(4).
- Bailey, A. 2018. On Anger, Silence, and Epistemic Injustice. *Royal Institute of Philosophy Supplement*, 84: 93–115.
- Bandy, J. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).
- Barabas, C. 2022. Refusal in Data Ethics: Re-Imagining the Code Beneath the Code of Computation in the Carceral State. *Engaging Science, Technology, and Society*, 8(2): 23.
- Benjamin, R. 2024. *Imagination: A Manifesto*. New York, USA: W. W. Norton & Company. ISBN 9781324020974.
- Bietti, E. 2020. From ethics washing to ethics bashing: a view on tech ethics from within moral philosophy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 210–219. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- Birhane, A. 2021. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2): 100205.
- Birhane, A.; Isaac, W.; Prabhakaran, V.; Diaz, M.; Elish, M. C.; Gabriel, I.; and Mohamed, S. 2022a. Power to the People? Opportunities and Challenges for Participatory AI. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394772.
- Birhane, A.; Ruane, E.; Laurent, T.; S. Brown, M.; Flowers, J.; Ventresque, A.; and L. Dancy, C. 2022b. The Forgotten Margins of AI Ethics. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 948–958. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Birhane, A.; Steed, R.; Ojewale, V.; Vecchione, B.; and Raji, I. D. 2024. AI auditing: The Broken Bus on the Road to AI Accountability. arXiv:2401.14462.
- Bond, E. 2020. Google Fixes Gender Bias in Google Translate (Again).
- Calacci, D.; and Pentland, A. 2022. Bargaining with the Black-Box: Designing and Deploying Worker-Centric Tools to Audit Algorithmic Management. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).
- Callon, M. 1984. Some elements of a sociology of translation: domestication of the scallops and the fishermen of St Brieuc Bay. *The Sociological Review*, 32(S1): 196–233.
- Cave, S.; Dihal, K.; and Dillon, S. 2020. Introduction: Imagining AI. In *AI Narratives: A History of Imaginative Thinking about Intelligent Machines*. Oxford University Press. ISBN 9780198846666.
- Cherry, M. 2020. Anger Can Build a Better World.
- Costanza-Chock, S.; Raji, I. D.; and Buolamwini, J. 2022. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1571–1583. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Dalal, S.; Hall, S. M.; and Johnson, N. 2024. Provocation: Who benefits from "inclusion" in Generative AI? arXiv:2411.09102.
- Deng, W. H.; Guo, B.; DeVrio, A.; Shen, H.; Eslami, M.; and Holstein, K. 2023. Understanding Practices, Challenges, and Opportunities for User-Engaged Algorithm Auditing in Industry Practice. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- DeVos, A.; Dhabalia, A.; Shen, H.; Holstein, K.; and Eslami, M. 2022. Toward User-Driven Algorithm Auditing: Investigating users' strategies for uncovering harmful algorithmic behavior. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450391573.
- DeVrio, A.; Eslami, M.; and Holstein, K. 2024. Building, Shifting, & Employing Power: A Taxonomy of Responses From Below to Algorithmic Harm. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1093–1106. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Diberardino, N.; Baleshta, C.; and Stark, L. 2024. Algorithmic Harms and Algorithmic Wrongs. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1725–1732. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Dieterich, W.; Mendoza, C.; and Brennan, T. 2016. COMPAS Risk Scales: Demonstrating Accuracy Equity and Predictive Parity.
- Doctor, N.; Fiennes, G.; and O'Connor, C. 2024. NazTok: An organized neo-Nazi TikTok network is getting millions of views.
- Eubanks, V. 2018a. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. St. Martin's Press. ISBN 9781250074317.
- Eubanks, V. 2018b. A Child Abuse Prediction Model Fails Poor Families.

- Eubanks, V. 2018c. A response to Allegheny County DHS.
- Fitzgerald, R. 2018. DHS Response to 'Automated Inequality' by Virginia Eubanks.
- Forsythe, D. E. 2001. *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press. ISBN 0804742030.
