Preprint, to appear in PACM Human-Computer Interaction, CSCW '21

Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits

JACK BANDY, Northwestern University, USA

While algorithm audits are growing rapidly in commonality and public importance, relatively little scholarly work has gone toward synthesizing prior work and strategizing future research in the area. This systematic literature review aims to do just that, following PRISMA guidelines in a review of over 500 English articles that yielded 62 algorithm audit studies. The studies are synthesized and organized primarily by behavior (discrimination, distortion, exploitation, and misjudgement), with codes also provided for domain (e.g. search, vision, advertising, etc.), organization (e.g. Google, Facebook, Amazon, etc.), and audit method (e.g. sock puppet, direct scrape, crowdsourcing, etc.). The review shows how previous audit studies have exposed public-facing algorithms exhibiting problematic behavior, such as search algorithms culpable of distortion and advertising algorithms culpable of discrimination. Based on the studies reviewed, it also suggests some behaviors (e.g. discrimination on the basis of intersectional identities), domains (e.g. advertising algorithms), methods (e.g. code auditing), and organizations (e.g. Twitter, TikTok, LinkedIn) that call for future audit attention. The paper concludes by offering the common ingredients of successful audits, and discussing algorithm auditing in the context of broader research working toward algorithmic justice.

Additional Key Words and Phrases: algorithm auditing, literature review, ethics, policy, critical algorithm studies, algorithmic bias, algorithmic authority, algorithmic accountability

ACM Reference Format:
Jack Bandy. 2021. Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 74 (April 2021), 34 pages. https://doi.org/10.1145/3449148

1 INTRODUCTION

Algorithm auditing has become an important research method for diagnosing problematic behavior in algorithmic systems. For example, do targeted advertising algorithms facilitate discrimination? [98, 146] Does YouTube's recommendation algorithm elevate extremist videos? [57] Do facial recognition algorithms perform worse on darker-skinned females? [132] Does Google's search algorithm favor certain news outlets? [136, 149] These questions continue to multiply as algorithmic systems become more pervasive and powerful in society.

Despite the growing commonality and public importance of algorithm audits that address these problems, relatively little work has gone toward clarifying the past trajectory and future agenda of algorithm auditing. Systematic literature reviews (SLRs) – sometimes referred to as literature surveys [163] – have been a prominent part of computing research [1, 39, 63, 76, 102], and help "identify trends and gaps in the literature" [46] to clarify shared research goals for future work. To this end, the present paper conducts a scoped SLR of algorithm audits, screening over 500 papers from a variety of journals and conferences. By thematically analyzing 62 studies that audited

Author's address: Jack Bandy, Northwestern University, Evanston, Illinois, USA, jackbandy@u.northwestern.edu.

2573-0142/2021/4-ART74 $15.00
https://doi.org/10.1145/3449148

public-facing algorithmic systems, we identify a taxonomy of problematic machine behaviors that helps organize prior work as well as strategize for future work.

Based on the review, we identify four major types of problematic behavior in algorithmic systems: discrimination, distortion, exploitation, and misjudgement. Out of the 62 studies reviewed, most focused on discrimination (N=21) or distortion (N=29). Audit studies also gave more attention to search algorithms (N=25), advertising algorithms (N=12), and recommendation algorithms (N=8), helping to diagnose a range of problematic behaviors on these systems. Taken together, these empirical studies provide substantial evidence that the public harms of algorithmic systems already exist in real-world systems, not merely in hypothetical scenarios. These previous studies also help chart a promising path forward for future audits that can serve as a diagnostic for "how social problems manifest in technical systems" [2].

Specifically, the review suggests that some behaviors, domains, and methods call for future audit attention. In terms of behaviors, future audits should further explore discrimination on the basis of intersectional identity. For example, dynamic pricing algorithms might be audited to directly measure the allocative harms of price discrimination on the basis of intersectional identities pertaining to race, age, sex, gender, and more. This paper joins a growing chorus of calls (e.g. [21, 45, 71, 79, 125, 127]) for computing research to recognize identity as multi-faceted rather than one-dimensional. In terms of domains that deserve further research attention, advertising algorithms form the economic backbone of large technology companies, and they pose a number of potential allocative and representational harms to the public. Code auditing appears to be under-explored as an audit method, even though a growing number of influential algorithms are open-sourced. Finally, the studies we reviewed offered limited insights on some organizations such as Twitter, LinkedIn, and TikTok, even though these organizations operate widely influential algorithmic systems.

Formally, this literature review reports what previous algorithm audits have done (RQ1) and what remains to be done in future audits (RQ2). It also outlines important ingredients for successful audit studies, then concludes by discussing algorithm auditing as a diagnostic tool within a broader research effort working toward algorithmic justice.

