# AI auditing: The Broken Bus on the Road to AI Accountability

Anonymous Authors

***Abstract*—One of the most concrete measures to take towards meaningful AI accountability is to consequentially assess and report the systems' performance and impact. However, the practical nature of the "AI audit" ecosystem is muddled and imprecise, making it difficult to work through various concepts and map out the stakeholders involved in the practice. First, we taxonomize current AI audit practices as completed by regulators, law firms, civil society, journalism, academia, consulting agencies. Next, we assess the impact of audits done by stakeholders within each domain. We find that only a subset of AI audit studies translate to the desired accountability outcomes. We thus assess and isolate practices necessary for effective AI audit results, articulating the observed connections between AI audit design, methodology and institutional context on its effectiveness as a meaningful mechanism for accountability.**

***Index Terms*—Evaluation, auditing, accountability, transparency, artificial intelligence, society, law, machine learning, data science**

## I. Introduction

The widespread use of artificial intelligence (AI) systems is heavily weighed down by its multitude of related risks. Functional failures [1], disparate performance [2]–[4], embedded stereotypes [5]–[7], legal incompatibility [8], privacy violations [9], model inscrutability [10] and many more issues plague almost every use.

Audits are a routine practice with well established standards in various sectors including banking, finance, public management, accounting, healthcare, anthropology, international development, and governmental bodies [11]–[13]. The adoption of audits within the AI space, however, is relatively new. Despite its imprecision, we use the term "AI" to encompass a wide range of deployed products with a significant algorithmic component, including but not limited to risk assessments, large base models for computer vision and language, classification models, "generative" models, and recommendation systems. The inspiration for audit practice in the field of data science, machine learning (ML), and AI derives from a variety of related disciplines. Online platform audits, for example, cite social science and critical race studies as inspiration [14], [15], as do several audits of automated decision systems (ADS) and risk assessments [16]. Some audits of large language models take after security audits [17]. Some risk assessment evaluations follow models from traditional experimental design [18], including clinical trials [19]. Meanwhile, internal auditors derive their practice from regulated industries such as finance, aerospace and medical devices [20].

Across disciplines and contexts, one of the main motivations for conducting audits of AI systems is establishing informed and consequential judgements of the deployed AI systems – that is algorithmic accountability [21]. We thus define an AI audit to be any evaluation of these AI systems, independent of the AI development process, executed for the purpose of accountability [22].

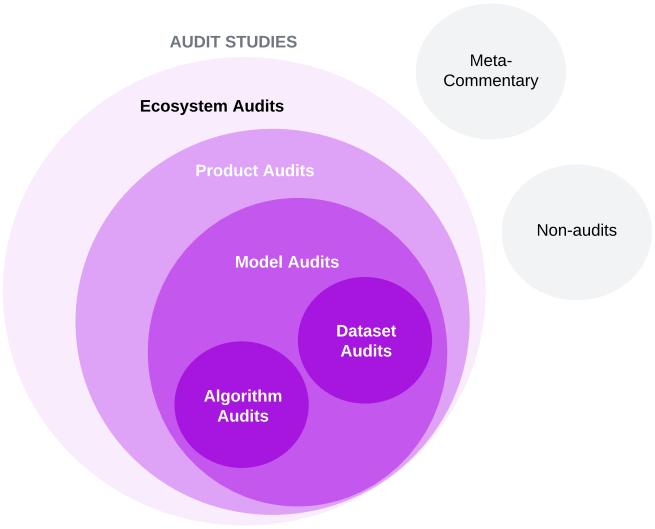


Fig. 1. Audit studies considered in our survey, classified by their scope.

However, unlike other, more mature audit industries, AI audit studies do not consistently translate into more concrete objectives to regulate system outcomes. This means AI audits rarely influence voluntary corporate action or internal corporate policies, inciting product recalls, informing product redesigns, governmental policy and regulation in the form of bans, and restrictions or moratoriums of use. In this paper, with the aim of enabling auditors and policymakers to achieve accountability outcomes, we taxonomize audit practices and identify the characteristics of audits that most directly and effectively contribute to audit objectives.

Some initial work has been done to taxonimize various aspects of AI auditing. Many of these efforts have focused on categorizing *methods* [14], [16], [23]–[25], or *types of audit organizations* [26], [27]. Some efforts have also looked at an AI audit's broader institutional [20], [28], policy [27], and societal context [16], through this analysis is often limited to one or at most a handful of case studies. In this regard Bandy [29] has systematically classifed audit studies by a broader set of criteria – audit method, audit target, audit domain and audit objective. However, the study is narrow in scope, consisting

of 62, exclusively academic audit studies.

In this paper, we take a much broader and more comprehensive view than past studies. We review the broad audit landscape consisting of academia and six other domains, taxonomizing the key characteristics of their context, goals and practice (Figure 1). Notably, as a mechanism for achieving accountability, we investigate the consequences of these audit investigations and how well the outcomes from the studies match those stated objectives.

## II. Background

### A. What is an AI audit?

In the context of this paper, we consider the operationalization of audits as a mechanism for accountability in computing — notably for AI, machine learning, data science, and related fields. We begin by proving definitions.

*Definition 1:* An ***audit*** is defined as any *independent assessment* of *an identified audit target* via *an evaluation of articulated expectations* with *the implicit or explicit objective of accountability*.

Care must be taken to translate this definition in the AI context. By *independent assessment*, we mean any measurement done by an entity operationally distinct from the team that engineered the examined AI system. Even if that team is within the same company and composed of corporate employees, the examination only counts as an audit if the auditors are adequately separate from those that built the system, or the audit process itself is distinct from the engineering process for the AI system [20], [27]. By *an identified audit target*, we look to studies and investigations that name a concrete and non-abstract, specific object of examination. Ideally, this target is connected to a real-world AI deployment, though sometimes a widely used open-source algorithm or dataset can operate as a stand-in or proxy. For instance, Steed and Caliskan [30] looked at bias in an open-source image generation model, reflecting issues later discovered in similar commercial products [31]; and an audit of Proctorio was conducted on the open source model OpenCV on which the commercial product is built [32]. Due to a lack of training data disclosure for most AI products (even supposedly "open source" models [33]), many data audits investigate open source datasets [34]–[36] and attempt to generalize conclusions on broader industry practice. Studies of prototypical algorithms (hypothetical models trained by the authors) without concrete, specific targets were not considered.

To be an *evaluation with the objective of accountability*, the audit must incorporate some implied or explicit objective to have the assessment play some role informing consequential judgements about the technology being examined. In order for these judgements to be more concrete, there needs to be some measurement between the reality of the AI deployment and articulated expectations held about a particular deployment. Many academic papers that we examined, for example, studying fairness, performance or safety in an abstract manner [37], without connecting it to anticipated accountability outcome, even implicitly, were not considered by us to be audits [38]. Note that the actual *type* of audit target or *criteria of assessment* are not part of the definition of what constitutes an audit. AI audits can thus encompass a wide range of targets, including automated decision systems (ADS) [39], [40], recommendation systems underlying online platforms or apps [41], [42], large base models in computer vision [30], [43], speech [44], text-based natural language processing [17], or multimodal models [45]. At times, the evaluations involve domain-specific considerations in hiring [46], healthcare [2], criminal justice [47] or social service delivery [48]. The expectations articulated for these systems can also vary in concreteness and specificity. For instance, some conduct audits specifically for legal compliance [49] while others declare expectations more normatively [38]. Others yet are not explicitly labelled as audit work yet satisfy our audit criteria and result in immense explicit and gradual structural change [50], [51]. As a result, we can observe audits that can evaluate and diagnose for a range of performance, safety concerns, as well as broader societal injustices and are not limited to fairness.

### B. Who conducts AI audits?

There is a wide range of possible audit practitioners that participate in the audit process [26], often described as below.

*Definition 2:* An *internal auditor* is an entity executing an audit or investigation with some contractual relationship with the audit target [27]. They typically seek to minimize corporate liability and test for compliance to corporate or industry-wide expectations [20]. In policy, internal auditors are typically those designated to carry out mandatory corporate audit requirements (e.g. the "independent auditors" in Article 37 of the Digital Services Act).

*Definition 3:* An *external auditor* is an entity executing an audit or investigation without any contractual relationship with the audit target [27]. They typically execute audits voluntarily with a broader mandate of identifying and minimizing the harm impacting their constituents.

*Internal audits* require a contractual relationship with the audit target. Internal audits are typically conducted by an organization hired by the audit target voluntarily or to maintain compliance with a required legal audit mandate. The auditor in these contexts are hired to operate in a professional capacity to audit the target. This often means they are selected and paid by the audit target, though that is not always necessarily the case (e.g., auditors selected and paid by the government). These are the auditors typically referenced in audit mandates. As the executors of more formal audit requirements, these auditors are ideally certified or otherwise qualified [27], and subject to some form of external oversight and quality control, including but not limited to auditor conduct and reporting standards.

*External auditors* typically conduct audits voluntarily by organizations, typically with a broader mandate of research or advocacy. These auditors are not assigned an audit target and do not execute audits on behalf of the audit target but choose to study systems based on the needs and concerns of their constituents. Without any formal connection to the audit target, these auditors typically struggle to access the information necessary to conduct a thorough investigation.

Furthermore, without any form of legal protection, data access regime, or oversight, these auditors tend to operate quite independently, taking on whatever methods suit their objectives. These auditors can be vulnerable to corporate retaliation and methodological skepticism after audit results are released.

### C. *How are AI audits conducted?*

Although the specific methods used in the execution of AI audits vary widely, some terminology are deployed regularly to describe how audits are executed. Audits can be conducted *ex ante* (before deployment), *in media res* (during an iterative process of design and restricted re-deployment) or *ex post* (after widespread adoption and use). Furthermore, although the details of the AI audit process varies widely, the high level structure of that process follows similar stages, which we describe using the following terminology. Note that not all audits follow all these processes.

*1) Harms Discovery:* This is the stage of discovering what to audit for. It involves identifying the audit target entity, possible targeted populations, and the anticipated form of harm or measurement required for a meaningful audit. This can happen, for example, via direct reporting from the impacted population [52], active investigation from the auditors, or other methods [53].

*2) Standards Identification:* This stage is about effectively articulating the requirements for an ideal AI audit outcome, by naming the standards the auditors will be holding the target to in the evaluation process. These expectations can be as vague as a set of named AI principles [54] or as precise as a specific threshold of performance (e.g., AI hiring tools' adherence to the 4/5ths rule [55]).

*3) Performance Analysis:* The core of the audit is the actual evaluation itself. There are a wide diversity of methods available to inform final assessments, ranging from qualitative to quantitative approaches [56]. Each method involves different degrees of complexity and challenges tied to data acquisition, model access, and measurement.

*4) Audit Communication and Advocacy:* Following the evaluation of the AI system, there is often also some activity to disseminate and translate audit results to some relevant stakeholders. That audience could include the audit target, regulators and the public [38], [57]—but could also include internal stakeholders within the organization such as a legal, business, product or engineering team [20].

## III. Methods

In order to review the AI audit research and practice landscape, we conducted a wide-ranging literature review of academic work published in a range of venues. We reviewed websites, reports, and relevant documentations from numerous non-academic audit practitioners. For academic research, we collected a comprehensive sample of $N = 341$ audit studies published in interdisciplinary computing and related venues from 2018–2022. Figure 2 shows the number of collected academic studies published per year, grouped by audit label type, while Table I describes our search method for each source and lists the number of identified pieces of literature collected for each. For audit practices outside academia, we identified major domains that are both relatively established in existing academic literature and currently emerging as key players in the AI audit ecosystem (inspired by the main categories identified by [26]): journalism, civil society, government, consulting agencies and corporate audits, and legal firms. We collected specific cases that serve as concrete examples of each domain. These methodological choices not only facilitate coherence with existing scholarship but also enhance the reliability and comparability of our findings within broader academic discourse on the subject.

Our list is not exhaustive of all academic audit literature or audit domains, but it provides the most comprehensive sample (to our knowledge) of the current state of AI audit ecosystem, encompassing from academic research to practice and everything in-between. The full list of audit studies we found using these methods is included in the supplementary information (see Appendix A).

### A. *Academic literature*

Audits in the academic context, conducted by authors from universities, non-profits, and tech companies, encompass a wide variety of disciplines, methods, and aims. Academic audits are published in various formats and venues, for example, as books, journals, and conference proceedings. Over the past five years, interdisciplinary computing conferences have become a central place where work around the topics of fairness, accountability, and transparency is discussed and published, especially within the ACM conferences, FAccT and AIES, two emerging main conferences in the space. In this regard, conference proceedings are not only the most common way of sharing work amongst the AI ethics (broadly defined) community, but such formats of publication also share commonalities such as relatively standardised style and presentation. We have, therefore, selected conference proceedings as a primary focus for our systematic reviews of academic audits.

Our search is not exhaustive—it did not, for example, include potential audit work that might have been covered in high-impact general audience journals such as Science, Nature, or PNAS. We also did not consider work published in non-interdisciplinary social science venues—in particular, work from economics or sociology. Instead, our sample represents the kinds of audit work recently published at interdisciplinary computing and related conferences.

*1) Searching for audit studies:* We found $N = 341$ academic papers that met our criteria for an audit study (Figure 2, Table I). We identified academic audit studies with two methods. First, we manually reviewed the conference proceedings from 2018 to 2022 for a selection of smaller interdisciplinary computing conferences: Fairness, Accountability, and Transparency (FAccT), Artificial Intelligence, Ethics, and Society (AIES), Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO), Association for the Advancement of Artificial Intelligence (AAAI), Computer-Supported Cooperative Work & Social Computing (CSCW), International

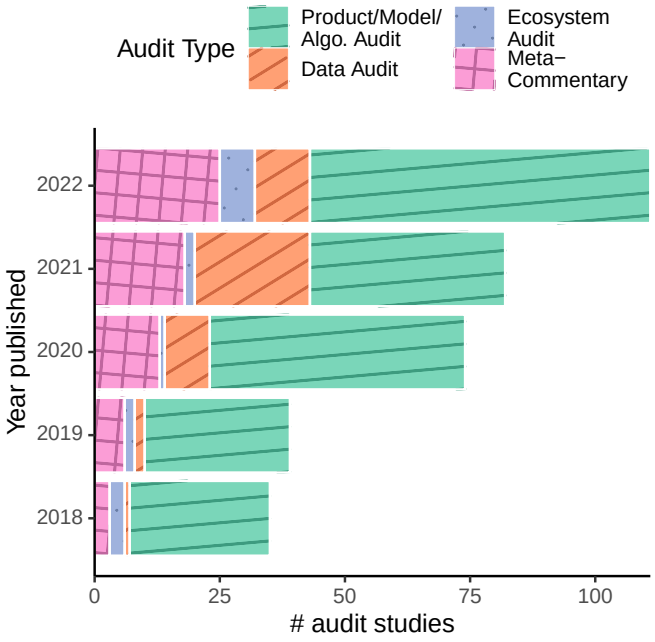


Fig. 2. Number of collected academic audit studies published in each year, grouped by manually labelled audit type. Studies that involve multiple audit types are counted only under the best-fitting category.

Conference on Computational Social Science (IC2S2), and the International World Wide Web Conference (WWW). We selected these conferences as venues where audit work is likely to appear, but still, audits are not the primary focus of these conferences. Of the 5 years of proceedings we reviewed, only $N = 237$ papers met our criteria for an audit study.

Second, to expand our search to other, larger conferences, we automated the inclusion process: we applied keyword searches to both the ACM Digital Library (excluding FAccT, AIES, and CSCW proceedings, which we reviewed manually) and the ACL Anthology. In these two libraries, we searched for papers published from 2018 to 2022 with one or more of six keywords in the title: "audit", "accountability", "case study", "bias", "fairness", and "assurance". For ACL, where the searches were web-based, we manually reviewed the search results for audit studies online, downloading only the few studies that met our criteria ($N = 6$). For ACM, we downloaded the search results. There were thousands of papers in the queries in total. Three of the keywords yielded relatively few ACM entries: "audit" ($N = 39$), "accountability" ($N = 87$), and "assurance" ($N = 65$), after removing duplicates. For the rest of the keywords, which returned over 500 results each, we narrowed our manual search by considering only the first 500 "most relevant" entries returned by the ACM search engine. After removing duplicates and other incomplete entries, each keyword yielded the following number of pieces of literature: "fairness" ($N = 404$), "bias" ($N = 318$), and "case study" ($N = 472$). We further excluded everything that is not a standard academic paper (such as panels, abstracts, workshops, and tutorials) for consistency. We then manually reviewed each abstract and, as before, excluded those that did not meet our criteria for an audit study, leaving a total of $N = 103$ audit studies from ACM conferences other than FAccT and AIES.