- Friedler, S.; Singh, R.; Blili-Hamelin, B.; Metcalf, J.; and Chen, B. J. 2023. AI Red-Teaming Is Not a One-Stop Solution to AI Harms: Recommendations for Using Red-Teaming for AI Accountability.
- Gansky, B.; and McDonald, S. 2022. CounterFAccTual: How FAccT Undermines Its Organizing Principles. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 1982–1992. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Geiger, R. S.; Tandon, U.; Gakhokidze, A.; Song, L.; and Irani, L. 2024. Making Algorithms Public: Reimagining Auditing from Matters of Fact to Matters of Concern. *International Journal of Communication*, 18: 634–655.
- Gerchick, M. K.; Encarnación, R.; Tanigawa-Lau, C.; Armstrong, L.; Gutiérrez, A.; and Metaxa, D. 2025. Auditing the Audits: Lessons for Algorithmic Accountability from Local Law 144's Bias Audits. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, 29–44. New York, NY, USA: Association for Computing Machinery. ISBN 9798400714825.
- Gonen, H.; and Goldberg, Y. 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 609–614. Minneapolis, Minnesota: Association for Computational Linguistics.
- Goodman, E. P.; and Trehu, J. 2022. AI Audit Washing and Accountability.
- Goodman, R. 2015. The Good Wife Tackles Algorithmic Discrimination. Meanwhile, in Real Life...
- Gorz, A. 1987. Strategy for Labor. In Nissen, B.; and Larson, S., eds., *Theories of the Labor Movement*, 100–116. Wayne State University Press.
- Green, B. 2021. Data Science as Political Action: Grounding Data Science in a Politics of Justice. *Journal of Social Computing*, 2(3): 249–265.
- Groves, L.; Metcalf, J.; Kennedy, A.; Vecchione, B.; and Strait, A. 2024. Auditing Work: Exploring the New York City algorithmic bias audit regime. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1107–1120. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Gururaja, S.; Bertsch, A.; Na, C.; Widder, D.; and Strubell, E. 2023. To Build Our Future, We Must Know Our Past: Contextualizing Paradigm Shifts in Natural Language Processing. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13310–13325. Singapore: Association for Computational Linguistics.
- Hamidi, F.; Scheuerman, M. K.; and Branham, S. M. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, 1–13. New York, NY, USA: Association for Computing Machinery. ISBN 9781450356206.
- Hamilton, A. M. 2021. Silicon Valley Pretends That Algorithmic Bias Is Accidental. It's Not.
- Harrington, C.; Erete, S.; and Piper, A. M. 2019. Deconstructing Community-Based Collaborative Design: Towards More Equitable Participatory Design Engagements. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).
- Harris, P. 2003. Gatekeeping and Remaking: The Politics of Respectability in African American Women's History and Black Feminism. *Journal of Women's History*, 15(1): 212–220, 238. Name - Harvard University Press; University of Houston; University of North Carolina Press; Copyright - Copyright Indiana University Press Spring 2003; People - Wolcott, Victoria W; Schechter, Patricia; White, Frances; Last updated - 2023-11-28; SubjectsTermNotLitGenreText - Wolcott, Victoria W; Schechter, Patricia; White, Frances; Detroit Michigan; United States-US.
- Haskins, C. 2019. Dozens of Cities Have Secretly Experimented With Predictive Policing Software.
- Hassein, N. 2017. Against Black Inclusion in Facial Recognition.
- Higginbotham, E. B. 1993. *Righteous Discontent: The Women's Movement in the Black Baptist Church: 1880-1920*. Harvard University Press.
- Hope, A.; D'Ignazio, C.; Hoy, J.; Michelson, R.; Roberts, J.; Krontiris, K.; and Zuckerman, E. 2019. Hackathons as Participatory Design: Iterating Feminist Utopias. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, 1–14. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359702.
- Huber, L. 2025. "A Bridge to Nowhere": A Healthcare Case Study for Non-Reformist Design. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Huber, L.; Reynolds-Cuellar, P.; DeVrio, A.; Raihan, J.; Sum, C. M.; Dombrowski, L.; Zhang, J.; Becker, C. B.; Irani, L.; Krafft, P. M.; and Hughes, M. 2025. From Tech Lash to Tech Fash: Strategic Reflections on a Decade of Collective Organizing in Computing. In *Adjunct Proceedings of the Sixth Decennial Aarhus Conference: Computing X Crisis*, AAR Adjunct '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400719684.