2 RELATED WORK

Literature reviews (sometimes referred to as literature surveys [163]) have been a prominent part of computing research [1, 39, 63, 76, 102], and a number of reviews are closely related to this work. A 2017 book by Cathy O'Neil [126] illustrated how various algorithmic systems create disparate impact, and the chapters discuss discrimination in hiring algorithms, loan approval algorithms, labor-scheduling algorithms, and more. Sandvig et al. [140] review methods for algorithm auditing, and include some examples of each method. A 2016 review article sought to "map the debate" around ethics and algorithms [114], and a 2018 review article [1] identified a research agenda for developing "Explainable, Accountable and Intelligible Systems." While they do overlap with the present work, these studies did not address algorithm auditing as a primary topic.

One closely-related literature review examined ethical considerations for data science [139]. The authors identified three challenges related to data (privacy/anonymity, misuse, and accuracy/validity), as well as three challenges related to mathematical modeling (personal/group harm, subjective model design, and model misuse/misinterpretation). Since these challenges are pertinent to both data science and algorithmic systems, many algorithm audits center around the same topics, especially personal/group harm.

There has also been important related work in developing frameworks and theories for algorithm auditing and accountability. Raji et al. [133] introduce a framework intended for internal auditing, which organizations could use when developing algorithms. This framework is exceedingly helpful,

although the focus on internal development leads to different audit considerations compared to this review with its focus on public-facing algorithms that have already been deployed. One related literature review by Wieringa [160] outlines a theory of "algorithmic accountability" through a synthesis of over 200 English articles. The study uses a conceptual framework for accountability to organize the findings, and concludes with a robust definition for algorithmic accountability: "Algorithmic accountability concerns a networked account for a socio-technical algorithmic system, following the various stages of the system's lifecycle." Building on this review, the current work focuses specifically on audit studies aimed at algorithmic accountability, and synthesizes the types of behavior that require accountability.

Finally, a 2017 literature review of "The sharing economy in computing" [46] overlaps with this work by including some algorithm audits of systems such as Uber and TaskRabbit. They labeled papers as "algorithmic auditing" if they described "how algorithms can rule the sharing economy sites," finding nine papers that met this definition in the sharing economy literature. This review expands the scope beyond just the sharing economy, including audits of search algorithms like Google, recommendation algorithms such as those used by Spotify, computer vision algorithms such as Amazon's "Rekognition," and more. Still, the literature review presented by Dillahunt et al. [46] was instructive for identifying potentially relevant studies. It also posed two clear research questions for SLRs in the field of Human-Computer Interaction, which were used in other SLRs [76, 130] and also served as a model for this review's research questions. The first question is about *what has been done* and the second is about *what is next to do*:

- **RQ1**: What kinds of problematic machine behavior have been diagnosed by previous algorithm audits?
- **RQ2**: What remains for future algorithm audits to examine the problematic ways that algorithms exercise power in society?

3 METHODS

To answer these two research questions, I conducted a scoped review using the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [117]. A scoping review is useful "when what is needed is not detailed answers to specific questions but rather an overview of a broad field" [118], and our broad research question aligned with this goal. PRISMA was originally developed for health sciences to systematically asses interventions through meta-analyses, and its evidence-based standards for transparency, comprehensiveness, and bias mitigation have led to usage in a number of other fields. This includes CSCW literature [76, 102] and other computing research [153, 160].

The first two main stages of PRISMA are *identification* – finding a pool of potentially relevant studies – and *screening* – manually reviewing article metadata for potential relevance. These first two steps required a definition of an algorithm audit, before moving to the third and fourth stages of *eligibility* – full-text review – and *inclusion* – the final analysis and synthesis stage.

3.1 What is an algorithm audit?

3.1.1 Social Change and Public Impact. While algorithm auditing can serve many purposes, this literature review focused on audits that provide public accountability as a potential means of positive social change. Abebe et al. [2] recently suggested that one of the major ways computing can play a role in social change is through diagnosing "how [social problems] manifest in technical systems." In this same spirit of social change, Raji and Buolamwini [132] offered the following definition of an algorithm audit:

> "An algorithmic audit involves the collection and analysis of outcomes from a fixed algorithm or defined model within a system. Through the stimulation of a mock user population, these audits can uncover problematic patterns in models of interest."

Social change and public impact have also become important topics not just for auditing researchers, but for computing researchers at large [62]. Rather than "assume that [computing] research will have a net positive impact on the world," [7], scholars are now grappling with the larger societal implications of technology, especially "the ways in which power, extraction, and oppression permeate socio-technical systems" [62]. Because of this growing focus on social change, our definition of an algorithm audit focuses on the potential for algorithm audits to provide meaningful accountability to the public.