TABLE I
AUDIT LITERATURE FROM ACADEMIC SOURCES

| Source | Search method | $N$ |
|---|---|---|
| FAccT | Manual review of proceedings — titles and abstracts, then full paper where classification still unclear | 55 |
| AIES | " | 40 |
| EAAMO, AAAI, IC2S2, and WWW | " | 137 |
| Other ACM proceedings | Keyword filters, then manual review | 103 |
| ACL Anthology | " | 6 |
| | **Total** | 341 |

*2) Classifying audit types:* We classified all the audit studies we collected into two main categories: audit studies and meta-commentary (depicted in Figure 1). We further classified audit studies into two kinds: *product/model/algorithm* audit studies and *data* audit studies. Some studies targeted specific products (e.g., YouTube comment moderation [58]); others targeted only the AI models behind the products (e.g., an investigation of tax audit models used by the IRS [59]); still others targeted only the algorithms used to build the model (e.g., pre-trained large language models such as GPT-3 [60]). Since the methodologies for these kinds of audits are similar, we classify them together. We distinguish these studies from *data* audit studies—also a kind of model audit—which specifically target training or benchmark datasets (e.g., ImageNet [61]) and tend to use different methodologies.

During the iterative classification process, an additional category emerged: *ecosystem* audit studies. Ecosystem audit go beyond datasets, models, and products and examine communities and sociotechnical environments (digital or physical) impacted by or are critical components to an AI system's operation. Brown *et al.* [62], for example, audited the impact of child welfare service algorithms by conducting interactive workshops with front-line service providers and families affected by such algorithms. *Meta-commentary* studies are those that examine auditing as a practice, including studies that putforward novel mechanisms and processes as a viable *method* for auditing certain systems as well as those proposing to improve existing audit processes. Meta-commentary also includes *critiques*, work that interrogates the effectiveness and merit of auditing as a practice.

Note that not all papers fit neatly into one of the above categories, and sometimes we find some papers incorporating elements of two—and, rarely, three—categories. For example, methodology papers often also use a specific case study to illustrate the method in practice. In our analysis (Fig. 2), we consider only the classification that best fits the paper.

### B. *Non-academic domains*

For audit work outside the academic domain, we identified the following categories where audits are regularly conducted

and becoming an established practice: journalism, civil society, law firms, regulatory audits, and consulting agencies and corporate audits. For each category, we gathered sources, often information on auditors websites, audit reports, and other relevant documentation produced by various bodies (whenever they are available), which we reference in our analysis (§V). Based on information from these sources, for each category we identified a number of institutions, organisations, agencies, or audit reports that served as concrete examples of the given category. Appendix A lists all the reports, webpages, and other documents we reviewed in each audit category and institution.

### C. *Analysis*

For both the academic audits and each non-academic audit source, we reviewed published audit studies, documents, and public webpages to summarize information on the following:

*1) Context:* Context includes the specific structural factors shaping the audit (all as stated by the auditor): what motivated the audit (*Motivations*), its goals (*Goals*), the main artefact or system under investigation (*Audit Target*), the specific harms or concerns investigated (*Types of Harms*), and ways in which the audit may or may not shift power between stakeholders [63] (*Institutional Context*).

*2) Methodology:* Methodology includes the main techniques and procedures used to investigate the target artefact—for example, quantitative evaluation (typical in academic audit studies) or forensic analysis or qualitative interviews (found more often in civil society audits).

*3) Impact:* Impact refers to changes that occur to the target artefact, the target audit, or the institutional environment as a direct consequence of the audit. Impacts could include policy developments, alterations to an algorithm, or monetary fines for certain violations. While impacts are well documented in some domains (e.g., journalism), the impact of an academic audit, for example, is not often clear at the time an audit is published. Detecting, measuring, and quantifying impact is challenging. Subsequently, we were as inclusive as possible of many different kinds of impacts, e.g., inspiring media coverage of an overlooked harm [64]. We considered notable documented evidence in the audit reports themselves or from related news stories, whenever available. In our analysis, we make a note wherever documented impacts were unstated and unclear (in academia, for example).

For nearly all the academic studies we found, our dataset includes abstracts ($N = 337$) and author-selected keywords ($N = 263$). In order to supplement our qualitative findings for academia (§IV) with quantitative statistics, we analysed key terms (e.g., "accountability") that were used most frequently by authors. Appendix B describes how we selected key terms related to the criteria above.

## IV. Results: Auditing in Academia

### A. *Audit types*

*1) Product/model/algorithm audits:* Most audit work fell under the category of product-level case studies. These audit studies target a mix of social media platforms, algorithms for administering public services, large language and vision models, and search engines. These case studies typically evaluate specific deployed systems. These case studies mainly diagnose failures, errors, disparities. For example, some variation of "bias" is mentioned in 32.5%, and "fairness" in 21.2% of abstracts (see Table B-IV). Less commonly, these studies call for model builders and practitioners to make amendments accordingly. For example, the term "accountability" appeared in only 14% of abstracts, less often than in ecosystem audits (33.3%) or meta-commentary (28.1%).

*2) Data audits:* Data audits typically focus on evaluating specific datasets, often targeting datasets used to train large models [65] and sometimes interrogate the benchmark datasets used for model evaluation, such as COMPAS [66]. Oftentimes, these studies emphasise (both implied and explicit) shifting norms around data use and benchmarking practices, while fewer explicitly emphasize holding dataset creators accountable. For example, "accountability" is mentioned in 9% of data audit abstracts, while "bias" is mentioned in 39.1% of abstracts (more in this category than any other). These studies typically focus on surfacing harms ranging from representation and stereotyping to privacy and, more recently, copyright protection [67]. The type of methods used include quantitative measurement (such as the incidence of NSFW images), simulation, ablation (removing or changing certain aspects of the dataset and measuring the result), and critical assessment. Like case studies, data audits tend to be conducted by a range of academic, non-profit, and corporate authors examining open-sourced and academic datasets.

*3) Ecosystem audits:* These studies target public services —predictive risk models used by child welfare agencies, for example [62], [68]—more often than any other kinds of audits. Abstracts of these studies also mention specific domains such as hiring and education more often than any other audit type (Table B-V). Harms and concerns are often concretely defined and audit are typically carried out with the expectation of subsequent change both to specific stakeholders and sometimes, society at large. The term "accountability", for example, is mentioned in over 33% of ecosystem audit abstracts, the highest of any category. These studies often utilise a wider range of methods, including qualitative interviews, surveys, workshops, and literature reviews [69]–[71]. Given the scope of investigation, methods are often all-encompassing and diffuse, less precise than more scoped audit investigations [72], where they are much more likely to employ qualitative and participatory methods than other audit studies. 6 out of 15 ecosystem audit abstracts, for example, mention ethnography, interviews, workshops, or other qualitative methods, while two explicitly use the term "participatory" (see Table B-VI).

*4) Meta-commentary & critique:* We found that 28.1% of meta-commentaries mention "accountability" in the abstract. The type of methods they used include surveys, interviews, and literature reviews. These kinds of studies were nearly exclusively conducted by academic, non-profit, or government authors—for example, guidance from government agencies such as NIST's AI Risk Management Framework [73].

Many meta-commentary papers aimed to develop a methodology and many considered the development of rigorous audit methods as a vital contribution. Audit methods these studies put forward include both qualitative and quantitative approaches. These studies also examine audit practitioners themselves, interrogating norms around algorithm evaluation and audit practice. The harms at issue tend to be less well-specified; most mention some form of independence, fairness, privacy, or recourse. These studies rarely mention affected stakeholders in method design. Similarly, a couple of tool development papers developed standardised tools, usually quantitative, for auditing or for algorithmic recourse.

### B. *Impact by audit type*

*1) Product/model/algorithm audits:* We found few audit studies that acknowledge and address power asymmetries. Many prominent papers we surveyed are collaborations between authors from academic institutions and large tech corporations, usually examining their own systems and datasets, and rarely explicitly calling for systemic change, auditing systems, often without involvement from affected stakeholders.

Unlike audit work by journalists and civil society, academic audit work tends to follow academic norms where the objective is academic publication. Subsequently, more often than not, audits are seen as an academic, intellectual exercise than practices directly linked to real world consequences. Audit findings are rarely presented with demands for concrete systemic change. This does not, however, mean that academic audits are not impactful. To the contrary, seminal audit case studies such as Gender Shades [43] have not only resulted in meaningful improvements to deployed systems [38] but also have come to establish algorithmic audits as a field of enquiry. The impact of such work is, however, rarely obvious at the time of publication but becomes apparent gradually over time. For industry collaborations, the audits may result in organisational reform. For example, case studies co-authored by authors with big tech affiliations such as [74] may have contributed to tightening Google Play Store app data access restrictions [75]. Generally, academic audit results that publicly call out audit targets [38], that tend to be picked up by news outlets such as MIT Tech Review, activists, and regulators tend to bring about the most observable changes.

*2) Data audits:* The impact of data audits are also often unclear, though prominent data audits sometimes resulted in changes to benchmarks — an audit of 80 Million Tiny Images, for example, resulted in the withdrawal [76] of the dataset [61] and a Financial Times investigation of Microsoft's dataset of "celebrity" faces even resulted in its discontinuation [77], [78].

*3) Ecosystem audits:* Ecosystem audits often reveal comprehensive institutional or legislative policy demands in advocacy. This includes, connecting performance and privacy concerns involved in facial recognition use by law enforcement [69], breaking down sources of bias throughout the model development cycle [79], systematic environmental costs [80], [81] or the labor issues across the entire supply chain of AI development [82].

*4) Meta-commentary & Critique:* Although the immediate impact is directly quantifiable, meta-commentaries provide important mechanisms that allow audit studies and practices to zoom out, self-reflect, and evaluate the overall picture and direction, all of which is a crucial element for grounding audits in concrete foundations and optimal accountability.

More specifically, a certain type of meta-commentary that seems to be of particular importance are critique studies. This category of work engages in reflexive commentary critiques of auditing as a practice. A few pieces of critical work interrogate the effectiveness, shortcomings, and limitations of audits. Such work highlights structural issues such as historical power asymmetries that might be reinforced by common academic auditing practices, or how audits might serve as a smokescreen for corporate responsibility, such as, audit washing [83], [84] or how audits might lead to simplistic technological solutionism [85]. They draw on qualitative interviews, literature reviews, and statistical analysis. Like meta-commentaries, the impacts of these works are not easily measured, though many of these studies are highly cited.

## V. Results: Auditing Outside Academia

In this section, we review a sample of audits and audit practices from outside academia — law firms, consulting agencies and corporate audits, journalism, civil society, government, and civil society — examining how factors such as institutional context affects the practice of auditing and accountability. These auditors operate in a variety of domains and deploy various methods throughout the audit design, development, and execution process. For each domain, we pay particular attention to the details of the audit *context* and *methodology*, then connect this to the observed *impact* derived from audits for that particular domain. Our analysis criteria and results are summarised in Table II (The full list of documents we analysed is found in Appendix A).

### A. *Law Firms*

There is a large industry devoted to data governance legal services, but firms offering legal services for AI auditing are less common. Three such merging boutique law firms are: Luminos.Law [86] (formerly called BHN.AI), Foxglove [87], and AWO [88]. These firms represent three different institutional arrangements in audit-related legal services.

*a) Context:* Law firms operate as both *internal* and *external* auditors. They can be hired by the audit target to conduct an internal assessment as part of a regulatory requirement or in legal defense. Law firms also work with representatives of impacted populations (often pro bono) to externally investigate an audit target to collect material evidence of perceived or reported harms. Both scenarios are typically prompted by individual or collective complaints post-deployment.

The organisational structure, business model, and subsequent selection of clientele determines the general objectives for these three organisations. Foxglove—a non-profit organisation advocating for the least empowered—aims to hold AI operators to account, aspiring to "stand up to tech giants

TABLE II
SUMMARY OF ANALYSIS ACROSS AI AUDITING DOMAINS.

| **Domain** | **Context** | | | | **Methods** | **Impacts** |
|---|---|---|---|---|---|---|
| | *Motivation* | *Target* | *Types of Harms* | *Inst. Context* | | |
| ACADEMIA | Accountability, make policy, establish standards | ADS, online platforms, training datasets, audit practitioners | Functionality failures, disparate impact, privacy harms, stereotyping, manipulation, ethics washing, unfairness | Academic, non-profit, & corp authors studying systems built by same, occasionally involving stakeholders | interview, survey, simulation, quant & qual eval, theory, lit review, framework devt, ablation | Often unclear; media coverage; occasionally, reforms to target systems |
| CIVIL SOCIETY | Accountability, advocacy, make policy | ADS, online platforms, surveillance tech, biometric data use | Fair use harms, privacy harms, censorship, disparate impact, HR violations, deceptive claims, hate speech | Non-profit, independents usually advocating on behalf of harmed stakeholders | Activism, interviews, visualisation/media, forensics, public records requests | Civil suits won, media coverage, moratoriums, abolition |
| JOURNALISM | Accountability | Online platforms, large tech companies, ADS (public services) | disparate impact, fraud, non-compliance, privacy harms, cultural genocide | Targets mostly tech firms & gov't agencies on behalf of/in consultation with public; funded by foundt'ns/private donors | Investigative reporting, quant. eval., interviews, data donation, scraping, docs review | Regulatory action, civil suits, reforms to target systems, inspiration for research |
| GOVERNMENT | Enforce regulations, establish standards | Data sharing and management; surveillance tech, ADS | Non-compliance, privacy harms, racial disparities, physical safety, functionality | Targets companies and gov't agencies; may or may not have enforcement authority | Interviews, testing, docs review, quant. eval. | Monetary penalties, standard-setting |
| CORPORATE AUDIT REPORTS | Identify ethical issues | Internal APIs, platforms | Human rights "impacts", racial injustice, voter suppression, hate speech | "Voluntary" reforms, often at request of & in consultation with external stakeholders | Stakeholder/expert consultation, interviews | Reforms to practices, commitments to future changes |
| LAW FIRMS | Compliance, impact assessment (BNH.AI, AWO); accountability (Foxglove) | Public services, online platforms, gig platforms, other private companies | Non-compliance, injustice/disparate impact, privacy harms, safety, digital manipulation | For-profits and public services (BNH.AI, AWO); consulting for harmed stakeholders (Foxglove) | Testimony, docs review, policy development, legal research | Reform, revision, abolition, forced disclosure of secret contracts, (Foxglove) |
| CONSULTING AGENCIES | Compliance, impact assessment, mitigate financial/PR risk | ADS, online platforms, gig work platforms, surveillance tech | Noncompliance, disparate impact, non-functionality, privacy harms, security, risk | Consulting for for-profit audit targets; may consult stakeholders (Babl AI) | Assessment frameworks, docs review, data governance planning | Set precedent for policy (ORCAA); otherwise not clear or unstated |

ADS: automated decision systems.

and governments" and "make tech fair for everyone" [87]. Accordingly, the main audit target for Foxglove includes data practices of social media platforms, governmental institutions, as well as the working conditions of content moderators, warehouse workers, and gig workers.

Luminos.Law and AWO—both for-profit firms—offer compliance and public policy services. They carry out audit work at the behest of their clients, with the goal of ensuring compliance with standards, data rights, data protection due diligence, and regulatory obligations. Luminos.Law's client testimonials feature Fortune 100 & 500 tech firms [89]; AWO's "commercial practice is balanced with giving those less-resourced a voice", according to the firm, and their client testimonials feature many non-profit and university clients [90]. The type of harm and concerns that underlie these firms' services vary accordingly. For Luminos.Law, these are assessment and assurance around legal compliance, legal defence and liability [91]. Similarly, AWO focuses on auditing for privacy and data rights, safety, surveillance violations, digital manipulation and exploitation [90], [92]. The types of harm and concerns that drive audit practice for Foxglove, on the other hand, include violations of justice, disparate performance [93], [94], and specific breaches of data governance law [95], [96].