- Kalluri, P. 2020. Don't ask if artificial intelligence is good or fair, ask how it shifts power.

- Katell, M.; Young, M.; Dailey, D.; Herman, B.; Guetler, V.; Tam, A.; Bintz, C.; Raz, D.; and Krafft, P. M. 2020. Toward situated interventions for algorithmic equity: lessons from the field. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 45–55. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- Keyes, O. 2018. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proc. ACM Hum.-Comput. Interact.*, 2(CSCW).
- Keyes, O.; and Austin, J. 2022. Feeling fixes: Mess and emotion in algorithmic audits. *Big Data & Society*, 9(2): 20539517221113772.
- Keyes, O.; Hutson, J.; and Durbin, M. 2019. A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI EA '19, 1–11. New York, NY, USA: Association for Computing Machinery. ISBN 9781450359719.
- Krafft, P. M.; Young, M.; Katell, M.; Lee, J. E.; Narayan, S.; Epstein, M.; Dailey, D.; Herman, B.; Tam, A.; Guetler, V.; Bintz, C.; Raz, D.; Jobe, P. O.; Putz, F.; Robick, B.; and Barghouti, B. 2021. An Action-Oriented AI Policy Toolkit for Technology Audits by Community Advocates and Activists. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 772–781. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Kuczmarski, J. 2018. Reducing gender bias in Google Translate.
- Kulynych, B.; Overdorf, R.; Troncoso, C.; and Gürses, S. 2020. POTs: protective optimization technologies. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 177–188. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- Longpre, S.; Klyman, K.; Appel, R. E.; Kapoor, S.; Bommasani, R.; Sahar, M.; McGregor, S.; Ghosh, A.; Blili-Hamelin, B.; Butters, N.; Nelson, A.; Elazari, A.; Sellars, A.; Ellis, C. J.; Sherrets, D.; Song, D.; Geiger, H.; Cohen, I.; McIlvenny, L.; Srikumar, M.; Jaycox, M. M.; Anderljung, M.; Johnson, N. F.; Carlini, N.; Mialhe, N.; Marda, N.; Henderson, P.; Portnoff, R. S.; Weiss, R.; Westerhoff, V.; Jernite, Y.; Chowdhury, R.; Liang, P.; and Narayanan, A. 2025. In-House Evaluation Is Not Enough: Towards Robust Third-Party Flaw Disclosure for General-Purpose AI. arXiv:2503.16861.
- Lorde, A. 1981. The Uses of Anger: Women responding to racism.
- Lorde, A. 1984. *Sister Outsider: Essays and Speeches*. Trumansburg, NY: The Crossing Press. ISBN 0895941422.
- Lum, K.; and Chowdhury, R. 2021. What is an “algorithm”? it depends whom you ask.
- MacMillan, D.; Ovalle, D.; and Schaffer, A. 2025. Arrested by AI: Police ignore standards after facial recognition matches.
- Meaker, M. 2024. Algorithms Policed Welfare Systems For Years. Now They're Under Fire for Bias.
- Metaxa, D.; Park, J. S.; Robertson, R. E.; Karahalios, K.; Wilson, C.; Hancock, J.; and Sandvig, C. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Foundations and Trends® in Human-Computer Interaction*, 14(4): 272–344.
- Metcalf, J.; Moss, E.; Watkins, E. A.; Singh, R.; and El-ish, M. C. 2021. Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 735–746. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Muller, M. 2004. HCI as Translation Work: How Translation Studies can Inform HCI Research and Practice. In *Presentations at the CHI 2004 Workshop on Reflective HCI*, CHI '04. New York, NY, USA: Association for Computing Machinery.
- Nafus, D.; and Sherman, J. 2014. This One Does Not Go Up To 11: The Quantified Self Movement as an Alternative Big Data Practice. *International Journal of Communication*, 8(0).