3.1.2 Definition. After reviewing proposed definitions in the auditing literature, the definition used for the review was *an empirical study investigating a public algorithmic system for potential problematic behavior*. Specifically, this entailed the following:

- An *empirical study* includes an experiment or analysis (quantitative or qualitative) that generates evidence-based claims with well-defined outcome metrics. It must not be purely an opinion/position paper, although position papers with substantial empirical components were included.
- An *algorithmic system* is any socio-technical system influenced by at least one algorithm. This includes systems that may rely on human judgement and/or other non-algorithmic components, as long as they include at least one algorithm.
- A *public* algorithmic system is one used in a commercial context or other public setting such as law enforcement, education, criminal justice, or public transportation.
- *Problematic behavior* in this study refers to discrimination, distortion, exploitation, or misjudgement, as well as various types of behaviors within each of these categories. A behavior is problematic when it causes harm (or potential harm). In the ACM Code of Ethics, examples of harm include "unjustified physical or mental injury, unjustified destruction or disclosure of information, and unjustified damage to property, reputation, and the environment."[1] See Rahwan et al. [131] for a discussion of "machine behavior" which guided the notion of "*problematic* machine behavior."

3.2 Identification

3.2.1 Keyword Search. The first author designed a keyword search to identify a wide body of records that were potentially relevant to the initial definition of algorithm audits. Keywords were generated iteratively through exploratory searches on Google Scholar. Early searches, based on initial domain knowledge, included "algorithm auditing," "platform audit," and "algorithmic bias." After inspecting results for these keyword searches, several keywords and keyphrases were added. Due to the empirical component of the definition, a boolean key was added to search for papers that included an analysis, experiment, a study, or an audit.

Due to its interdisciplinary database, flexible search options, and reproducible results, the Scopus database [52] was used to identify records. Given the various disciplines relevant to algorithm auditing, it was important to search records in fields like economics, journalism, and law, rather than only computing. Scopus is known to have an expansive database [58], and other literature reviews in computing have utilized it during the identification phase [1, 114].

Scopus' flexible search options also allowed an expanded search by identifying papers that reference any of three influential papers in algorithm auditing. The first author selected these

[1] https://www.acm.org/code-of-ethics

Relevant to algorithm auditing		**Empirical Study**	
Title, abstract, or keyword contains:	"algorithmic bias"	Title, abstract, or keyword contains:	"study"
OR Title, abstract, or keyword contains:	"algorithmic discrimination"	OR Title, abstract, or keyword contains:	"experiment"
OR Title, abstract, or keyword contains:	"algorithmic fairness"	OR Title, abstract, or keyword contains:	"audit"
OR Title, abstract, or keyword contains:	"algorithmic accountability"	OR Title, abstract, or keyword contains:	"analysis"
OR Full text contains:	"algorithm audit*"		
OR Full text contains:	"algorithmic audit*"		
OR The paper references:	"Auditing Algorithms: Research Methods..." [140]		
OR The paper references:	"Thinking critically about and researching algorithms" [92]		
OR The paper references:	"The Relevance of Algorithms" [65]		

Table 1. The boolean search used to identify potential articles in the Scopus database

papers after noting they were frequently cited as motivation in relevant studies, and confirming each paper was a highly-cited contribution to the field of algorithm auditing. Sandvig et al. [140] (300 citations) provides the first methodological overview of algorithm auditing, Gillespie [65] (1,500 citations) discusses the "relevance of algorithms" as having significant societal impact, and Kitchin [92] (500 citations) discusses both the relevance of algorithms to society and potential methods to critically examine them (citation counts based on Google Scholar in August 2020). The final boolean search string (Table 1) generated 506 initial records.

3.2.2 Additional Sources. To improve coverage of the review, additional sources were identified and included throughout the writing process, including citations encountered during full-text screening and during the peer review process. The first author maintained a list of papers that were referenced as algorithm audits during full text screening, and added them to the review pipeline. Some papers were also added to the review pipeline during peer review, when reviewers recommended additional algorithm audit studies. Including these additional studies helps minimize any bias introduced in the keyword search, for example, many papers identified from these sources were published in or before 2015. This suggests the initial Scopus search may have suffered from a recency bias which excluded many studies prior to 2016. In total, additional sources identified 36 studies as potentially relevant to the review. As with the keyword search, the papers were screened and filtered based on titles and abstracts.

3.3 Title and Abstract Screening

During the initial screening, the first author examined the title and abstracts for 503 papers (3 duplicates were removed from the original set of 506). Based on this review, 416 papers did not fit the initial definition, which filtered the set down to 87 potential studies that met or likely met the criteria. 36 additional papers were identified through additional sources, so in total, 123 papers were reviewed in full text screening.