We see a stark difference in power asymmetries amongst these three organisations. As for profit forms, both Luminos.Law and AWO's objectives, missions, and practices are shaped by their business models that prioritise the needs of their clients, which are often wealthy and powerful corporations. Foxglove's work directly or indirectly pushes to shift power from the most to the least powerful. Subsequently, Foxglove frequently works with groups such as content moderators, warehouse workers and gig-workers, groups that are often underpaid, over-exploited, and disfranchised.

*b) Methodology:* Foxglove's cases often involve specific campaigns, petitions, and/or case studies. Foxglove uses methods such as legal compliance analysis, interviewing, and anecdotal evidence, or leans on previous research to diagnose and highlight harms, concerns, and to pursue legal challenges. Both Luminos.Law and AWO work within the needs of their clients,

which include big tech companies and start-ups. Within such a context, the main audit target for Luminos.Law are models and data, while audit target include social media platforms [97] and hiring algorithms [98], based on media coverage of the firm [86]. AWO similarly carries out documentation analysis, legal compliance analysis, and policy development using previous research as a basis [92].

*c) Impact:* As firms that provide audits as a service to clients, some details of their practices are inaccessible—particularly for Luminos.Law—because much of their work is explicitly marketed as "privileged and confidential" [89]. Information on the impacts of their commercial work is therefore difficult to identify. AWO's blog, for example, mostly features work done in collaboration with civil society organizations like the Ada Lovelace Institute [99]. Luminos's public portfolio includes some standards guidance, a bias calculator for New York's new audit law [100], [101], and an audit of the open-source large language model RoBERTa [102], [103], but the impacts of these audits on practice are not evident.

Foxglove's work, on the other hand, often results in significant impactful changes including the reversal of Ofqual's A level grading algorithm [104], halting the use of visa-streaming algorithm that was deployed by the UK Home Office [105], and forcing disclosure of a secrete contracts between corporations and the UK government; for example, the "NHS Covid-19 data deals" [106] and exposing contracts between Palantir and the UK government.

### *B. Consulting Agencies & Corporate Audits*

Other organizations offer audit services beyond legal and policy advice. A crop of recent startups, such as, Arthur AI and Fiddler AI, offer model monitoring services with fairness and privacy components. Others offer consulting specifically, auditing as a service, including boutiques such as, Parity Consulting [107] as well as large consulting firms such as Deloitte [108], McKinsey [109], and Accenture [110].

Because the product-specific work stemming from an internal audit team is rarely published, corporations occasionally publish a report in conjunction with an external consulting agency. For example, Business for Social Responsibility (BSR) has conducted audits on behalf of Google for its celebrity facial recognition system [111], and Facebook regarding its human rights impact in Myanmar [112]. At times these consulting agencies are also part of a regulatory process. For example, the management consulting firm called Guidehouse Inc. played the role of an "independent third-party reviewer" in the Department of Justice (in the US) settlement with Facebook regarding bias in its advertising [113].

Here, we examine three consulting agencies that have been involved with prominent audit case studies: ORCAA [114], Eticas [115], and BABL AI [116].

*a) Context:* All three agencies provide *internal* audits as a service with the main objective of algorithmic accountability, and engage primarily in case-by-case consulting with private and public sector clients such as HireVue, Airbnb, Proctorio [116], the states of Illinois and Colorado [114], the Allegheny County Health Department [115], the cities of Barcelona [115] and Amsterdam [114], and the University of Iowa [116]. Within the bounds of client-agency agreement, the main audit targets for these agencies include social media platforms (such as TikTok and YouTube), ADS, FRT, ride hailing apps (such as Uber, Cabify, Bolt), predictive scoring systems, hiring algorithms, and healthcare algorithms [114], [115], [117]. Typically, the audits are ex post, though some can also be completed pre-deployment (e.g. BSR celebrity facial recognition audit was conducted before model release). ORCAA and Eticas also undertake some internal audit methodology development work.

ORCAA conducts audits in order to assess regulatory compliance, performance testing, as well as to measure and mitigate disparate performance [114]. BABL AI conducts audits for bias, risk, and impact assessment [117]. Like Luminos, ORCAA and BABL AI, both offer services to help companies comply legal requirements, for example, in the US, the New York Local Law 144 [101], which requires annual audits of AI hiring tools. Eticas similarly conducts audits to ensure security and data protection and to assess issues such as fairness, bias, and model accuracy [115]. The type of concerns and harms these consulting agencies mention in public materials vary. ORCAA, for example, mainly focuses on measuring bias—such as discrimination, race, and gender—and assessing for regulatory compliance [114]. Eticas mainly focuses on assessing algorithmic systems, for example, for functionality and detecting unfair practices towards protected groups [115]. Similarly, the main types of harm BABL AI focuses on include bias, fairness, transparency as well as assessment for standards and privacy compliance [116].

It is difficult to asses how these three consulting agencies might shift power with clarity. However, they target discrimination, fairness, and gender and ethnicity disparities and aspire to lofty goals—BABL AI, for example, seeks to "prioritize human flourishing" [116]. But, like the legal consulting firms, these audits are subordinate to client needs, and those clients include large corporations, startups, and government bodies. Ultimately, these audits primarily serve those entities—for example, by protecting clients from concerns such as organisational and reputational crisis—and help clients build trust with stakeholders [115], [116], [118].

*b) Methodology:* We have sparse information on audit methodologies used within these agencies. From the information we can gather on their web-pages, these agencies sometimes develop internal audit tools. ORCAA, for example uses "Ethical Matrix Framework" [114], while BABL AI has developed a set of criteria that can be used to conduct bias testing in their "process audit" [117]. Other methods include reviewing documents, interacting with stakeholders, and ethics due diligence vetting, the details of which are *not* provided.

*c) Impact:* The impact of these types of audits, more particularly impact as a direct consequence of these three agencies, is not clear, particularly, with Eticas and BABL AI. ORCAA's work has seen some impact (albeit as an indirect influence) on policy on the US White House AI Ethics

Blueprint [119]. ORCAA also conducted an audit of HireVue's early career and campus hire assessment tools [118], [120]. HireVue subsequently declared its intention to make changes to its practices (but only following a formal complaint from the Electronic Privacy Information Center to the U.S. FTC [121]), including halting the use of facial analysis in its tools, but received criticism for misrepresenting ORCAA's analysis and not taking bolder steps [120]. HireVue also commissioned the audit and defined both the scope of evaluation and the extent of its impact—a problem of independence raised frequently in prior work [26], [27], [84], [120].

### C. *Journalism*

Journalists have a long history of investigating and reporting on AI systems [122]. Investigative journalists at the Wall Street Journal, the New York Times, the MIT Technology Review, and many other outlets have unearthed concrete examples of systematic AI harms [123]–[127].

We examined work from two of the most prominent outlets that have conducted extensive AI audits: ProPublica and The Markup. ProPublica [128] carried out a foundational investigation into criminal risk assessment in 2016, that has set precedence for the field of AI auditing [39]. The Markup [129] is a relatively newer outlet focused on data-driven investigations.

*a) Context:* Audits from both The Markup and ProPublica tend to be *external* focusing on specific types of targets. These include, AI systems that are commonly used by large corporations or social media platforms for advertising, hiring, and ranking; government ADS, or public programs with digital sites; as well as content moderation systems and related labor conditions. Both organisations carry out audits with the stated objective of accountability. The Markup's slogan reads, "Big Tech is watching you. We're watching big tech" [129]. The type of harms and concerns both organisations investigate, surface, diagnose, and evaluate include disparities in performance (for instance, along the dimensions of gender, race, ethnicity, age), injustice, discrimination, privacy violations, fraud, and legal compliance breaches. As harm discovery happens through publicly submitted journalistic "tips", the audits are typically ex-post.

*b) Methodology:* Both Propublica and The Markup utilise quantitative statistical analysis in addition to investigative reporting, interviews, and document analysis. The Markup's investigation of Amazon's product ranking system, for example, involved training a model to predict where products would appear based on various factors [130]. Compared to other domains, these outlets engage in an extensive amount of data collection using a variety of custom scraping, data donation, and analysis tools, often built in-house. The Markup's Citizen Browser project, for example, used data donations to investigate discriminatory targeted advertising [131]–[133] on Facebook [134].

*c) Impact:* Of the various domains we have examined, journalistic audits result in the most impactful outcomes. Audits carried out both by The Markup and ProPublica have resulted in subsequent changes in numerous domains including shifting the audit discourse in academia, altering practices in industry (including big tech corporations such as Facebook, Amazon, and Google) [135], and inspiring activism [136]. Audits both from both outlets have also directly inspired legislative and regulatory action [137], [138], including a settlement in which the U.S. Department of Justice required Facebook to stop using a special audience tool for housing ads [139], [140]. Those actions included abolishing deployed tools—such as an algorithm governing liver transplants that favored rich, urban patients [141], [142]. ProPublica's audit of COMPAS set the precedent for algorithmic auditing and remains a canonical work not only for academic research but also as a prime example of algorithmic audit [39].

### D. *Civil society*

Activist and other civil society organisations—such as the Electronic Privacy Information Center (EPIC), Data & Society, and the AI Now Institute—conduct algorithm audits studies as part of their work. We examined work from six prominent organisations: the Electronic Frontier Foundation (EFF) [143], Refugee Law Lab (RLL) [144], The Citizen Lab [145], Migration Tech Monitor (MTM) [146], the Ada Lovelace Institute [147] and the American Civil Liberties Union (ACLU) [148].

*a) Context:* These organisations and institutions primarily conduct case studies, while the Ada Lovelace Institute in particular, also conducts meta-commentary work, especially on audit methods. Though the audits are consistently *external*, the harms and concerns these civil society audits aim to surface, diagnose, and mitigate vary. Diagnosing and mitigating security vulnerabilities, spying technology, and illegal surveillance are some of the main focuses for EFF. Similarly, the Citizen Lab audits are driven by overarching goals such as investigating and exposing security vulnerabilities and defending free speech online [145]. RLL and MTM primarily focus on investigative and advocacy work around the questions of transparency and the impacts of technology on refugees [144], [149]. MTM particularly aims to document, map, and monitor migration technology and dismantle and destabilise hierarchical power structures [149]. Audits at Ada Lovelace Institute are often driven by the objective of policy change with the primary focus of diagnosis and remedy for concerns such as privacy, transparency, participation, and disparate impact. Accountability through legal action, litigation, and legal objectives are central to ACLU.

The audits we examined target mostly ADS in government services, online platforms, and—more than any other domain—surveillance tech. RLL in particular targets technology such as lie detectors and border patrolling drones [144]. The Citizen Lab focuses on surveillance and biometric technologies (e.g. for facial recognition), hardware (such as phones and other devices used by politicians and activists), apps, and code [145]. Biometric data, digital ID systems, surveillance drones, facial recognition, iris scan data, algorithmic motion detectors, ankle monitors, GPS tags, AI powered satellites as well as border surveillance vendors themselves are

the central audit targets for MTM [146]. The Ada Lovelace Institute also has a diverse target of audit, with a general focus on biometric data and healthcare in particular [147]. The ACLU similarly targets facial recognition technology, medical algorithms, welfare algorithms, insurance pricing algorithms, and redlining algorithms as well as advertising algorithms on platforms such as Facebook [148].

*b) Methodology:* Civil society audits utilize the most diverse audit methods. EFF, for example, uses policy analysis, grass-root activism, and technology development. Some of the methods used by RLL include interviews with refugees, film making, data collection (from the Immigration and refugee board through Access to Information request, for instance) and analysis, interactive visualisation of data, and documentation of issues such as deportation and refused refugee claims. The Citizen Lab audit methods include in-house developed tools, interviews, and forensic analysis of devices. Similar to RLL, MTM also uses methods such as interviews with people crossing borders, photography, investigative analysis, as well as documenting and archiving migration tech. The main audit methods for Ada Lovelace Institute are participatory methods, including opinion polling, as well as policy research and development. The ACLU often uses quantitative evaluation of, for example operational systems, analysis of data acquired through privileged access, public record requests, and scraping.

*c) Impact:* Similar to journalistic investigations, civil society audits often result in significant impact, often through legal action. These include nuanced and difficult to measure impacts such as shifting public attitude towards surveillance technology or drawing public and media attention towards harmful or controversial tech, as well as more concrete outcomes such as moratoriums and abolition. The EFF, for example, won a student's civil lawsuit against the exam surveillance company Proctorio [150]. In *K.W. v. Armstrong*, the ACLU obtained an injunction stopping algorithmically-determined welfare cuts targeting individuals with developmental disabilities in Idaho [151]. Civil society organisations often represent those harmed by AI systems and in doing so, shift power from the most to the least powerful.

### *E. **Government***

Several government agencies, especially in Europe, have begun to engage in AI audits. The EU's recent Digital Services Act and an AI hiring bill passed in New York City [101], for example, both require some form independent audit. We looked at two government organizations, one with an initially prominent role in AI auditing in the UK, the Information Commissioner's Office (ICO) [152], and another in the U.S., the National Institute for Standards and Technology (NIST) [153].

*a) Context:* ICO conducts audits with the broader objective of upholding information rights and enforcing legal requirements [154], [155]. NIST, on the other hand, is primarily focused on establishing standards and best practice principles for ensuring the development and deployment of socially responsible algorithmic systems [155], [156]. ICO conducts audits on a case by case basis, often published as a report. Comparatively, NIST tends to produce meta-commentary on audits. The central goal for engaging in audit practices for ICO is primarily to investigate and enforce regulations, while NIST lacks such authoritative power and is mainly focused on establishing standards.

The type of audits regulators carry out tend to be procedural rather than substantive. Audits are carried out with specific objectives which often involve ensuring that a given organisation, institution or corporation is in compliance with established standards or legal requirements. While some organisations like ICO, for example, have the authority to assess and enforce regulatory compliance, others, organisations like NIST, focus on establishing best practice standards and principles that tech developers and vendors are only encouraged to follow. Guidelines from both ICO [155] and NIST [157] tend to apply to informing and standardizing practice amongst *internal* audit actors. As these guidelines explicitly mention model development interventions such as fairness or explainability mitigation strategies, it is implied that they apply to ex ante and in media res audits, as well as ex post audits.

The main audit targets for ICO are data governance practices within public and private companies, public authorities, and government departments, especially those considered to have major impacts [155], [158]. Organisation can request to be audited but also the ICO audits organisations that the Commissioner believes require audits, that is organisations and corporations with major impacts on society. The ICO audits are a mechanism to check for compliance with data protection legislation, how personal data is managed, data sharing agreements with third parties, and security measures, amongst other things. NIST, on the other hand, focuses on artefacts such as facial recognition technology and "general" AI as the main target for audit with the aim of assessing harms such as racial disparities, basic functionality, privacy harms, as well as physical safety of these systems [159], [160]. To the degree that government actors participate in the audit process themselves, they tend to execute *external* ex-post audits, separate from formal corporate participation.

*b) Methodology:* The ICO has developed a set of definitions of what an audit consists of as well as setting out mechanisms for assessing compliance in accordance with regulations based on four assurance ratings: high, reasonable, limited, and very limited assurance. Following an audit, ICO makes recommendations on how to improve on the audit result. The main method used to conduct audits for ICO include examining documents, testing, and interviews with key personnel [155]. For NIST, the main audit methods include data trust, accuracy evaluation, and benchmarking [156], [161].

*c) Impact:* ICO has issued several monetary penalties for data protection legislation breaches, including a fine of £12.7 million to TikTok for misusing children's data [154] and £17 million fine to Clearview AI inc [162], in addition to issuing a notice to stop further processing of the personal data of people in the UK and a request to delete such data. NIST has made significant impact in the U.S. regulatory conversation with its AI Risk Management Framework (RMF), though it has not

conducted any audits with specific impacts and has no power to enforce these guidelines [73].

## VI. Discussion

While many of the academic studies we reviewed formed the foundation for important methods and topics for AI auditing, they also often lacked in achieving comparable levels of impact to the audit work we analyzed in journalism, civil society, government, and industry. In this section, we outline several practical takeaways from our analysis for researchers, policymakers, and practitioners. In particular, we identify several ways that audit work outside academia could serve as a guide for more impactful audit studies within academia: first, by expanding the aims of audit work beyond evaluation, and towards accountability; second, by encouraging more explicit and forward-looking engagement with often excluded stakeholders, acknowledging power asymmetries [63], and fostering mechanisms for collective action and ecological change; and third, by offering specific improvements for audit methodology, including specificity, practitioner diversity, and interdisciplinary collaboration.