- Noble, S. U. 2012. *Searching for black girls: old traditions in new media*. Phd thesis, University of Illinois at Urbana-Champaign.
- Noble, S. U. 2013. Google Search: Hyper-visibility as a Means of Rendering Black Women and Girls Invisible. In *Visible Culture: A Journal for Visual Culture*, 19.
- Noble, S. U. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press. ISBN 9781479849949.
- of Queer in AI, O.; Dennler, N.; Ovalle, A.; Singh, A.; Soldaini, L.; Subramonian, A.; Tu, H.; Agnew, W.; Ghosh, A.; Yee, K.; Peradejordi, I. F.; Talat, Z.; Russo, M.; and Pinhal, J. D. J. D. P. 2023. Bound by the Bounty: Collaboratively Shaping Evaluation Processes for Queer AI Harms. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 375–386. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.
- Ojewale, V.; Steed, R.; Vecchione, B.; Birhane, A.; and Raji, I. D. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, CHI '25. New York, NY, USA: Association for Computing Machinery. ISBN 9798400713941.
- Olteanu, A.; Ekstrand, M.; Castillo, C.; and Suh, J. 2023. Responsible AI Research Needs Impact Statements Too. arXiv:2311.11776.
- Pitcan, M.; Marwick, A. E.; and boyd, d. 2018. Performing a Vanilla Self: Respectability Politics, Social Class, and the Digital World. *Journal of Computer-Mediated Communication*, 23(3): 163–179.
- Postman, N. 1998. Five Things We Need to Know About Technological Change.

- Prates, M. O. R.; Avelar, P. H.; and Lamb, L. C. 2019. Assessing gender bias in machine translation: a case study with Google Translate. *Neural Computing and Applications*, 32: 6363–6381.
- Raji, I. D.; and Buolamwini, J. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 429–435. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.
- Raji, I. D.; and Buolamwini, J. 2022. Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. *Commun. ACM*, 66(1): 101–108.
- Raji, I. D.; Gebru, T.; Mitchell, M.; Buolamwini, J.; Lee, J.; and Denton, R. 2020a. Saving Face: Investigating the Ethical Concerns of Facial Recognition Auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, AIES '20, 145–151. New York, NY, USA: Association for Computing Machinery. ISBN 9781450371100.
- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020b. 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*, FAT* '20, 33–44. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.
- Raji, I. D.; Xu, P.; Honigsberg, C.; and Ho, D. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '22, 557–571. New York, NY, USA: Association for Computing Machinery. ISBN 9781450392471.
- Sandvig, C.; Hamilton, K.; Karahalios, K.; and Langbort, C. 2014. Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and Discrimination: Converting Critical Concerns into Productive Inquiry*.
- Sankin, A.; Mehrotra, D.; Mattu, S.; and Gilbertson, A. 2021. Crime Prediction Software Promised to Be Free of Biases. New Data Shows It Perpetuates Them.
- Scheuerman, M. K.; Paul, J. M.; and Brubaker, J. R. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.*, 3(CSCW).
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68. New York, NY, USA: Association for Computing Machinery. ISBN 9781450361255.
- Shelby, R.; Rismani, S.; Henne, K.; Moon, A.; Roshtamzadeh, N.; Nicholas, P.; Yilla-Akbari, N.; Gallegos, J.; Smart, A.; Garcia, E.; and Virk, G. 2023. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, 723–741. New York, NY, USA: Association for Computing Machinery. ISBN 9798400702310.
- Shen, H.; DeVos, A.; Eslami, M.; and Holstein, K. 2021. Everyday Algorithm Auditing: Understanding the Power of Everyday Users in Surfacing Harmful Algorithmic Behaviors. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).
- Sloane, M.; Moss, E.; Awomolo, O.; and Forlano, L. 2022. Participation Is not a Design Fix for Machine Learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '22. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394772.
- Smith IV, J. 2016. (Exclusive) Crime-prediction tool PredPol amplifies racially biased policing, study shows.