3.4 Full Text Screening

3.4.1 Exclusion Reasons. The first author screened each of the texts that were potentially relevant and excluded papers that did not match the initial definition of an algorithm audit. While many studies were relevant and important for the research area, many did not constitute an algorithm audit for a variety of reasons:

- **Theory or methods** (n=16): the paper focused on theoretical or methodological topics, including proposed fairness metrics and other potential solutions to mitigate problematic behavior. For example, Badami et al. [9] propose a method for mitigating polarization in recommender systems, but do not conduct an algorithm audit.

- **Non-public** (n=13): the study focused on an algorithmic system that does not directly face the public. For example, Möller et al. [119] simulate hypothetical recommendation algorithms to audit for echo chamber effects, rather than auditing a real-world algorithm.
- **Non-algorithm** (n=12): the study audited an aspect that may *pertain* to an algorithmic system, such as an interface or input data, but did not substantially audit the algorithm. For example, May et al. [107] investigate gender disparities in participation on Stack Overflow, but do not explore how Stack Overflow's algorithms may be implicated in the disparities.
- **User study** (n=9): the paper focused on how people experience or perceive an algorithmic system, but did not audit the algorithm (i.e. it was a study of human behavior, not machine behavior). For example, Díaz and Diakopoulos [43] conducted an interview study about an algorithm that calculated neighborhood "walkability" scores, but did not conduct an algorithm audit by testing inputs and outputs.
- **Development history** (n=5): the study examined the development, design, and/or implementation of an algorithmic system, but did not substantially audit the algorithm. For example, DeVito [40] provides insight about Facebook's News Feed algorithm by exploring the history and context of its development.

Furthermore, 6 papers could not be reviewed due to lack of access to the full text, resulting in 62 total papers included in the final review. Figure 1 shows the flowchart of exclusion and inclusion in terms of the PRISMA guidelines.

3.5 Analysis and Synthesis

With 62 relevant studies identified, the author read, summarized and thematically analyzed the papers. Based on categories from two meta-analyses of CSCW scholarship, the first author coded for general data categories: year, domain (e.g. search, vision, recommendation, etc.), and organization audited (e.g. Google, YouTube, Uber, etc.). The first author also coded the audit method used, using the following definitions for audit methods from Sandvig et al. [140]:

- **Code audit**: researchers obtain and analyze the code that makes up the algorithm
- **Direct scrape**: researchers collect data directly via an API or other systematic query
- **Sock puppet**: researchers collect data by creating computer programs which impersonate users who then test the algorithm
- **Carrier puppet**: similar to the sock puppet method, except that the impersonated users affect the real-world system and may "carry" effects onto end users
- **Crowdsourcing**: researchers collect data by hiring end users to test the algorithm

Finally, to identify themes from the data, the author worked with another coder using inductive thematic analysis [20], which aims to link themes directly to the data. In our case, given the focus on problematic behavior, this meant linking problematic machine behaviors directly to algorithm audit studies. As in prior CSCW literature reviews (e.g. [76]), one author developed an open and iterative coding scheme using language directly from the included papers. When a study explicitly named a problematic behavior, the phrase was incorporated verbatim into the coding process. The author then iteratively grouped, synthesized, and renamed these themes until the process produced a final scheme of behaviors which worked "in relation to the coded extracts and the entire data set" [20].

After the initial behaviors were defined and named, the second coder used the scheme to independently code 12 random studies (roughly a 20% sample) from the full set of 62. Inter-rater reliability was moderate in this first round (Cohen's $\kappa = 0.42$), after which the reviewers discussed discrepancies and potential modifications to the coding scheme. The first author then coded all papers with a modified coding scheme, and the second coder applied this scheme to another set of