*1) **Power asymmetries & auditor-stakeholder relationships**:* One of the most influential factors in determining an audit's impact was revealed in the power analysis of how the auditors interacted with various stakeholders—including impacted populations, the audit target and others. Subsequently, methodological innovation and evaluation of AI systems without considerations of structural factors—such as the uneven distribution of power, control, benefit and harm—is unlikely to result in significant impact. Future research could further explore tools and strategies to map out, define and foster auditor relationships with various stakeholders. Similarly, policy-makers and other authorities could step in to empower auditors – for example, by incorporating audit results into more consequential repercussions for companies or open-sourcing, publishing, disseminating, and reviewing submitted audit results.

*2) **Prioritizing audit execution stages beyond evaluation**:* Framing details such as the selection of audit targets, named motivations, types of harms investigated and the scope of audit goals are more likely to determine the effectiveness of an audit's outcome than the details of audit execution. Most of the academic work we reviewed focused on the process of evaluating AI systems for bias, fairness, or disparate impacts. Conversely, these studies rarely focused on other stages of auditing crucial to accountability in non-academic work, such as discovering harms, communicating audit results, or organizing non-technical interventions and collective action. As detailed in Section V, these factors are critical in influencing audit success outside of the auditor's control [26], [27]. For example, some auditors, such as regulators and law firms, have the legal authority to demand access. Meanwhile, for others, this degree of internal visibility is rare, requiring the auditors to rely on proxy evaluations and open up their audit results to critique, denial and retaliation from the audit targets.

Even as policymakers work to install legal protections for external auditors and direct resources towards harms discovery and audit reporting, future research should investigate the under-studied aspects of the AI auditing process and practitioners should prioritize these aspects of auditing in addition to evaluation.

*3) **Expanding audit scope beyond the product, model or algorithm**:* In our survey, we use the ecosystem audit classification to describe studies that consider the entire AI pipeline in a holistic way— communities and socio-technical environments (digital or physical) defining or impacted by critical components to an AI system's operation—(§IV-A). While these ecosystem audits exemplify many of the elements of the socio-technical auditing advocated for in various academic critiques [69], [83], [163], most academic audit studies we found extremely rarely involved a comprehensive analysis of involved stakeholders and any holistic view of multiple interacting AI systems. Future work should continue to incorporate and expand broader and holistic perspectives in order to craft a richer account of algorithmic systems and their impacts.

*4) **Impact increases with specificity**:* Even as audit studies broaden their scope to include the ecosystem surrounding and defining an AI system, most audits need to be specific in order to be more impactful. As has been previously discussed in the aftermath of the Gender Shades audit Raji and Buolamwini [38], naming a specific audit target, advocacy objectives and intended target responses ensures that the audit speaks to more specific demands from the audit target. This makes that target more likely to to respond to the audit and informs advocates on precise policy demands. This strategy was seen in many civic audits, such as the ACLU, the Markup and regulators like ICO, where the scope and demands tied to their investigations are specific. When audit scope was too diffused, the audit seemed to hold less effectiveness for accountability. ORCAA's Hirevue audit, for example, was criticized for being too high level to inform any specific allegations or demands for remedy [164].

Constraining audit objectives and scope has benefits as well as limitations. Too narrow a scope may obfuscate broader systematic factors and make it even more difficult to advocate for more structural changes beyond the scope of the examined audit target [38], [165]. Therefore, future work should aim to be highly specific without losing focus on holistic and ecological observations.

*5) **Utilising a wider range of audit methodologies**:* Within CS and CS-adjacent academic venues, quantitative methodological approaches to AI auditing were often front and center. These methods are evident in the range of fairness metrics developed and debated within these communities [166]. Existing critiques already articulate the limitations of these approaches, in terms of its abstraction [167], legal incompatibilities [168], [169] and lack of connection to substantive outcomes [170]. It is particularly noteworthy that in many non-academic domains, the methods employed are far from what is being explored or commonly discussed in academia, including a range of qualitative methodological approaches such

as investigative reporting, document review, and stakeholder consultation. Granted, these other institutions often possess different skill sets and operate within different constraints. However, especially for academics hoping to connect their work to the practice of other audit practitioner communities, future AI audit research should explore how extra-disciplinary methods and tactics like these—especially those with a proven record of impact—could bolster academic audit work.

*6) **Appreciating the diversity of audit practitioners**:* Different audit communities possess different strengths and weaknesses—civil society auditors, for example, typically have mature and well-developed methods around audit communication and advocacy, but at times may lack the technical capacity to execute more complex quantitative analyses. On the other hand, in academia, auditors might be well-equipped for innovating technically but struggled with public and other stakeholder engagement.

Our survey supports prior findings [26] that audit practitioners within a given domain are not a monolith. For instance, there are law firms that provide audits as a service on behalf of audit targets while others execute civic audits on behalf of, and in collaboration with impacted populations. Similarly, we found a wide range of technical competence within domains. Audits from the journalism organization the Markup, for instance, were often more thorough and reproducible than some academic studies. Similarly, auditors can have differing motivations and goals, which deeply informs their practice [171].

Future research on the practice of AI auditing should still consider the strengths and weaknesses of different auditing institutions. These findings are informative not only for improving audit practice but also 1) facilitating avenues for collaboration with auditors outside academia and 2) informing a growing set of enacted and proposed legislation requiring audits for AI systems [101], [172]–[175].

*7) **Timing and auditor type are not the primary determining factors of audit impact**:* The distinction of *when* the audit should ideally be conducted (i.e. ex ante, in media res, ex post) has recently been the subject of some policy debate [176], [177]. As expected, those with increased access (i.e. internal auditors) typically have the opportunity to execute audits pre-deployment, while others (i.e. external auditors) are often restricted to conducting audits post-deployment. However, when it comes to translating audit results to accountability, both scenarios come with distinct challenges and it seems that other contextual factors outside of auditor type or timing, including audit goal, target, design, communication and scope, can ultimately be much more meaningful.

For example, despite increased and early access, internal auditors may still struggle to convince key organizational stakeholders to act on audit results [178]–[180]. Furthermore, such auditors can become vulnerable to internal corporate retaliation, and corporate censorship [181] as well as being mired in conflict of interest [182]. All of this can interfere with the auditors' ability to externally communicate results, or force auditors to scope results too narrowly to be meaningfully. In policy, internal auditors are typically those designated to carry out mandatory corporate audit requirements (e.g. the "independent auditors" in Article 37 of the Digital Services Act). The particular challenges of internal auditors indicate that requirements for external reporting, product standards and auditor conduct standards are especially important in such contexts.

On the other hand, external auditors are free to set their scope and communicate results publicly. However, they typically face issues around external legitimacy and visibility that can make them easy to ignore completely [38]. The serious issues they face around auditor capacity and information access are further hindrances to audit quality and thus impact [27], and the public-facing nature of their advocacy makes them even more vulnerable to corporate retaliation. External auditors are typically designated as voluntary "investigators" or "researchers", rather than actual auditors (e.g. "vetted researchers" in Article 40 of the Digital Services Act), and are often mentioned in data access or safe harbor clauses, revealing the importance of such policy interventions in addressing their particular challenges.

*8) **Recognizing the limits of audits**:* Some societal impacts of AI are not amenable to audits. What can be evaluated and audited depends on numerous factors including normative and pragmatic considerations as well as what is prioritised and anticipated, and whether harms and risks are known or anticipated [183]. As there are several serious, although gradual, and pernicious harms emanating from AI – for instance, the chilling effect of surveillance, or corporate power concentration [184] – it is clear that not all pressing issues regarding the technology can be addressed with audits. Although the ecological audits we have introduced in this work are an encouraging first step towards significant structural change in some of these cases, other, likely moral or rights-based arguments are required for this kind of advocacy. Thus, it is clear that audits are just one element of a necessarily broader set of AI accountability strategies. We caution against any approaches that treat audits as any kind of all-encompassing solution to technological ills.

## VII. Conclusion

Audits have become widely popular in the AI field as an aspirational accountability measure to inform our decision-making and regulation of AI deployments. However, lessons from the broader range of AI audit practitioners beyond academia reveal that there is still much to learn in order for these audit studies to operate as actual consequential judgements. Future work will hopefully continue to investigate this relationship between contextual factors, audit design, audit execution and the underlying shared vision of audit practitioners across all domains: a substantively meaningful path towards AI accountability.

## References


[1] I. D. Raji, I. E. Kumar, A. Horowitz, and A. Selbst, "The fallacy of ai functionality," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 959–972.

[2] Z. Obermeyer and S. Mullainathan, "Dissecting racial bias in an algorithm that guides health decisions for 70 million people," in *Proceedings of the conference on fairness, accountability, and transparency*, 2019, pp. 89–89.

[3] C. H. Chu, R. Nyrup, K. Leslie, *et al.*, "Digital ageism: Challenges and opportunities in artificial intelligence for older adults," *The Gerontologist*, vol. 62, no. 7, pp. 947–955, 2022.

[4] B. Imana, A. Korolova, and J. Heidemann, "Auditing for discrimination in algorithms delivering job ads," in *Proceedings of the web conference 2021*, 2021, pp. 3767–3778.

[5] A. S. Luccioni, C. Akiki, M. Mitchell, and Y. Jernite, "Stable bias: Analyzing societal representations in diffusion models," *arXiv preprint arXiv:2303.11408*, 2023.

[6] P. Barlas, K. Kyriakou, O. Guest, S. Kleanthous, and J. Otterbacher, "To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off," *Proceedings of the ACM on Human-Computer Interaction*, vol. 4, no. CSCW3, pp. 1–31, 2021.

[7] M. K. Scheuerman, J. M. Paul, and J. R. Brubaker, "How computers see gender: An evaluation of gender classification in commercial facial analysis services," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–33, 2019.

[8] I. Kilovaty, "Legally cognizable manipulation," *Berkeley Tech. LJ*, vol. 34, p. 449, 2019.

[9] C. Cadwalladr and E. Graham-Harrison, "Revealed: 50 million facebook profiles harvested for cambridge analytica in major data breach," *The guardian*, vol. 17, no. 1, p. 22, 2018.

[10] F. Pasquale, *The black box society: The secret algorithms that control money and information*. Harvard University Press, 2015.

[11] M. Power, *The Audit Society: Rituals of Verification*. Oxford, New York: Oxford University Press, Oct. 28, 1999, 200 pp., ISBN: 978-0-19-829603-4.

[12] M. Strathern, *Audit cultures: Anthropological studies in accountability, ethics, and the academy*. Psychology Press, 2000.

[13] K. Reichborn-Kjennerud and S. I. Vabo, "Performance audit as a contributor to change and improvement in public administration," *Evaluation*, vol. 23, no. 1, pp. 6–23, 2017.

[14] C. Sandvig, K. Hamilton, K. Karahalios, and C. Langbort, "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms," in *Data and Discrimination: Converting Critical Concerns into Productive: A Preconference at the 64th Annual Meeting of the International Communication Association*, Seattle, WA, 2014, p. 23.

[15] F. Cherry and M. Bendick, "Making it count: Discrimination auditing and the activist scholar tradition," *Audit studies: Behind the scenes with theory, method, and nuance*, pp. 45–62, 2018.

[16] B. Vecchione, S. Barocas, and K. Levy, "Algorithmic Auditing and Social Justice: Lessons from the History of Audit Studies," Sep. 14, 2021. DOI: 10.1145/3465416.3483294. arXiv: 2109.06974 [cs]. [Online]. Available: http://arxiv.org/abs/2109.06974 (visited on 10/08/2021).

[17] N. Carlini, J. Hayes, M. Nasr, *et al.*, "Extracting training data from diffusion models," in *32nd USENIX Security Symposium (USENIX Security 23)*, 2023, pp. 5253–5270.

[18] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-ai decision making: A survey of empirical studies," *arXiv preprint arXiv:2112.11471*, 2021.

[19] S. Rivera, X. Liu, A. Chan, A. Denniston, and M. Calvert, *Guidelines for clinical trial protocols for interventions involving artificial intelligence: The spirit-ai extension. bmj 370*, 2020.

[20] I. D. Raji, A. Smart, R. N. White, *et al.*, "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*, ser. FAT* '20, New York, NY, USA: Association for Computing Machinery, Jan. 27, 2020, pp. 33–44, ISBN: 978-1-4503-6936-7. DOI: 10.1145/3351095.3372873. [Online]. Available: https://dl.acm.org/doi/10.1145/3351095.3372873 (visited on 09/14/2023).

[21] M. Wieringa, "What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 1–18.

[22] I. D. Raji, S. C. CHOCK, and D. BUOLAMWINI, "Change from the outside: Towards credible third-party audits of ai systems," *MISSING LINKS IN AI GOVERNANCE*, p. 5, 2023.

[23] S. Brown, J. Davidovic, and A. Hasan, "The algorithm audit: Scoring the algorithms that score us," *Big Data & Society*, vol. 8, no. 1, p. 2 053 951 720 983 865, 2021.

[24] Ada Lovelace Institute, "Examining the Black Box," Ada Lovelace Institute, Apr. 2020. [Online]. Available: https://www.adalovelaceinstitute.org/report/examining-the-black-box-tools-for-assessing-algorithmic-systems/ (visited on 10/10/2023).

[25] Ada Lovelace Institute, "Technical methods for regulatory inspection of algorithmic systems," Ada Lovelace

Institute, Dec. 2021. [Online]. Available: https://www.adalovelaceinstitute . org / report / technical - methods - regulatory-inspection/ (visited on 10/10/2023).

[26] S. Costanza-Chock, I. D. Raji, and J. Buolamwini, "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*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, Jun. 20, 2022, pp. 1571–1583, ISBN: 978-1-4503-9352-2. DOI: 10 . 1145 / 3531146 . 3533213. [Online]. Available: https://doi.org/10.1145/3531146.3533213 (visited on 09/14/2023).

[27] I. D. Raji, P. Xu, C. Honigsberg, and D. E. Ho. "Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance." arXiv: 2206 . 04737 [cs]. (Jun. 9, 2022), [Online]. Available: http://arxiv.org/abs/2206.04737 (visited on 08/04/2022), preprint.

[28] A. D. Selbst, "An institutional view of algorithmic impact," *Harvard Journal of Law & Technology*, vol. 35, no. 1, 2021.

[29] J. Bandy, "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits," *Proceedings of the ACM on Human-Computer Interaction*, vol. 5, 74:1–74:34, CSCW1 Apr. 22, 2021. DOI: 10.1145/3449148. [Online]. Available: https://doi.org/10.1145/3449148 (visited on 09/14/2023).

[30] R. Steed and A. Caliskan, "Image representations learned with unsupervised pre-training contain human-like biases," in *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 2021, pp. 701–713.

[31] M. Heikkilä, *How it feels to be sexually objectified by an ai*, 2022.

[32] T. Feathers, *Proctorio is using racist algorithms to detect faces*, 2021.

[33] H. Touvron, T. Lavril, G. Izacard, *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[34] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, IEEE, 2021, pp. 1536–1546.

[35] O. Solon, "Facial recognition's 'dirty little secret': Millions of online photos scraped without consent," *NBC News*, vol. 12, 2019.

[36] A. Reisner, "Revealed: The authors whose pirated books are powering generative ai," *The Atlantic*, 2023.

[37] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[38] I. D. Raji and J. Buolamwini, "Actionable Auditing Revisited: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," *Communications of the ACM*, vol. 66, no. 1, pp. 101–108, Dec. 20, 2022, ISSN: 0001-0782. DOI: 10.1145/3571151. [Online]. Available: https://dl.acm.org/doi/10.1145/3571151 (visited on 09/14/2023).

[39] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias," *ProPublica*, May 23, 2016. [Online]. Available: https : / / www . propublica . org / article / machine-bias-risk-assessments-in-criminal-sentencing (visited on 09/14/2023).

[40] A. Chouldechova, D. Benavides-Prado, O. Fialko, and R. Vaithianathan, "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions," in *Conference on Fairness, Accountability and Transparency*, PMLR, 2018, pp. 134–148.

[41] L. Chen, A. Mislove, and C. Wilson, "Peeking beneath the hood of uber," in *Proceedings of the 2015 internet measurement conference*, 2015, pp. 495–508.

[42] L. Chen, A. Mislove, and C. Wilson, "An empirical analysis of algorithmic pricing on amazon marketplace," in *Proceedings of the 25th international conference on World Wide Web*, 2016, pp. 1339–1349.