- Suchman, L. 2002. Located accountabilities in technology production. *Scand. J. Inf. Syst.*, 14(2): 91–105.
- Suchman, L. 2023. The uncontroversial 'thingness' of AI. *Big Data & Society*, 10(2): 20539517231206794.
- Sum, C. M.; and DeVrio, A. 2024. Shifting Power Through Resistance: Social Justice in Practice. *XRDS*, 30(4): 7–9.
- Sum, C. M.; Tran, A.-T.; Lin, J.; Kuo, R.; Bennett, C. L.; Harrington, C.; and Fox, S. E. 2023. Translation as (Re)Mediation: How Ethnic Community-Based Organizations Negotiate Legitimacy. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI '23. New York, NY, USA: Association for Computing Machinery. ISBN 9781450394215.
- Suresh, H.; Movva, R.; Dogan, A. L.; Bhargava, R.; Cruxen, I.; Cuba, A. M.; Taurino, G.; So, W.; and D'Ignazio, C. 2022. Towards Intersectional Feminist and Participatory ML: A Case Study in Supporting Femicide Counterdata Collection. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, 667–678. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Sweeney, L. 2013. Discrimination in Online Ad Delivery: Google ads, black names and white names, racial discrimination, and click advertising. *Queue*, 11(3): 10–29.
- Terzis, P.; Veale, M.; and Gaumann, N. 2024. Law and the Emerging Political Economy of Algorithmic Audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1255–1267. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Upadhyay, S.; Pruss, D.; DeVrio, A.; Gajos, K. Z.; and Karusala, N. 2026. "It just requires so much more creativity": Barriers and Workarounds to Gathering Information for AI Contestation. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, CHI '26. New York, NY, USA: Association for Computing Machinery. ISBN 9798400722783.
- Vecchione, B.; Levy, K.; and Barocas, S. 2021. Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies. In *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, EAAMO '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450385534.

- Vincent, N.; Li, H.; Tilly, N.; Chancellor, S.; and Hecht, B. 2021. Data Leverage: A Framework for Empowering the Public in its Relationship with Technology Companies. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 215–227. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Vinsel, L. 2021. You're Doing It Wrong: Notes on Criticism and Technology Hype.
- Whitney, C. D.; Naval, T.; Quepons, E.; Singh, S.; Rick, S. R.; and Irani, L. 2021. HCI Tactics for Politics from Below: Meeting the Challenges of Smart Cities. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21. New York, NY, USA: Association for Computing Machinery. ISBN 9781450380966.
- Widder, D. G. 2024. Epistemic Power in AI Ethics Labor: Legitimizing Located Complaints. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1295–1304. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Widder, D. G.; and Goues, C. L. 2024. What Is a 'Bug'? *Commun. ACM*, 67(11): 32–34.
- Widder, D. G.; and Nafus, D. 2023. Dislocated accountabilities in the “AI supply chain”: Modularity and developers' notions of responsibility. *Big Data & Society*, 10(1): 20539517231177620.
- Widder, D. G.; Whittaker, M.; and West, S. M. 2024. Why 'open' AI systems are actually closed, and why this matters. *Nature*, 635: 827–833.
- Wilkinson, D.; Ekstrand, M.; Vertesi, J. A.; and Olteanu, A. 2023. Theories of Change in Responsible AI.
- Wong, R. Y. 2021. Tactics of Soft Resistance in User Experience Professionals' Values Work. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).
- Wright, L.; Muenster, R. M.; Vecchione, B.; Qu, T.; Cai, P. S.; Smith, A.; Investigators, C. . S.; Metcalf, J.; and Matias, J. N. 2024. Null Compliance: NYC Local Law 144 and the challenges of algorithm accountability. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1701–1713. New York, NY, USA: Association for Computing Machinery. ISBN 9798400704505.
- Yin, L.; and Sankin, A. 2020. Google Ad Portal Equated “Black Girls” with Porn.
- Zong, J.; and Matias, J. N. 2024. Data Refusal from Below: A Framework for Understanding, Evaluating, and Envisioning Refusal as Design. *ACM J. Responsib. Comput.*, 1(1).