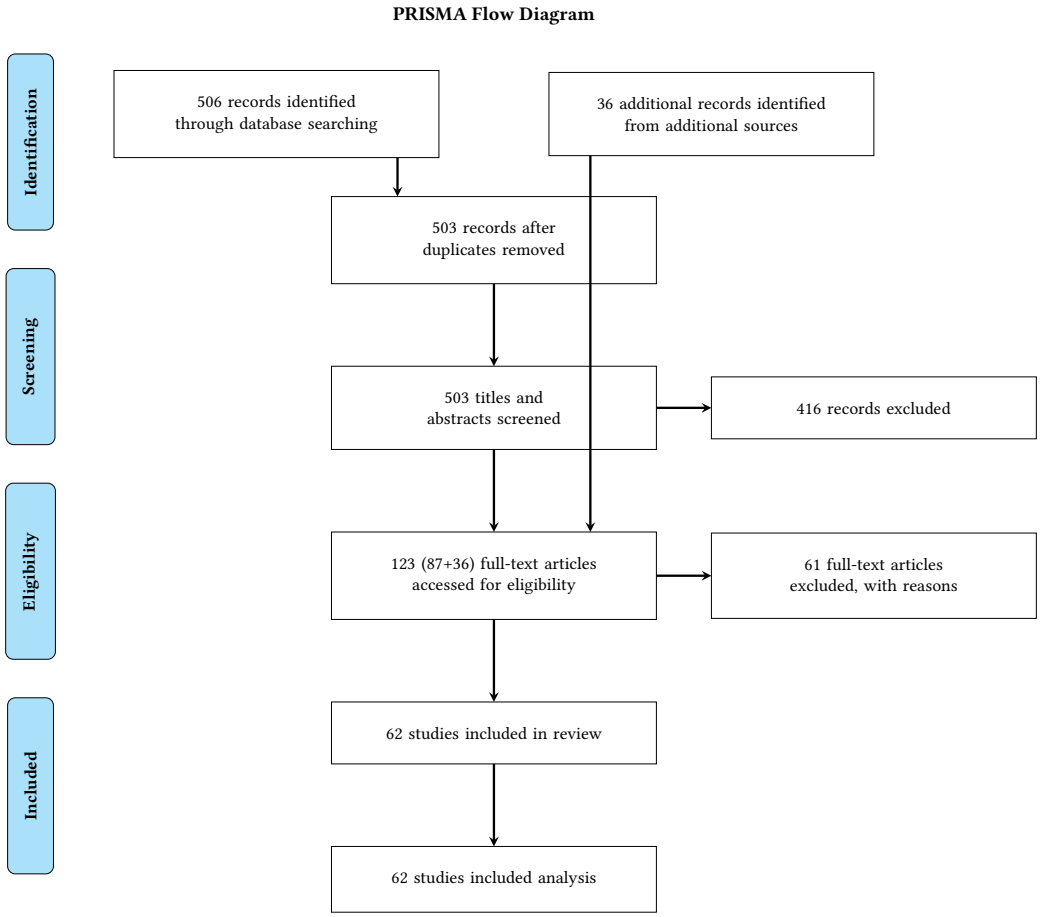

Fig. 1. PRISMA Diagram

12 random studies. In this second round with the revised scheme, inter-rater reliability was perfect (Cohen's $\kappa = 1.0$), and no further revisions were made to the coding scheme for behaviors.

The final scheme included four main types of problematic behavior, each with some variations which are discussed in the results section. Generally, coders found that each of the audits primarily addressed one of four major types of problematic behavior:

- **Discrimination**: the algorithm disparately treats or disparately impacts people on the basis of their race, age, gender, location, socioeconomic status, and/or intersectional identity. For example, an algorithm implicated in discrimination may systematically favor people who identify as males, or reinforce harmful stereotypes about elderly people.
- **Distortion**: the algorithm presents media that distorts or obscures an underlying reality. For example, an algorithm implicated in distortion may favor content from a given political perspective, hyper-personalize output for different users, change its output frequently and without good reason, or provide misleading information to users.
- **Exploitation**: the algorithm inappropriately uses content from other sources and/or sensitive personal information from people. For example, an algorithm implicated in exploitation may

infer sensitive personal information from users without proper consent, or feature content from an outside source without attribution.

- **Misjudgement**: the algorithm makes incorrect predictions or classifications. Notably, misjudgement can often lead to discrimination, distortion, and/or exploitation, but some studies in the review focused on this initial error of misjudgement without exploring second-order problematic effects. An algorithm implicated in misjudgement may incorrectly classify a user's employment status or mislabel a piece of political news as being primarily about sports, for example.

Studies that addressed more than one of the above behaviors were coded based on their primary focus. The 62 audit studies and their codes are available in a public repository[2] and will be updated as an effort to account for overlooked studies, as well as studies that were published after this review.

4 RESULTS FOR RQ1: PREVIOUS ALGORITHM AUDITS

This section provides an overview of the findings for RQ1, "what kinds of problematic machine behavior have been exposed by previous algorithm audits?" It organizes results according to behavior:

- Discrimination (Section 4.2)
- Distortion (Section 4.3)
- Exploitation (Section 4.4)
- Misjudgement (Section 4.5)

4.1 Overview

Thematic analysis produced a taxonomy with four main types of problematic machine behavior: discrimination, distortion, exploitation, and misjudgement. We propose this taxonomy as a potential high-level framework for the different harms caused by algorithmic systems, although future audit work will likely diagnose additional behaviors that could expand this taxonomy.