[43] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, S. A. Friedler and C. Wilson, Eds., ser. Proceedings of Machine Learning Research, vol. 81, New York, NY, USA: PMLR, Jan. 2018, pp. 77–91. [Online]. Available: http://proceedings.mlr.press/v81/buolamwini18a.html.

[44] A. Koenecke, A. Nam, E. Lake, *et al.*, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.

[45] A. Mandal, S. Little, and S. Leavy, "Gender bias in multimodal models: A transnational feminist approach considering geographical region and culture," *arXiv preprint arXiv:2309.04997*, 2023.

[46] C. Wilson, A. Ghosh, S. Jiang, *et al.*, "Building and auditing fair algorithms: A case study in candidate screening," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 666–677.

[47] K. Lum and W. Isaac, "To predict and serve?" *Significance*, vol. 13, no. 5, pp. 14–19, 2016.

[48] A. Koenecke, E. Giannella, R. Willer, and S. Goel, "Popular support for balancing equity and efficiency in resource allocation: A case study in online advertising to increase welfare program awareness," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 17, 2023, pp. 494–506.

[49] P. Sapiezynski, A. Ghosh, L. Kaplan, A. Rieke, and A. Mislove, "Algorithms that" don't see color" measuring biases in lookalike and special ad audiences," in *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 2022, pp. 609–616.

[50] S. U. Noble, "Algorithms of oppression," in *Algorithms of oppression*, New York university press, 2018.

[51] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[52] S. McGregor, "Preventing repeated real world ai failures by cataloging incidents: The ai incident database," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, 2021, pp. 15 458–15 463.

[53] K. Fink, "Opening the government's black boxes: Freedom of information and algorithmic accountability," *Information, Communication & Society*, vol. 21, no. 10, pp. 1453–1471, 2018.

[54] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of ai ethics guidelines," *Nature machine intelligence*, vol. 1, no. 9, pp. 389–399, 2019.

[55] I. Ajunwa, S. Friedler, C. E. Scheidegger, and S. Venkatasubramanian, "Hiring by algorithm: Predicting and preventing disparate impact," *Available at SSRN*, 2016.

[56] J. Bandy, "Problematic machine behavior: A systematic literature review of algorithm audits," *Proceedings of the acm on human-computer interaction*, vol. 5, no. CSCW1, pp. 1–34, 2021.

[57] P. Krafft, M. Young, M. Katell, *et al.*, "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*, 2021, pp. 772–781.

[58] S. Jiang, R. E. Robertson, and C. Wilson, "Bias Misperceived:The Role of Partisanship and Misinformation in YouTube Comment Moderation," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 278–289, Jul. 6, 2019, ISSN: 2334-0770. DOI: 10.1609/icwsm.v13i01.3229. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/3229 (visited on 10/09/2023).

[59] E. Black, H. Elzayn, A. Chouldechova, J. Goldin, and D. Ho, "Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea: ACM, Jun. 21, 2022, pp. 1479–1503, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533204. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533204 (visited on 10/09/2023).

[60] A. Abid, M. Farooqi, and J. Zou, "Persistent Anti-Muslim Bias in Large Language Models," in *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, ser. AIES '21, New York, NY, USA: Association for Computing Machinery, Jul. 30, 2021, pp. 298–306, ISBN: 978-1-4503-8473-5. DOI: 10.1145/3461702.3462624. [Online]. Available: https://doi.org/10.1145/3461702.3462624 (visited on 10/09/2023).

[61] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Jan. 2021, pp. 1536–1546. DOI: 10.1109/WACV48630.2021.00158.

[62] A. Brown, A. Chouldechova, E. Putnam-Hornstein, A. Tobin, and R. Vaithianathan, "Toward Algorithmic Accountability in Public Services: A Qualitative Study of Affected Community Perspectives on Algorithmic Decision-making in Child Welfare Services," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, ser. CHI '19, New York, NY, USA: Association for Computing Machinery, May 2, 2019, pp. 1–12, ISBN: 978-1-4503-5970-2. DOI: 10.1145/3290605.3300271. [Online]. Available: https://dl.acm.org/doi/10.1145/3290605.3300271 (visited on 10/09/2023).

[63] P. Kalluri *et al.*, "Don't ask if artificial intelligence is good or fair, ask how it shifts power," *Nature*, vol. 583, no. 7815, pp. 169–169, 2020.

[64] J. Metcalf, E. Moss, E. A. Watkins, R. Singh, and M. C. Elish, "Algorithmic Impact Assessments and Accountability: The Co-construction of Impacts," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '21, New York, NY, USA: Association for Computing Machinery, Mar. 1, 2021, pp. 735–746, ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445935. [Online]. Available: https://dl.acm.org/doi/10.1145/3442188.3445935 (visited on 10/10/2023).

[65] A. Birhane, V. U. Prabhu, and E. Kahembwe. "Multimodal datasets: Misogyny, pornography, and malignant stereotypes." arXiv: 2110.01963 [cs]. (Oct. 5, 2021), [Online]. Available: http://arxiv.org/abs/2110.01963 (visited on 10/05/2023), preprint.

[66] M. Bao, A. Zhou, S. Zottola, *et al.*, "It's COMPASlicated: The Messy Relationship between RAI Datasets and Algorithmic Fairness Benchmarks," Jun. 10, 2021. arXiv: 2106.05498 [cs]. [Online]. Available: http://arxiv.org/abs/2106.05498 (visited on 09/10/2021).

[67] P. Samuelson, "Generative ai meets copyright," *Science*, vol. 381, no. 6654, pp. 158–161, 2023.

[68] L. Stapleton, M. H. Lee, D. Qing, *et al.*, "Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 1162–1177.

[69] E. Radiya-Dixit and G. Neff, "A Sociotechnical Audit: Assessing Police Use of Facial Recognition," in *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '23, New York, NY, USA: Association for Computing Machinery, Jun. 12, 2023, pp. 1334–1346, ISBN: 9798400701924. DOI: 10.1145/3593013.3594084. [Online]. Available: https://dl.acm.org/doi/10.1145/3593013.3594084 (visited on 06/16/2023).

[70] L. Stapleton, M. H. Lee, D. Qing, *et al.*, "Imagining new futures beyond predictive systems in child welfare: A qualitative study with impacted stakeholders," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, Jun. 20, 2022, pp. 1162–1177, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533177. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533177 (visited on 10/11/2023).

[71] A. Woodruff, S. E. Fox, S. Rousso-Schindler, and J. Warshaw, "A Qualitative Exploration of Perceptions of Algorithmic Fairness," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, ser. CHI '18, New York, NY, USA: Association for Computing Machinery, Apr. 21, 2018, pp. 1–14, ISBN: 978-1-4503-5620-6. DOI: 10.1145/3173574.3174230. [Online]. Available: https://dl.acm.org/doi/10.1145/3173574.3174230 (visited on 10/11/2023).

[72] I. Solaiman, Z. Talat, W. Agnew, *et al.*, "Evaluating the social impact of generative ai systems in systems and society," *arXiv preprint arXiv:2306.05949*, 2023.

[73] E. Tabassi, "AI Risk Management Framework: AI RMF (1.0)," National Institute of Standards and Technology, Gaithersburg, MD, error: NIST AI 100-1, 2023, error: NIST AI 100–1. DOI: 10.6028/NIST.AI.100-1. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf (visited on 09/15/2023).

[74] D. Ramesh, V. Kameswaran, D. Wang, and N. Sambasivan, "How Platform-User Power Relations Shape Algorithmic Accountability: A Case Study of Instant Loan Platforms and Financially Stressed Users in India," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, ser. FAccT '22, New York, NY, USA: Association for Computing Machinery, Jun. 20, 2022, pp. 1917–1928, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533237. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533237 (visited on 10/05/2023).

[75] J. Singh. "Google to prohibit personal loan apps from accessing user photos, contacts," TechCrunch. (Apr. 6, 2023), [Online]. Available: https://techcrunch.com/2023/04/05/google-personal-loan-apps-update/ (visited on 10/05/2023).

[76] A. Torralba, R. Fergus, and B. Freeman. "80 Million Tiny Images." (Jun. 29, 2020), [Online]. Available: https://groups.csail.mit.edu/vision/TinyImages/ (visited on 10/12/2023).

[77] M. Murgia and M. Harlow, "Who's using your face? The ugly truth about facial recognition," *Financial TimesFT Magazine*, Sep. 18, 2019. [Online]. Available: https://www.ft.com/content/cf19b956-60a2-11e9-b285-3acd5d43599e (visited on 10/05/2023).

[78] M. Murgia, "Microsoft quietly deletes largest public face recognition data set," *Financial TimesMicrosoft Corp*, Jun. 6, 2019. [Online]. Available: https://www.ft.com/content/7d3e0d6a-87a0-11e9-a028-86cea8523dc2 (visited on 10/12/2023).

[79] H. Suresh and J. V. Guttag, "A framework for understanding unintended consequences of machine learning," *arXiv preprint arXiv:1901.10002*, vol. 2, no. 8, 2019.

[80] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.

[81] B. Rakova and R. Dobbe, "Algorithms as social-ecological-technological systems: An environmental justice lens on algorithmic audits," *arXiv preprint arXiv:2305.05733*, 2023.

[82] K. Crawford, *The atlas of AI: Power, politics, and the planetary costs of artificial intelligence*. Yale University Press, 2021.

[83] B. Gansky and S. McDonald, "CounterFAccTual: How FAccT Undermines Its Organizing Principles," in *2022 ACM Conference on Fairness, Accountability, and Transparency*, Seoul Republic of Korea: ACM, Jun. 21, 2022, pp. 1982–1992, ISBN: 978-1-4503-9352-2. DOI: 10.1145/3531146.3533241. [Online]. Available: https://dl.acm.org/doi/10.1145/3531146.3533241 (visited on 06/21/2022).

[84] E. P. Goodman and J. Trehu. "AI Audit Washing and Accountability." (Sep. 22, 2022), [Online]. Available: https://papers.ssrn.com/abstract=4227350 (visited on 03/10/2023), preprint.

[85] M. Sloane, E. Moss, and R. Chowdhury, "A Silicon Valley love triangle: Hiring algorithms, pseudo-science, and the quest for auditability," *Patterns (New York, N.Y.)*, vol. 3, no. 2, p. 100425, Feb. 11, 2022, ISSN: 2666-3899. DOI: 10.1016/j.patter.2021.100425. pmid: 35199067.

[86] Luminos.Law. "Public Resources," Luminos.Law — A Boutique Law Firm Focused on AI and Analytics. (), [Online]. Available: https://luminos.law/resources (visited on 10/11/2023).

[87] Foxglove. "Who We Are," Foxglove. (), [Online]. Available: https://www.foxglove.org.uk/who-we-are/ (visited on 10/11/2023).

[88] AWO. "AWO." (2023), [Online]. Available: https://awo.agency/ (visited on 10/11/2023).

[89] Luminos.Law. "Clients," Luminos.Law — A Boutique Law Firm Focused on AI and Analytics. (), [Online]. Available: https://luminos.law/our-clients (visited on 10/11/2023).

[90] AWO. "Services." (), [Online]. Available: https://awo.agency/ (visited on 10/11/2023).

[91] Luminos.Law. "Our Work," luminos.law. (), [Online]. Available: https://luminos.law/our-work (visited on 10/12/2023).

[92] AWO. "Blog," awo.agency. (), [Online]. Available: https://awo.agency/ (visited on 10/12/2023).

[93] M. Dark. "Home Office says it will abandon its racist visa algorithm - after we sued them," Foxglove. (Aug. 4, 2020), [Online]. Available: https://www.foxglove.org.uk/2020/08/04/home-office-says-it-will-abandon-its-racist-visa-algorithm-after-we-sued-them/ (visited on 10/12/2023).

[94] M. Dark. "We put a stop to the A Level grading algorithm!" Foxglove. (Aug. 17, 2020), [Online]. Available: https://www.foxglove.org.uk/2020/08/17/we-put-a-stop-to-the-a-level-grading-algorithm/ (visited on 10/12/2023).

[95] Foxglove. "Areas of Work," foxglove.org.uk. (), [Online]. Available: https://www.foxglove.org.uk/who-we-are/areas-of-work/ (visited on 10/12/2023).

[96] T. Hegarty. "The government has scrapped the deadline for the NHS Data Grab," Foxglove. (Jul. 22, 2021), [Online]. Available: https://www.foxglove.org.uk/2021/07/22/the-government-has-scrapped-the-deadline-for-the-nhs-data-grab/ (visited on 10/12/2023).

[97] B. Tau, "Banning TikTok in the U.S. Is Easier Said Than Done," *Wall Street JournalBusiness*, Mar. 25, 2023, ISSN: 0099-9660. [Online]. Available: https://www.wsj.com/articles/tiktok-ban-legal-explained-bbeb21c2 (visited on 10/12/2023).

[98] S. Lynch, "A New Anti-Bias A.I. Hiring Law Is Now in Effect. How to Know If You're in Compliance," *Inc.com*, Jul. 19, 2023. [Online]. Available: https://www.inc.com/sarah-lynch/new-anti-bias-ai-hiring-law-now-in-effect-how-to-comply.html (visited on 10/12/2023).

[99] AWO. "AWO analysis shows gaps in effective protection from AI harms," Articles. (Jul. 17, 2023), [Online]. Available: https://awo.agency/ (visited on 10/11/2023).

[100] Luminos.Law. "Microwave," Luminos. (), [Online]. Available: https://www.luminos.ai/microwave (visited on 10/11/2023).

[101] *A Local Law to amend the administrative code of the city of New York, in relation to automated employment decision tools*, in collab. with L. A. Cumbo, A. Ampry-Samuel, H. K. Rosenthal, *et al.*, Dec. 11, 2021. [Online]. Available: https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page (visited on 09/15/2023).

[102] A. Brennen, R. Ashley, R. Calix, J. J. Ben-Joseph, G. Sieniawski, and M. Gogia, "AI Assurance Audit of RoBERTa, an Open source, Pretrained Large Language Model," IQT Labs, Dec. 2022. [Online]. Available: https://assets.iqt.org/pdfs/IQTLabs_RoBERTaAudit_Dec2022_final.pdf/web/viewer.html (visited on 10/11/2023).

[103] R. A. Calix, J. Ben-Joseph, N. Lopatina, *et al.*, "Saisiyat Is Where It Is At! Insights Into Backdoors And Debiasing Of Cross Lingual Transformers For Named Entity Recognition," in *2022 IEEE International Conference on Big Data (Big Data)*, Dec. 2022, pp. 2940–2949. DOI: 10.1109/BigData55660.2022.10020403. [Online]. Available: https://ieeexplore.ieee.org/document/10020403 (visited on 10/11/2023).

[104] M. Lee, N. Stringer, and Z. Nadir, *Student-level equalities analyses for gcse and a level. ofqual*, 2020.

[105] The Joint Council for the Welfare of Immigrants. "We won! Home Office to stop using racist visa algorithm," Joint Council for the Welfare of Immigrants. (Aug. 4, 2020), [Online]. Available: https://www.jcwi.org.uk/News/we-won-home-office-to-stop-using-racist-visa-algorithm (visited on 10/11/2023).

[106] M. Fitzgerald and C. Crider. "Under pressure, UK government releases NHS COVID data deals with big tech," openDemocracy. (Jun. 5, 2020), [Online]. Available: https://www.opendemocracy.net/en/ournhs/under-pressure-uk-government-releases-nhs-covid-data-deals-big-tech/ (visited on 10/11/2023).

[107] Parity Consulting. "Parity Consulting," Parity Consulting. (), [Online]. Available: https://www.get-parity.com (visited on 10/12/2023).

[108] B. Cassidy, R. Hittner, B. Crowley, Z. Bowman, and J. Fogarty, "An auditor's mindset in an AI-driven world," 2022. [Online]. Available: https://www2.deloitte.com/content/dam/Deloitte/us/Documents/deloitte-analytics/us-ai-institute-auditors-mindset.pdf (visited on 10/12/2023).

[109] K. Buehler, R. Dooley, L. Grennan, and A. Singla. "Identifying and managing your biggest AI risks — McKinsey," Quantum Black: AI by McKinsey. (May 3, 2021), [Online]. Available: https://www.mckinsey.com/capabilities/quantumblack/our-insights/getting-to-know-and-manage-your-biggest-ai-risks (visited on 10/12/2023).

[110] Accenture. "AI ethics & governance," Accenture. (2023), [Online]. Available: https://www.accenture.com/us-en/services/applied-intelligence/ai-ethics-governance (visited on 10/12/2023).

[111] Business for Social Responsibility, "Google Celebrity Recognition API Human Rights Assessment: Executive Summary," Oct. 2019. [Online]. Available: https://www.bsr.org/reports/BSR-Google-CR-API-HRIA-Executive-Summary.pdf (visited on 10/12/2023).

[112] A. Warofka, "An independent assessment of the human rights impact of facebook in myanmar," *Facebook Newsroom, November*, vol. 5, 2018.

[113] Office of Public Affairs. "Justice Department Secures Groundbreaking Settlement Agreement with Meta Platforms, Formerly Known as Facebook, to Resolve Allegations of Discriminatory Advertising," United States Department of Justice. (Jun. 21, 2022), [Online]. Available: https://www.justice.gov/opa/pr/justice-department-secures-groundbreaking-settlement-agreement-meta-platforms-formerly-known (visited on 10/12/2023).

[114] ORCAA. "ORCAA," ORCAA. (Jul. 28, 2023), [Online]. Available: https://orcaarisk.com (visited on 10/11/2023).

[115] Eticas. "Algorithmic Audits," eticasconsulting.com. (2022), [Online]. Available: https://eticas.tech/algorithmic-audits (visited on 10/11/2023).

[116] BABL AI. "About Us," babl.ai. (), [Online]. Available: https://babl.ai/about-us/ (visited on 10/11/2023).

[117] BABL AI. "Services." (), [Online]. Available: https://babl.ai/services/ (visited on 10/12/2023).

[118] ORCAA, "Description of Algorithmic Audit: Pre-built Assessments," ORCAA, Dec. 2020. [Online]. Available: https://www.hirevue.com/resources/template/orcaa-report (visited on 10/12/2023).

[119] S. Samuel, "There's something missing from the White House's AI ethics blueprint," *Vox*, Oct. 5, 2022. [Online]. Available: https://www.vox.com/future-perfect/23387228/ai-bill-of-rights-white-house-artificial-intelligence-bias (visited on 10/11/2023).

[120] A. C. Engler, "Independent auditors are struggling to hold AI companies accountable," *Fast Company*, Jan. 26, 2021. [Online]. Available: https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue (visited on 10/12/2023).

[121] The Electronic Privacy Information Center (EPIC), *Complaint and Request for Investigation, Injunction, and Other Relief in the Matter of HireVue, Inc.* Federal Trade Commission, Nov. 6, 2019. [Online]. Available: https://epic.org/wp-content/uploads/privacy/ftc/hirevue/EPIC_FTC_HireVue_Complaint.pdf (visited on 10/12/2023).

[122] N. Diakopoulos, "Algorithmic accountability: Journalistic investigation of computational power structures," *Digital journalism*, vol. 3, no. 3, pp. 398–415, 2015.

[123] K. Hao and D. Seetharaman, "Cleaning Up ChatGPT Takes Heavy Toll on Human Workers," *Wall Street JournalTech*, Jul. 24, 2023, ISSN: 0099-9660. [Online]. Available: https://www.wsj.com/articles/chatgpt-openai-content-abusive-sexually-explicit-harassment-kenya-workers-on-human-workers-cf191483 (visited on 10/12/2023).

[124] K. Hill, "Wrongfully Accused by an Algorithm," *The New York TimesTechnology*, Jun. 24, 2020, ISSN: 0362-4331. [Online]. Available: https://www.nytimes.com/2020/06/24/technology/facial-recognition-arrest.html (visited on 10/12/2023).

[125] K. Hao and H. Swart, "South Africa's private surveillance machine is fueling a digital apartheid," *MIT Technology Review*, Apr. 19, 2022. [Online]. Available: https://www.technologyreview.com/2022/04/19/1049996/south-africa-ai-surveillance-digital-apartheid/ (visited on 10/12/2023).

[126] K. Hao and A. P. Hernández, "How the AI industry profits from catastrophe," *MIT Technology Review*, Apr. 20, 2022. [Online]. Available: https://www.technologyreview.com/2022/04/20/1050392/ai-industry-appen-scale-data-labels/ (visited on 10/12/2023).

[127] M. Heikkilä, "The viral AI avatar app Lensa undressed me—without my consent," *MIT Technology Review*, Dec. 12, 2022. [Online]. Available: https://www.technologyreview.com/2022/12/12/1064751/the-viral-ai-avatar-app-lensa-undressed-me-without-my-consent/ (visited on 12/12/2022).

[128] ProPublica. "Investigative Journalism and News in the Public Interest," propublica.org. (Oct. 11, 2023), [Online]. Available: https://www.propublica.org/ (visited on 10/11/2023).

[129] The Markup. "About Us," themarkup.org. (), [Online]. Available: https://themarkup.org/about (visited on 10/11/2023).

[130] A. Jeffries and L. Yin, "Amazon Puts Its Own "Brands" First Above Better-Rated Products – The Markup," *The Markup*, Oct. 14, 2021. [Online]. Available: https://themarkup.org/amazons-advantage/2021/10/14/amazon-puts-its-own-brands-first-above-better-rated-products (visited on 10/12/2023).

[131] J. Angwin and T. Parris Jr., "Facebook Lets Advertisers Exclude Users by Race," *ProPublica*, Oct. 28, 2016. [Online]. Available: https://www.propublica.org/article/facebook-lets-advertisers-exclude-users-by-race (visited on 10/12/2023).

[132] J. Keegan, "Facebook Got Rid of Racial Ad Categories. Or Did It? – The Markup," *The Markup*, Jul. 9, 2021. [Online]. Available: https://themarkup.org/citizen-browser/2021/07/09/facebook-got-rid-of-racial-ad-categories-or-did-it (visited on 10/12/2023).

[133] C. Faife and A. Ng, "Credit Card Ads Were Targeted by Age, Violating Facebook's Anti-Discrimination Policy – The Markup," *The Markup*, Apr. 29, 2021. [Online]. Available: https://themarkup.org/citizen-browser/2021/04/29/credit-card-ads-were-targeted-by-age-violating-facebooks-anti-discrimination-policy (visited on 10/12/2023).

[134] The Markup. "Citizen Browser," The Markup. (Oct. 25, 2022), [Online]. Available: https://themarkup.org/series/citizen-browser (visited on 10/12/2023).

[135] A. Ng and C. Faife, "Facebook Pledges to Remove Discriminatory Credit and Loan Ads Discovered by The Markup – The Markup," *The Markup*, May 4, 2021. [Online]. Available: https://themarkup.org/citizen-browser/2021/05/04/facebook-pledges-to-remove-discriminatory-credit-and-loan-ads-discovered-by-the-markup (visited on 10/12/2023).

[136] L. Yin, "Citing Markup Investigation, Civil Rights Group Demands Racial Equity Audit at Google – The Markup," *The Markup*, May 4, 2021. [Online]. Available: https://themarkup.org/google-the-giant/2021/05/04/citing-markup-investigation-civil-rights-group-demands-racial-equity-audit-at-google (visited on 10/12/2023).

[137] M. Carollo, "The Markup's Work Cited in Effort to Outlaw Discriminatory Algorithms – The Markup," *The Markup*, Dec. 17, 2021. [Online]. Available: https://themarkup.org/locked-out/2021/12/17/the-markups-work-cited-in-effort-to-outlaw-discriminatory-algorithms (visited on 10/12/2023).

[138] C. Lecher and J. Keegan, "Nevada Lawmakers Introduce Privacy Legislation After Markup Investigation into Vaccine Websites – The Markup," *The Markup*, May 18, 2021. [Online]. Available: https://themarkup.org/blacklight/2021/05/18/nevada-lawmakers-introduce-privacy-legislation-after-markup-investigation-into-vaccine-websites (visited on 10/12/2023).

[139] K. Benner, G. Thrush, and M. Isaac, "Facebook Engages in Housing Discrimination With Its Ad Practices, U.S. Says," *The New York TimesU.S.*, Mar. 28, 2019, ISSN: 0362-4331. [Online]. Available: https://www.nytimes.com/2019/03/28/us/politics/facebook-housing-discrimination.html (visited on 10/12/2023).

[140] L. Feiner, "DOJ settles lawsuit with Facebook over allegedly discriminatory housing advertising," *CNBC*, Jun. 21, 2022. [Online]. Available: https://www.cnbc.com/2022/06/21/doj-settles-with-facebook-over-allegedly-discriminatory-housing-ads.html (visited on 10/12/2023).

[141] M. Carollo, "An Algorithm Decides Who Gets a Liver Transplant. Here Are 5 Things to Know. – The Markup," May 20, 2023. [Online]. Available: https://themarkup.org/hello-world/2023/05/20/an-algorithm-decides-who-gets-a-liver-transplant-here-are-5-things-to-know (visited on 10/12/2023).

[142] M. Carollo and B. Tanen, "Poorer States Suffer Under New Organ Donation Rules, As Livers Go to Waste – The Markup," *The Markup*, Mar. 21, 2023. [Online]. Available: https://themarkup.org/organ-failure/2023/03/21/poorer-states-suffer-under-new-organ-donation-rules-as-livers-go-to-waste (visited on 10/12/2023).

[143] Electronic Frontier Foundation. "About EFF," Electronic Frontier Foundation. (Jul. 10, 2007), [Online]. Available: https://www.eff.org/about (visited on 10/11/2023).

[144] Refugee Law Lab. "Refugee Law Lab," refugeelab.ca. (Jul. 13, 2020), [Online]. Available: https://refugeelab.ca/ (visited on 10/11/2023).

[145] The Citizen Lab. "About the Citizen Lab," citizenlab.ca. (), [Online]. Available: https://citizenlab.ca/about/ (visited on 10/11/2023).

[146] Migration Tech Monitor. "Migration and Technology Monitor," migrationtechmonitor.com. (Feb. 14, 2023), [Online]. Available: https://www.migrationtechmonitor.com (visited on 10/11/2023).

[147] Ada Lovelace Institute. "About," adalovelaceinstitute.org. (2023), [Online]. Available: https://www.adalovelaceinstitute.org/about/ (visited on 10/11/2023).

[148] American Civil Liberties Union. "Home," aclu.org. (), [Online]. Available: https://www.aclu.org/ (visited on 10/11/2023).

[149] Migration Tech Monitor. "About Us," migrationtechmonitor.com. (), [Online]. Available: https://www.migrationtechmonitor.com/about-us (visited on 10/12/2023).

[150] K. Gullo. "EFF Client Erik Johnson and Proctorio Settle Lawsuit Over Bogus DMCA Claims," Electronic Frontier Foundation. (Mar. 25, 2022), [Online]. Available: https://www.eff.org/deeplinks/2022/03/eff-client-eric-johnson-and-proctorio-settle-lawsuit-over-bogus-dmca-claims (visited on 10/11/2023).

[151] ACLU of Idaho. "K.W. v. Armstrong — ACLU of Idaho," ACLU of Idaho. (Jul. 18, 2016), [Online]. Available: https://www.acluidaho.org/en/cases/kw-v-armstrong (visited on 10/12/2023).

[152] Information Commissioner's Office. "Information Commissioner's Office (ICO)," ico.org.uk. (Oct. 6, 2023), [Online]. Available: https://ico.org.uk/ (visited on 10/11/2023).

[153] National institute of Standards and Technology. "National Institute of Standards and Technology," nist.gov. (Oct. 11, 2023), [Online]. Available: https://www.nist.gov/ (visited on 10/11/2023).

[154] Information Commissioner's Office. "ICO fines TikTok £12.7 million for misusing children's data." (May 15, 2023), [Online]. Available: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2023/04/ico-fines-tiktok-127-million-for-misusing-children-s-data/ (visited on 10/11/2023).

[155] Information Commissioner's Office. "Audits and overview reports." (Jan. 5, 2022), [Online]. Available: https://ico.org.uk/action-weve-taken/audits-and-overview-reports/ (visited on 10/12/2023).

[156] National institute of Standards and Technology AIRC Team. "NIST Trustworthy & Responsible AI Resource Center," NIST. (), [Online]. Available: https://airc.nist.gov/home (visited on 10/10/2023).

[157] N. AI, "Artificial intelligence risk management framework (ai rmf 1.0)," 2023.

[158] Information Commissioner's Office, "Annex A: Fairness in the AI Lifecycle," Guidance on the AI auditing framework Draft guidance for consultation, Mar. 2023. [Online]. Available: https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/.

[159] P. Grother, P. Grother, M. Ngan, and K. Hanaoka, *Face recognition vendor test (frvt) part 2: Identification*, 2019.

[160] P. J. Phillips, P. Grother, R. Micheals, D. M. Blackburn, E. Tabassi, and M. Bone, "Face recognition vendor test 2002," in *2003 IEEE International SOI Conference. Proceedings (Cat. No. 03CH37443)*, IEEE, 2003, p. 44.

[161] M. Ngan, P. Grother, and K. Hanaoka, "Ongoing Face Recognition Vendor Test (FRVT) Part 6B: Face recognition accuracy with face masks using post-COVID-19 algorithms," National Institute of Standards and Technology, Nov. 30, 2020. DOI: 10.6028/NIST.IR.8331. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2020/NIST.IR.8331.pdf (visited on 08/23/2023).

[162] Information Commissioner's Office. "ICO issues provisional view to fine Clearview AI Inc over £17 million." (Dec. 1, 2021), [Online]. Available: https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2021/11/ico-issues-provisional-view-to-fine-clearview-ai-inc-over-17-million/ (visited on 10/11/2023).

[163] M. S. Lam, A. Pandit, C. H. Kalicki, R. Gupta, P. Sahoo, and D. Metaxa, "Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising," *Proceedings of the ACM on Human-Computer Interaction*, vol. 7, 360:1–360:37, CSCW2 Oct. 4, 2023. DOI: 10.1145/3610209. [Online]. Available: https://dl.acm.org/doi/10.1145/3610209 (visited on 10/09/2023).

[164] A. C. Engler, "Independent auditors are struggling to hold ai companies accountable," *Fast Company*, 2021.

[165] I. D. Raji, T. Gebru, M. Mitchell, J. Buolamwini, J. Lee, and E. Denton, "Saving face: Investigating the ethical concerns of facial recognition auditing," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 145–151.

[166] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian, "The (im) possibility of fairness: Different value systems require different mechanisms for fair decision making," *Communications of the ACM*, vol. 64, no. 4, pp. 136–143, 2021.

[167] A. D. Selbst, d. boyd, S. Friedler, S. Venkatasubramanian, and J. Vertesi, "Fairness and Abstraction in Sociotechnical Systems," Aug. 2018. [Online]. Available: https://papers.ssrn.com/abstract=3265913.

[168] A. Xiang. "Reconciling Legal and Technical Approaches to Algorithmic Bias." (Jan. 4, 2021), [Online]. Available: https://papers.ssrn.com/abstract=3650635 (visited on 02/20/2023), preprint.

[169] A. Xiang and I. D. Raji, "On the legal compatibility of fairness definitions," *arXiv preprint arXiv:1912.00761*, 2019.

[170] B. Green, "Escaping the impossibility of fairness: From formal to substantive algorithmic fairness," *Philosophy & Technology*, vol. 35, no. 4, p. 90, 2022.

[171] B. Laufer, J. Kleinberg, K. Levy, and H. Nissenbaum. "Strategic Evaluation: Subjects, Evaluators, and Society." arXiv: 2310.03655 [cs]. (Oct. 5, 2023), [Online]. Available: http://arxiv.org/abs/2310.03655 (visited on 10/11/2023), preprint.

[172] *Algorithmic Accountability Act of 2022*, in collab. with Y. D. Clarke, Apr. 11, 2019. [Online]. Available: https://www.congress.gov/117/bills/hr6580/BILLS-117hr6580ih.pdf (visited on 02/07/2022).

[173] A. Lenhart, "Federal AI Legislation: An Analysis of Proposals from the 117th Congress Relevant to Generative AI tools," Institute for Data, Democracy & Politics, George Washington University, Jun. 2023. [Online]. Available: https://iddp.gwu.edu/sites/g/files/zaxdzs5791/files/2023-06/federal_ai_legislation_v3.pdf (visited on 09/15/2023).

[174] B. Perrigo, "California Bill Proposes Regulating AI at State Level," *Time*, Sep. 13, 2023. [Online]. Available: https://time.com/6313588/california-ai-regulation-bill/ (visited on 09/15/2023).