Discrimination is an important central focus of algorithm audit studies, and 21 studies in our review specifically audited for discrimination. As noted by Sandvig et al. [140], while "the word 'audit' may evoke financial accounting, the original audit studies were developed by government economists to detect racial discrimination in housing." Racial discrimination has continued to be a key focus of algorithm audits, and of critical algorithm studies in general, as scholars have introduced different terms for racially discriminatory algorithms: "coded gaze" (by Buolamwini [21]), "digital poorhouse" (by Eubanks [56]), "algorithms of oppression" (by Noble [124]), "algorithmic inequity" (by Wachter-Boettcher [157]), and "the New Jim Code" (by Benjamin [17]), among others. In addition to racial discrimination, discrimination can also occur on the basis of age, sex, gender, location, socioeconomic status, and/or intersectional identities, all of which have been the subject of algorithm audits, as discussed in Section 4.2.

While discrimination was the central focus of algorithm audits reviewed, overall, more studies (N=29) in our review focused on distortion (Section 4.3), especially for search algorithms (N=18) and recommender systems (N=7). Broadly, these audit studies scrutinize curated media from algorithmic systems, checking for distortion in terms of political partisanship, false information, suppression of sources, and more. The third behavior was exploitation. While many scholars have expressed concern over exploitation on algorithmic systems, especially in terms of personal information (e.g. see Zuboff [165], Couldry and Mejias [31], and West [159]), only five studies addressed how algorithmic systems exploit media content and/or personal information (Section 4.4). Finally, the

[2] https://github.com/comp-journalism/list-of-algorithm-audits/

Year	Discrimination	Distortion	Exploitation	Misjudgement	Total
2012	1				1
2013	3	1			4
2014	1				1
2015	1	1	1		3
2016	1	1			2
2017	2	3	2	1	8
2018	3	9	1	1	14
2019	7	10	1	4	22
2020*	2*	4*		1*	7*
Total	21	29	5	7	62

Table 2. Number of studies, by year and behavior.

Domain	Discrimination	Distortion	Exploitation	Misjudgement	Total
Search	5	18	2		25
Advertising	3	2	3	4	12
Recommendation	1	7			8
Pricing	5				5
Vision	5				5
Criminal Justice	1			3	4
Language Processing	1	1			2
Mapping		1			1
Total	21	29	5	7	62

Table 3. Number of studies, by domain and behavior.

more general "misjudgement" behavior was the focus of seven studies in our review (Section 4.5), all of which audited algorithmic systems for errant decisions or classifications, without exploring more specific harms related to discrimination, distortion, or exploitation. The following subsections discuss all four behaviors in more detail.

4.2 Discrimination

This review defined *discrimination* as an algorithmic system disparately treating or disparately impacting people with respect to their race, age, sex, gender, location, socioeconomic status, and/or intersectional identity. This section details how discrimination manifested in advertising algorithms, computer vision algorithms, search algorithms, and pricing algorithms. Importantly, algorithmic discrimination can happen through allocation or through representation [13].

Discrimination is often thought of in terms of allocation, and especially for "a narrow set of goods, namely rights, opportunities, and resources" [79]. However, discrimination can also play out through representational disparities, as evidenced in the work by Noble [123] and Kay et al. [88] and further discussed in 4.2.3. Representational discrimination is sometimes referred to as "bias," however, that word has vastly different connotations and definitions in statistics (e.g. estimator

bias), machine learning (e.g. bias term), psychology (e.g. cognitive bias), physics (e.g. tape bias), and other fields related to computing. This review therefore refrains from using the word "bias" whenever possible, recognizing the fact that public-facing algorithms exist in real-world social contexts where bias leads to harmful discrimination.

4.2.1 Discrimination in Advertising. Researchers have mainly found discrimination in advertising with regards to *who* is searching. Datta et al. [36] showed that protected classes directly affect ad targeting on Google, and later used this evidence to argue that platforms' immunity to Title VII should be reconsidered [35]. Corroborating these findings, Cabañas et al. [23] showed that some targeting features explicitly rely on protected and/or sensitive personal attributes such as sexual orientation, political preferences, and religious beliefs. Most recently, Asplund et al. [8] conducted a sock-puppet audit which yielded evidence of "differential treatment in the number and type of housing ads served based on the user's race, as well as bias in property recommendations based on the user's gender." For example, sock puppets emulating Caucasians saw more housing-related advertisements, while sock puppets emulating African Americans saw more ads for predatory rent-to-own programs [8]. Algorithmic advertising has also been shown to discriminate on the basis of sexual and/or gender identity. Lambrecht and Tucker [98] found that Facebook advertisements for STEM career ads, even with gender-neutral targeting and bidding options, reached fewer women than men.

An earlier audit in our review showed an additional way that advertising can discriminate through representational harms, regardless of who is searching. The study, conducted by Sweeney [146], showed through case studies and a large-scale analysis of 2,184 name searches that "ads suggesting arrest tend to appear with names associated with blacks, and neutral ads or no ads appear with names associated with whites." We discuss this phenomenon further in the search section, in light of audits by Noble [123]. Overall, algorithmic advertising systems often enable harmful discrimination which poses significant risks to users' safety, privacy, and economic opportunity.