[175] *Stop Discrimination by Algorithms Act of 2021*, in collab. with P. Mendelson, Dec. 9, 2021. [Online]. Available: https://lims.dccouncil.gov/Legislation/B24-0558 (visited on 09/15/2023).

[176] E. Accountable Tech AI Now Institute, *Zero trust ai governance*, 2023. [Online]. Available: https://ainowinstitute.org/publication/zero-trust-ai-governance.

[177] A. L. Institute, *Fda-style oversight for foundation models*, forthcoming.

[178] T. Phan, J. Goldenfein, M. Mann, and D. Kuch, "Economies of virtue: The circulation of 'ethics' in big tech," *Science as culture*, vol. 31, no. 1, pp. 121–135, 2022.

[179] W. Boag, H. Suresh, B. Lepe, and C. D'Ignazio, "Tech worker organizing for power and accountability," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 452–463.

[180] D. G. Widder, D. Zhen, L. Dabbish, and J. Herbsleb, "It's about power: What ethical concerns do software engineers have, and what do they (feel they can) do about them?" In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 2023, pp. 467–479.

[181] M. Ryan, E. Christodoulou, J. Antoniou, and K. Iordanou, "An ai ethics 'david and goliath': Value conflicts between large tech companies and their employees," *AI & SOCIETY*, pp. 1–16, 2022.

[182] H. Schellmann, "Auditors are testing hiring algorithms for bias, but there's no easy fix," *MIT Technology Review, February*, vol. 11, p. 2021, 2021.

[183] L. Weidinger, M. Rauh, N. Marchal, *et al.*, "Sociotechnical safety evaluation of generative ai systems," *arXiv preprint arXiv:2310.11986*, 2023.

[184] A. Birhane, P. Kalluri, D. Card, W. Agnew, R. Dotan, and M. Bao, "The values encoded in machine learning research," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 2022, pp. 173–184.

[185] National institute of Standards and Technology. "Face Recognition Vendor Test (FRVT)," NIST. (Nov. 30, 2020), [Online]. Available: https://www.nist.gov/

programs-projects/face-recognition-vendor-test-frvt (visited on 10/10/2023).
[186] National institute of Standards and Technology. "Face Recogntion Vendor Test (FRVT) 2000," NIST. (2000), [Online]. Available: https://www.nist.gov/itl/iad/image-group/face-recogntion-vendor-test-frvt-2000 (visited on 10/10/2023).
[187] National institute of Standards and Technology. "Face Recognition Vendor Test (FRVT) 2002," NIST. (2002), [Online]. Available: https://www.nist.gov/itl/iad/image-group/face-recognition-vendor-test-frvt-2002 (visited on 10/10/2023).
[188] National institute of Standards and Technology. "Face Recognition Vendor Test (FRVT) 2006," NIST. (2006), [Online]. Available: https://www.nist.gov/itl/iad/image-group/face-recognition-vendor-test-frvt-2006 (visited on 10/10/2023).
[189] National institute of Standards and Technology. "Face Recognition Vendor Test (FRVT) 2013," NIST. (2013), [Online]. Available: https://www.nist.gov/itl/iad/image-group/face-recognition-vendor-test-frvt-2013 (visited on 10/10/2023).
[190] P. Grother, M. Ngan, and K. Hanaoka, "Face Recognition Vendor Test Part 3: Demographic Effects," National Institute of Standards and Technology, Gaithersburg, MD, National Institute of Standards and Technology Interagency or Internal Report (NISTIR) 8280, Dec. 2019, NIST IR 8280. DOI: 10.6028/NIST.IR.8280. [Online]. Available: https://nvlpubs.nist.gov/nistpubs/ir/2019/NIST.IR.8280.pdf (visited on 10/10/2023).
[191] AWO. "Announcing our algorithm governance services." (May 26, 2023), [Online]. Available: https://www.awo.agency/blog/announcing-our-algorithm-governance-services/ (visited on 10/10/2023).
[192] L. Groves, "Algorithmic impact assessment: A case study in healthcare," Ada Lovelace Institute, Feb. 8, 2022. [Online]. Available: https://www.adalovelaceinstitute.org/report/algorithmic-impact-assessment-case-study-healthcare/ (visited on 10/10/2023).
[193] J. Brennan and A. Circiumaru, "Getting under the hood of big tech," Ada Lovelace Institute, Mar. 2022. [Online]. Available: https://www.adalovelaceinstitute.org/blog/getting-under-the-hood-of-big-tech/ (visited on 10/10/2023).

## APPENDIX A
## METHODS: SEARCHING IN NON-ACADEMIC DOMAINS

To produce the analysis in §V, we analyzed reports, press releases, publications, blogs, and web postings from multiple institutions outside of academia. Here, we describe in more detail the specific search methods used to produce our analysis for each institution.

### A. Regulators

*1) National Institute of Standards and Technology (NIST):* We reviewed documents ($N = 16$) from two programs: NIST's Trustworthy & Responsible AI Resource Center (airc.nist.gov) and NIST's Face Recognition Vendor Test (FRVT) program [185]. From the Resource Center, we reviewed the NIST AI Risk Management Framework (RMF) [73] (as well as the accompanying Playbook and Roadmap) and the $N = 5$ special publications and internal reports in the RMF Knowledge Base [156]. From the FRVT program, we reviewed the executive summary of all the pages linked under sections with titles pre-pended "FRVT" ($N = 8$) on the FRVT page [185]–[189], including the homepages for FRVT 2000, 2002, 2006, and 2013, as well as NIST Interagency Reports (NISTIR) 8331, 8280, 7709, and 7830. NISTIR 8280 in particular describes analyses of demographic differences across a large number of algorithms, developers, and test images [190].

*2) Information Commissioner's Office (ICO):* The ICO details audit information on its website ico.org.uk, including definition of audit, assurance ratings, audit target, as well as audit reports. We reviewed the "Audits and overview reports" section yielding 48 results in total.

### B. Law Firms

*1) AWO:* We reviewed publicly available pages on AWO's website (awo.agency), particularly the descriptions of their services at awo.agency/services. We also reviewed all the posts on their blog up to July 2023 ($N = 62$), only a few of which were related to AWO's audit work (AWO [191], for example).

*2) Foxglove:* In order to extract relevant information on the organization, we reviewed the website, mainly the Who We Are and News sections.

*3) Luminos.Law:* Our information on Luminos.Law—called BHN.AI at the time of our study—is sourced from the firm's website, and more particularly, in the AI Audits section.

### C. Civil Society

*1) Electronic Frontier Foundation (EFF):* We looked at the About page to extract general information on the Foundation. We also reviewed the Our Work, and Tools page in order to extract detailed information on including Methods and Tools.

*2) Refugee Law Laboratory (RLL):* The Refugee Law Lab website About page provided the initial general information. For reports, datasets, data visualizations, and impact, we reviewed the Projects and News and Recent Publications sections.

*3) The Citizen Lab:* The About the Citizen Lab provided general information about this group. In order to extract detailed relevant information, we reviewed the Research and News pages.

*4) Migration and Tech Monitor (MTM):* We started our survey with the Home page of MTM's website. The Methodology section details the methods used by the organization, the Resources page lists various case studies carried out by

the organization, and the Snapshots section presents numerous pictorial evidence supporting the organization's work.

*5) Ada Lovelace Institute:* We surveyed the "Projects" listed at adalovelaceinstitute.org/our-work, selecting only those related to AI auditing. We followed any reports listed in those project overviews. In particular, we analyzed the executive summary of $N = 4$ reports on algorithmic impact assessments in healthcare [192], tools and methods for assessing algorithmic systems [24], [25], and auditing standards [193]. We also surveyed $N = 47$ blog posts published July 2023 or earlier at adalovelaceinstitute.org/blog, filtering for the programmes "Biometrics", "Enabling a responsible AI ecosystem", "Ethics and accountability in practice", and "Public-sector use of data & algorithms". We considered only blog posts that discussed the Institute's own AI auditing work.

*6) The American Civil Liberties Union (ACLU):* The ACLU publishes details about its work on its website aclu.org, including news, publications and reports. We reviewed all documents labelled as "reports and other assets" published 2018–2022 ($N = 70$), using aclu.org/search. Most of these did not relate to AI, so we also used Google to search aclu.org/news using the keywords from our academic search as well as other keywords such as "algorithm", "machine learning", and "automated decision" (e.g., `site:aclu.org/news after:2017 ''algorithm''`).

### D. Journalism

*1) Propublica:* We initially scraped the ProPublica website using the keyword "audit". However, this did not yield fruitful as significant number of articles that the keyword returned were not relevant, for example, articles on financial audit and web pages with reports of financial audits. Subsequently, we manually reviewed the website looking specifically for algorithm (and technology) related audits under the Technology section of the website.

*2) The Markup:* We manually sifted through the main page of the Markup's website identifying audit investigation report articles. We reviewed each article and extracted thematic information.

### E. Consulting Agencies

*1) O'Neil Risk Consulting & Algorithmic Auditing (ORCAA):* We reviewed the website starting from the Home page and scrolling through the main pages such as What We Do, NYC Bias Audit and Principles.

*2) Eticas:* We surveyed the Eticas website, more particularly focusing on the Research page of the Eticas Library to glean insights into the organization's algorithmic audit practice.

*3) BABL AI:* We started by reviewing the About Us page of BABL.IA'S website. We then looked at the Services page and extracted contextual information such as the kind of information the agency provides. We also reviewed the Research page for more detailed information on past audits carried out by the agency.

## APPENDIX B
## METHODS: QUANTITATIVE CONTENT ANALYSIS

We supplement our qualitative analysis in §IV by analyzing words and phrases commonly used in the abstracts and keywords of the academic audit studies we collected. All but 4 of the $N = 341$ studies we found were published with an abstract; 78 were published without keywords.

Many of the most-used terms are fairly generic (e.g. "algorithm", or "system), so we show only a selection of key terms relevant to our study. We manually filtered the most frequent 1-grams and 2-grams across all abstracts and keywords in our dataset of academic audits, keeping only those terms relevant to one of the following dimensions of our analysis (see Table II): 1) motivation (e.g., "accountability"); 2) target (including specific entities, e.g., "Facebook"; types of systems, e.g., "computer vision"; and domains, e.g., "healthcare" or "hiring"); 3) types of harms (e.g., "discrimination" or "privacy"); 4) methods (e.g., "qualitative" or "ethnography"). We drop 2-grams that are already encompassed by a more general 1-gram unless they indicate a meaningfully different concept or subset (e.g., "criminal justice" is much different than "justice" alone, whereas, e.g., "algorithm auditing" is encompassed by "auditing", given the scope of our study). We include only terms that were mentioned in more than one keyword list (more than 0.27% of papers) or in more than 10 abstracts (more than 2.7% of papers).

For further analysis, the entire dataset of academic audits we collected and analyzed can be accessed here.[1]

Tables B-I–B-III list the most frequent terms related to audit motivations and harms, audit targets, and audit methods, respectively. Tables B-IV–B-VI further compare the most frequent keywords across audit types.

[1]https://drive.google.com/file/d/1M4QnksZCccaALijjmnL9LeLB8redwhou/view?usp=sharing

TABLE B-I
MOST FREQUENT TERMS: AUDIT MOTIVATION AND TYPES OF HARMS.

| **Term(s)** | **# times used** | **# studies (% of total)** |
|---|---|---|
| *Terms in keywords (only terms in $\geq 1\%$ of studies)* | | |
| fairness/fair | 75 | 66 (25.1%) |
| bias/biases/biased | 75 | 65 (24.7%) |
| audit/audits/auditing | 56 | 49 (18.6%) |
| accountability/accountable | 42 | 41 (15.6%) |
| ethics/ethical | 26 | 26 (9.9%) |
| impact/impacts | 17 | 15 (5.7%) |
| evaluate/evaluation/evaluations/evaluating | 17 | 13 (4.9%) |
| privacy | 16 | 12 (4.6%) |
| transparency | 12 | 12 (4.6%) |
| explainability/explanations/explanation/xai | 12 | 10 (3.8%) |
| discrimination/discriminatory | 8 | 8 (3.0%) |
| justice | 8 | 8 (3.0%) |
| governance | 6 | 6 (2.3%) |
| hate speech | 6 | 6 (2.3%) |
| impact assessment(s) | 6 | 6 (2.3%) |
| disparate impact / disparity / disparities | 5 | 5 (1.9%) |
| gender bias | 5 | 5 (1.9%) |
| harm/harms | 5 | 5 (1.9%) |
| risk | 5 | 5 (1.9%) |
| data protection | 5 | 4 (1.5%) |
| interpretability | 4 | 4 (1.5%) |
| misinformation | 6 | 4 (1.5%) |
| security | 4 | 4 (1.5%) |
| surveillance | 4 | 4 (1.5%) |
| trust | 4 | 4 (1.5%) |
| accessibility | 3 | 3 (1.1%) |
| *Terms in abstract (only terms in $\geq 5\%$ of studies)* | | |
| bias/biases/biased | 325 | 99 (29.4%) |
| audit/audits/auditing | 201 | 87 (25.8%) |
| impact/impacts | 135 | 85 (25.2%) |
| fairness/fair | 271 | 78 (23.1%) |
| evaluate/evaluation/evaluations/evaluating | 114 | 69 (20.5%) |
| accountability/accountable | 112 | 56 (16.6%) |
| harm/harms | 74 | 44 (13.1%) |
| transparency | 62 | 44 (13.1%) |
| ethics/ethical | 65 | 37 (11.0%) |
| discrimination/discriminatory | 57 | 34 (10.1%) |
| risk | 55 | 34 (10.1%) |
| mitigate | 31 | 28 (8.3%) |
| privacy | 56 | 25 (7.4%) |
| explainability/explanations/explanation/xai | 63 | 22 (6.5%) |
| disparate impact / disparity / disparities | 37 | 21 (6.2%) |
| justice | 31 | 21 (6.2%) |

TABLE B-II
MOST FREQUENT TERMS: TARGET SYSTEMS AND DOMAINS.

| **Term(s)** | **# times used** | **# studies (% of total)** |
|---|---|---|
| *Terms in keywords (only terms in $\geq 1\%$ of studies)* | | |
| social media / social networks / facebook / twitter / youtube | 18 | 17 (6.5%) |
| computer vision | 13 | 13 (4.9%) |
| health/healthcare/medical | 14 | 11 (4.2%) |
| natural language / language processing | 22 | 11 (4.2%) |
| advertising/ads | 8 | 7 (2.7%) |
| platform/platforms | 6 | 6 (2.3%) |
| hiring/employment | 5 | 5 (1.9%) |
| policing/police | 5 | 5 (1.9%) |
| credit | 4 | 4 (1.5%) |
| criminal justice | 4 | 4 (1.5%) |
| multimodal | 4 | 4 (1.5%) |
| search engines | 4 | 4 (1.5%) |
| speech recognition / speaker recognition / speaker verification | 6 | 4 (1.5%) |
| automated decision | 3 | 3 (1.1%) |
| facial recognition / face recognition | 3 | 3 (1.1%) |
| generative | 3 | 3 (1.1%) |
| google | 3 | 3 (1.1%) |
| government | 3 | 3 (1.1%) |
| risk assessment | 3 | 3 (1.1%) |
| social credit | 3 | 3 (1.1%) |
| welfare | 3 | 3 (1.1%) |
| *Terms in abstract (only terms in $\geq 5\%$ of studies)* | | |
| platform/platforms | 99 | 45 (13.4%) |
| health/healthcare/medical | 89 | 42 (12.5%) |
| social media / social networks / facebook / twitter / youtube | 99 | 40 (11.9%) |
| advertising/ads | 79 | 21 (6.2%) |
| hiring/employment | 23 | 18 (5.3%) |
| google | 22 | 17 (5.0%) |
| government | 22 | 17 (5.0%) |

TABLE B-III
MOST FREQUENT TERMS: METHODS.

| **Term(s)** | **# times used** | **# studies (% of total)** |
|---|---|---|
| *Terms in keywords (only terms in $\geq 1\%$ of studies)* | | |
| sociotechnical | 9 | 9 (3.4%) |
| participatory | 8 | 7 (2.7%) |
| hci | 7 | 6 (2.3%) |
| qualitative/ethnography/interviews/interviewed/workshop(s) | 7 | 5 (1.9%) |
| participatory design | 5 | 5 (1.9%) |
| community/communities | 4 | 4 (1.5%) |
| human centered | 4 | 4 (1.5%) |
| interdisciplinary | 5 | 3 (1.1%) |
| *Terms in abstract (only terms in $\geq 5\%$ of studies)* | | |
| community/communities | 70 | 48 (14.2%) |
| qualitative/ethnography/interviews/interviewed/workshop(s) | 37 | 26 (7.7%) |
| benchmark/benchmarks | 50 | 24 (7.1%) |
| sociotechnical | 27 | 22 (6.5%) |