4.2.2 Discrimination in Vision. The audits reviewed show computer vision algorithms discriminating on the basis of sex, race, geography, and intersectional identities. Many audits in this area stem from the seminal "gender shades" study by Buolamwini [21], which showed that facial analysis on darker-skinned females performed significantly worse than on light-skinned males. The study demonstrated that disparities often arise from imbalanced training data, for example, benchmark datasets exhibited "overrepresentation of lighter males, underrepresentation of darker females, and underrepresentation of darker individuals in general" [21]. Because computer vision training data often exhibits imbalance, and imbalanced data often creates discriminatory performance disparities, researchers have audited a variety of systems for these problematic disparities.

In the case of sexual discrimination, performance disparities occur when a system exhibits higher error rates for a given sexual identity. Similar to findings by Buolamwini [21], Kyriakou et al. [96] found that several commercial systems for image tagging perform worse on female faces. In terms of representation disparities, one of the earliest algorithm audits by Kay et al. [88] found that women were consistently underrepresented in Google's image search results compared to metrics from the U.S. Bureau of Labor Statistics.

Racial discrimination in photographic technology has been present for decades in cinema and photography (e.g. [99], see also chapter 3 in *Race After Technology*, by Benjamin [17]), and in some ways, this appears to continue with computer vision algorithms. In a well-publicized incident involving the Google Photos app, one headline summarized "Google Mistakenly Tags Black People as 'Gorillas'" [14]. While not always so overt and egregious, formal audits of computer vision systems have consistently surfaced similar discrimination. An audit of five commercial image tagging systems [12, 96] found worse performance on people with dark skin, a finding consistent

Reference	Year	Method	Domain	Class	Organization
Mikians et al. [111]	2012	Sock puppets	Pricing	Class Geography	Google Bing
Mikians et al. [112]	2013	Crowdsourcing	Pricing	Geography	Amazon Hotels.com Other websites
Sweeney [146]	2013	Direct scrape	Advertising	Race	Google
Noble [123]	2013	Direct scrape	Search	Race	Google
Hannak et al. [74]	2014	Crowdsourcing Sock puppets	Pricing	Generic Device	WalMart Expedia Orbitz
Kay et al. [88]	2015	Direct scrape	Search	Gender	Google
Chen et al. [27]	2016	Direct scrape	Pricing	Generic	Amazon
Eriksson and Johansson [54]	2017	Sock puppets	Recommendatio	Gender	Spotify
Hannák et al. [72]	2017	Direct scrape	Search	Intersectional	TaskRabbit Fiverr
Hupperich et al. [81]	2018	Sock puppets	Pricing	Geography	Booking.com Hotels.com Other websites
Chen et al. [26]	2018	Direct scrape	Search	Gender	Indeed Monster CareerBuilder
Buolamwini [21]	2018	Direct scrape	Vision	Intersectional	Microsoft IBM Face++
Raji and Buolamwini [132]	2019	Carrier puppet	Vision	Intersectional	Amazon Kairos
Barlas et al. [12]	2019	Direct scrape	Vision	Gender Race	(Same as below)
Kyriakou et al. [96]	2019	Direct scrape	Vision	Gender Race	(Same as below)
DeVries et al. [41]	2019	Direct scrape	Vision	Geography	Amazon Google IBM Microsoft Clarifai
Lambrecht and Tucker [98]	2019	Carrier puppet	Advertising	Gender	Facebook Google Instagram Twitter
Geyik et al. [64]	2019	Code	Search	Gender	LinkedIn
Tolan et al. [148]	2019	Code	Criminal Justice	Gender Race	Catalonia, Spain
Asplund et al. [8]	2020	Sock puppets	Advertising	Gender Race	Google
Sap et al. [141]	2020	Direct scrape	Language Processing	Race	Jigsaw (Google)

Table 4. Papers included in the review that audited for discrimination. Sorted by year.

with the original gender shades study [21] as well as a follow-up study that found performance disparities in three other commercial systems [132].

More recently, some studies have found computer vision algorithms can discriminate on the basis of geography. As with sexual and racial discrimination, this may arise from skewed training data, in particular, many popular training datasets consist of mostly images from North America and Europe, with fewer images from the global south. This imbalanced training data leads to discriminatory performance disparities, as a recent study noted that "object-classification accuracy in recognizing household items is substantially higher for high-income households than it is for low-income households" [41].