TABLE B-IV
FREQUENT TERMS BY AUDIT TYPE: AUDIT MOTIVATION AND TYPES OF HARMS.

| **Term(s)** | **# studies (% of studies of the same audit type)** | | | |
|---|---|---|---|---|
| | *Data Audit* | *Product/Model/ Algo. Audit* | *Ecosystem Audit* | *Meta-Commentary* |
| *Terms in keywords* | | | | |
| bias/biases/biased | 8 (**29.6**%) | 48 (**28.4**%) | 3 (**20.0**%) | 6 (11.5%) |
| fairness/fair | 8 (**29.6**%) | 40 (**23.7**%) | 3 (**20.0**%) | 15 (**28.8**%) |
| audit/audits/auditing | 1 (3.7%) | 35 (**20.7**%) | 1 (6.7%) | 12 (**23.1**%) |
| accountability/accountable | 1 (3.7%) | 16 (9.5%) | 3 (**20.0**%) | 21 (**40.4**%) |
| ethics/ethical | 3 (11.1%) | 13 (7.7%) | 0 | 10 (**19.2**%) |
| *Terms in abstract* | | | | |
| bias/biases/biased | 18 (**39.1**%) | 69 (**32.5**%) | 2 (13.3%) | 10 (**15.6**%) |
| fairness/fair | 8 (**17.4**%) | 45 (**21.2**%) | 4 (**26.7**%) | 21 (**32.8**%) |
| audit/audits/auditing | 4 (8.7%) | 57 (**26.9**%) | 1 (6.7%) | 25 (**39.1**%) |
| evaluate/evaluation/evaluations/evaluating | 11 (**23.9**%) | 45 (**21.2**%) | 2 (13.3%) | 11 (**17.2**%) |
| impact/impacts | 12 (**26.1**%) | 52 (**24.5**%) | 4 (**26.7**%) | 17 (**26.6**%) |
| accountability/accountable | 4 (8.7%) | 29 (13.7%) | 5 (**33.3**%) | 18 (**28.1**%) |
| transparency | 5 (10.9%) | 26 (12.3%) | 4 (**26.7**%) | 9 (14.1%) |
| ethics/ethical | 7 (**15.2**%) | 19 (9.0%) | 2 (13.3%) | 9 (14.1%) |
| harm/harms | 8 (**17.4**%) | 19 (9.0%) | 4 (**26.7**%) | 13 (**20.3**%) |
| risk | 2 (4.3%) | 20 (9.4%) | 3 (**20.0**%) | 9 (14.1%) |
| mitigate | 3 (6.5%) | 18 (8.5%) | 3 (**20.0**%) | 4 (6.2%) |

Showing only terms that appear in $\geq 15\%$ of studies for at least one audit type (bolded).

TABLE B-V
FREQUENT TERMS BY AUDIT TYPE: TARGET SYSTEMS AND DOMAINS.

| **Term(s)** | **# studies (% of studies of the same audit type)** | | | |
|---|---|---|---|---|
| | *Data Audit* | *Product/Model/ Algo. Audit* | *Ecosystem Audit* | *Meta-Commentary* |
| *Terms in keywords* | | | | |
| social media / social networks / facebook / twitter / youtube | 0 | 17 (**10.1**%) | 0 | 0 |
| natural language / language processing | 3 (**11.1**%) | 6 (3.6%) | 1 (6.7%) | 1 (1.9%) |
| computer vision | 7 (**25.9**%) | 5 (3.0%) | 0 | 1 (1.9%) |
| welfare | 1 (3.7%) | 0 | 2 (**13.3**%) | 0 |
| child welfare | 0 | 0 | 2 (**13.3**%) | 0 |
| *Terms in abstract* | | | | |
| social media / social networks / facebook / twitter / youtube | 2 (4.3%) | 38 (**17.9**%) | 0 | 0 |
| platform/platforms | 1 (2.2%) | 42 (**19.8**%) | 1 (6.7%) | 1 (1.6%) |
| health/healthcare/medical | 5 (**10.9**%) | 30 (**14.2**%) | 1 (6.7%) | 6 (9.4%) |
| computer vision | 7 (**15.2**%) | 8 (3.8%) | 0 | 1 (1.6%) |
| natural language / language processing | 5 (**10.9**%) | 9 (4.2%) | 0 | 1 (1.6%) |
| hiring/employment | 2 (4.3%) | 11 (5.2%) | 2 (**13.3**%) | 3 (4.7%) |
| government | 2 (4.3%) | 8 (3.8%) | 3 (**20.0**%) | 4 (6.2%) |
| welfare | 1 (2.2%) | 3 (1.4%) | 2 (**13.3**%) | 0 |
| child welfare | 0 | 1 (0.5%) | 2 (**13.3**%) | 0 |
| education | 1 (2.2%) | 6 (2.8%) | 2 (**13.3**%) | 1 (1.6%) |

Showing only terms that appear in $\geq 10\%$ of studies for at least one audit type (bolded).

TABLE B-VI
FREQUENT TERMS BY AUDIT TYPE: METHODS.

| **Term(s)** | **# studies (% of studies of the same audit type)** | | | |
|---|---|---|---|---|
| | *Data Audit* | *Product/Model/ Algo. Audit* | *Ecosystem Audit* | *Meta-Commentary* |
| *Terms in keywords* | | | | |
| qualitative/ethnography/interviews/interviewed/workshop(s) | | 3 (1.8%) | 2 (**13.3**%) | 0 |
| participatory | | 1 (0.6%) | 3 (**20.0**%) | 3 (5.8%) |
| participatory design | | 0 | 3 (**20.0**%) | 2 (3.8%) |
| human centered | | 2 (1.2%) | 2 (**13.3**%) | 0 |
| *Terms in abstract* | | | | |
| benchmark/benchmarks | 14 (**30.4**%) | 6 (2.8%) | 0 | 4 (6.2%) |
| community/communities | 9 (**19.6**%) | 21 (9.9%) | 6 (**40.0**%) | 12 (**18.8**%) |
| qualitative/ethnography/interviews/interviewed/ workshop(s) | 3 (6.5%) | 13 (6.1%) | 6 (**40.0**%) | 4 (6.2%) |
| sociotechnical | 1 (2.2%) | 12 (5.7%) | 1 (6.7%) | 8 (**12.5**%) |
| hci | 0 | 3 (1.4%) | 2 (**13.3**%) | 1 (1.6%) |
| participatory | 1 (2.2%) | 0 | 2 (**13.3**%) | 1 (1.6%) |

Showing only terms that appear in $\geq 10\%$ of studies for at least one audit type (bolded).

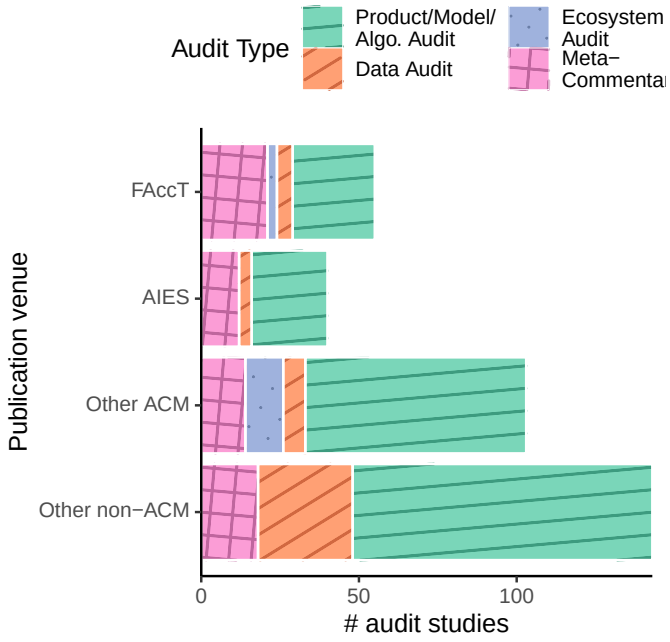


Fig. C-I. Number of collected academic audit studies published in each year, grouped by publication source.

## APPENDIX C
## ADDITIONAL RESULTS

Here, we include more detailed summaries of the non-academic auditors we reviewed.

TABLE C-VII
CONSULTING AGENCIES

| Consulting Agencies | ORCAA | Eticas | Babl AI |
|---|---|---|---|
| **Motivation** | regulatory compliance, performance testing, measure and mitigate disparate performance | help companies comply with legal requirements, security and data protection, fairness, bias and model accuracy | bias assessment, risk and impact assessment |
| **Target** | predictive scoring systems, hiring algorithms, healthcare AI, AI platforms, ADS, FRT | risk assessment algorithms; social media platforms such as tiktok and youtube; ride hailing apps, such as uber, cabify and bolt; algorithms used by governments and companies | their clients include universities, small AI companies, startups like Proctorio |
| **Types of harm** | discrimination, bias, measure gender and race/ethnicity bias, assess for regulatory compliance | algorithmic systems (their functioning), to detect anomalies or practices that could be unfair towards protected groups or society | bias, fairness, effectiveness, transparency, best policies and standards, privacy, compliance |
| **Institutional context** | The agency assists tech giants and governments, while its audits prioritize the concerns of marginalized communities | The consulting agency champions fairness for protected groups, yet also safeguards corporations from financial and reputational risks. | Ultimately, they are consultants that work with/for clients to help orgs build trust with stakeholders. |
| **Methods** | internal tool (Ethical Matrix framework), review documents | data management planning, encryption, ethics due diligence and vetting | in-house developed set of criteria used to conduct conduct bias testing, review documentations, interact with stakeholders |
| **Impact** | some influence (albeit indirect) on policy (on the White House AI ethics blueprint) | not clear | none stated |

TABLE C-VIII
CORPORATE AUDITS

| Corporate Audits | Google | Facebook |
|---|---|---|
| **Motivation** | identify potential issues relating to celebrity recognition products | help the company identify, prioritise, and implement improvements in accordance with civil rights ”at the behest of the civil rights community and members of the US congress” |
| **Target** | Google’s celebrity recognition API | Facebook |
| **Types of harm** | human rights *impacts* | racial injustice, voter suppression, hate speech, algorithmic bias |
| **Methods** | consultation with stakeholders, dialogue with expert resources | interview with hundreds of civil rights orgs and advocates and members of congress |
| **Impact** | incorporation of some of the findings into product design | facebook committed to implement (no information if these changes were in fact implemented) a new advertising system in accordance with civil rights |

TABLE C-IX
JOURNALISM

| Journalism | The MarkUp | Propublica |
|---|---|---|
| **Motivation** | Accountability, expose disparities injustice and illegal practice | accountability, expose disparities in performance, privacy violations, discrimination in deployed systems |
| **Target** | arge corporations/institutions, social media platforms, algorithms, content moderators | Gov't ADS, VLOPs (Facebook advertising, Amazon product placement) public programs with digital sites |
| **Types of harm** | Disparities, fraud, discrimination, legal compliance, privacy breach | discrimination, privacy harms, cultural genocide, injustice |
| **Institutional context** | Targeting mostly tech firms/platforms, on behalf of general public; funded by foundations and private donors | Targeting public agencies and private companies, on behalf of general public; funded by foundations and private donors |
| **Methods** | Inhouse audit tools, interview, data donation, reviewing documents | scraping, data donation, API exploitation, evaluation, interviews, qualitative evidence |
| **Impacts** | high impact resulting in legislative and systemic (in the US) | High impact, legislative action, company policy changes, COMPAS a canonical work that set president for academic research |

TABLE C-X
LAW FIRMS

| Law firms | BNH.AI | Foxglove | AWO |
|---|---|---|---|
| **General Objectives** | | | |
| **Audit type** | information not available | Case study. algorithms, social media platforms, living conditions of tech workers | evaluative reviews of general industries or technologies, case studies |
| **Audit Goal** | compliance with standards | keeping tech giants and governments accountable | evaluate/design governance, protect data rights |
| **Audit target** | models and data | social media platforms, governments, content moderator, warehouse workers, gig-workers | VLOPs, private companies, FRT |
| **Type of harm** | legal compliance/liability/legal defence | ustice, disparate performance, enforcing legal requirements | privacy/data rights/surveillance violations, safety, digital manipulation/exploitation |
| **Methods** | information not available | legal compliance, anecdotal evidence | documentation, critical commentary, policy development, legal/compliance research |
| **Impact** | information not available | Sforced disclosure of secret contract between tech corps and governments, stopped the UK Home Office use of visa-streaming algo, reversed Ofqual's A level grading algorithm | litigation, change to govt's and corporate policy/strategy |
| **Power Analysis** | corporate/for profit, confidential and privileged | non-profit | corporate/for profit, consulting for both corporate and public agencies |

TABLE C-XI
REGULATORS

| **Regulators** | **ICO** | **NIST** |
|---|---|---|
| **General Objectives** | Uphold information rights. Enforcing legal requirements. Regulatory interventions through guidance, enforcement notice, and issuing monetary penalties. | Quality assurance, develop best standards and regulatory practices |
| **Audit type** | Case studies | Meta-commentary |
| **Audit Goal** | Investigate and enforce regulations | Establish standards |
| **Audit target** | Data, data sharing and management documents | FRT, ADS |
| **Type of harm** | Legal compliance, privacy breach, data management | Racial disparities, physical safety, functionality, privacy |
| **Methods** | Interview, testing, reviewing documents | Data trust, accuracy evaluation, benchmarking |
| **Impact** | Significant monetary penalties issued | Significant contribution on standardisation |
| **Power Analysis** | Targets companies and gov't agencies; has authority to enforce laws with fines | Targets companies and gov't agencies; establishes standards but no enforcement authority |

TABLE C-XII
CIVIL SOCIETY

| Civil Society | EFF | RLL | The Citizen Lab | MTM | Ada Lovelace | ACLU |
|---|---|---|---|---|---|---|
| **Motivation** | accountability, non-profit and independent | accountability, non-profit and independent | accountability | accountability | accountability | accountability |
| **Target** | ADS, online platforms, large corporations | ADS, lie detectors, border patrolling drones | ADS, apps, code, websites, FRT, surveillance tech, hardware (phones and other devices used by politicians and activists), software | biometric data, digital ID systems, border surveillance vendors, surveillance drones, FRT, iris scan data, lie detector tech, thermal cameras, algorithmic motion detectors, AI powered satellites, ankle monitors and GPS tags, voice recognition tech | AI and data-driven systems generally, biometric data (facial recognition) and healthcare in particular | ADS, platform advertising, facial recognition tech, medical algorithms, online advertising (particularly Facebook), welfare programs, insurance pricing, redlining algorithms |
| **Types of harm** | security vulnerabilities, fair use, user rights, privacy, free speech, resisting surveillance, encryption, consumer protection | disparities in refugee claim recognition rates, human rights violations, digital rights violations, xenophobic and racial discrimination, privacy | privacy, security, transparency, surveillance, spyware, censorship, freedom of expression | racial discrimination, border surveillance, xenophobia, human rights violations | privacy harms, due process harms, disparate impact | fallacy of functionality, cultural hegemony, hate speech / toxicity |
| **Methods** | litigation, policy analysis, grassroot activism, technology development | interviews with refugees, data collection, analysis and interactive visualisation from the Immigration and Refugee Board (IRB) through Access to Information Request, documentations of deportation and refused refugee claims, film making | in-house tools, forensic analysis of devices, interviews | document and archive migration tech, interviews with people crossing borders, investigative analysis, photography | participatory methods, including opinion polling; also policy research and policy development | quantitative evaluation of operational systems (but not necessarily as rigorous as academic papers; may be simple metrics); data acquired through privileged access, public records requests, scraping |
| **Impact** | high impact. Sued and won a case for First Amendment protection for software code for privacy protection, class-action lawsuit that resulted in removing Sony BMG off the market, won a rulings against the US government's attempt to track location of mobile phone users, won Proctorio lawsuit on behalf of a student | some impact around advocacy | significant media coverage and attention to some investigations | some impact. TUI stopped deportation flights as a result of MTM's campaigns | | |