Finally, the original gender shades study by Buolamwini [21] demonstrated the importance of considering intersectional identities when identifying discrimination. The three commercial systems audited in the gender shades study exhibited *some* disparities between male faces and female faces, as well as between darker-skinned and lighter-skinned faces, however, disparities amplified when they were considered simultaneously. For example, the Face++ error rate among lighter-skinned males was 0.8%, while the error rate among darker-skinned females was 34.5%. A follow-up study [132] also used intersectional subgroups to audit for potential improvements in commercial vision algorithms, measuring corporations' success by their ability "to significantly reduce error gaps in the intersectional performance of their commercial APIs" [132]. Crenshaw [33] coined the term intersectionality for exactly these kinds of disparities, and Ogbonnaya-Ogburu et al. [125] recently discussed their importance to HCI research: "each person represents a unique and even potentially conflicting set of overlapping identities... we must be anti-essentialist and incorporate an understanding that these intersecting identities create unique contexts."

4.2.3 Discrimination in Search. Within our review, one of the earliest and most important audits of discrimination in search algorithms was performed by Noble [123] in 2013. The audit vividly demonstrated how search engines can reinforce harmful stereotypes with respect to intersectional identities: in a Google search for "black girls," five of the top ten results were sexualized or pornified, while only three of the top ten results were "blogs focused on aspects of social or cultural life for Black women and girls." Noble developed these findings of representational discrimination in a book about search engines and racism, titled "Algorithms of Oppression" [124].

Another important work focused on representational discrimination is the aforementioned audit by Kay et al. [88] focused on Google's image search. As mentioned in the context of discriminatory computer vision algorithms, the audit found that image search results exaggerated gender stereotypes when searching for occupations, such as "nurse" and "doctor," compared to baselines from the U.S. Bureau of Labor and Statistics. The authors suggest that this behavior "risks reinforcing or even increasing perceptions of actual gender segregation in careers," [88] especially given that users focus on highly-ranked results [69].

While Google provides search results from across the internet, some search algorithms are specific to content within a given platform. Discrimination in these search algorithms can lead to egregious allocative harms on platforms that influence employment opportunities, and audit studies have directed their attention accordingly. Hannák et al. [72] investigated discrimination in search results on TaskRabbit and Fiverr, finding evidence for discrimination on the basis of gender and race. Namely, the audit found that "workers perceived to be Black tend to be shown at lower ranks relative to those perceived to be White" [72]. Another study by Chen et al. [26] audited Indeed, Monster, and CareerBuilder, also surfacing evidence of gender discrimination ("overall, men rank higher than women with equivalent features"). On a more positive note, Geyik et al. [64] conducted a similar audit of discrimination in LinkedIn talent search, and demonstrated a method for improving gender representation on the platform based on notions of equal opportunity and demographic parity. In

A/B testing, their fairness-aware search algorithm produced gender-representative search results for 95% of all queries, compared to 50% for the original search algorithm, and the fairness-aware search algorithm was then deployed to LinkedIn Recruiter users worldwide.

4.2.4 Discrimination in Pricing. Lastly, some studies have audited for discrimination in pricing algorithms. E-commerce websites such as Orbitz, Amazon, and Home Depot sometimes show different prices to different people for the same item – a phenomenon known explicitly as *price discrimination*. This behavior presents clear potential for harm, by disproportionately burdening some populations while systematically advantaging others. As Jiang et al. [85] point out, ethical pitfalls in price discrimination surface even in random A/B/N tests that do not explicitly discriminate on the basis of race or class, for example. And while price discrimination is often legal, laws in the United States prohibit similar kinds of opportunistic price changes [50], making algorithmic price discrimination an important behavior to capture and understand through algorithm audits. To this end, Mikians et al. [111] conducted the earliest audit in our review, with similar follow-up studies conducted by Mikians et al. [113], Hannak et al. [74], Chen et al. [27], and Hupperich et al. [81].

These price discrimination audits were primarily concerned with *detecting* price discrimination and demonstrating its extent, and they offer fewer details about the specific discriminatory behavior. For example, the audit by Chen et al. [27] focused on proving the existence of dynamic pricing, noting that dynamic price discrimination threatens to disrupt the market, exacerbate inequalities between sellers, and create a confusing shopping experience for customers. In other words, price discrimination causes harms even if it is not definitively on the basis of race, age, sex, gender, location, socioeconomic status, and/or intersectional identity. Still, some audits provide more specific analyses of such discrimination, such as the study by Mikians et al. [111] which found evidence of discrimination on the basis of location and class (e.g. "affluent customers" versus "budget conscious customers").

4.3 Distortion

The *distortion* behavior refers to an algorithm presenting media in a way that distorts or obscures an underlying reality. In many ways, distortion is analogous to discrimination, but rather than discriminating against different groups of people, the algorithm discriminates against different groups of media. This may involve political partisanship, unpredictable fluctuation, dissemination of misinformation, and/or hyper-personalized content that can lead to "echo chambers." Distortion is mainly of interest with regards to search algorithms and recommendation algorithms.

4.3.1 Distortion in Search. Search audits were common among the papers reviewed, and many of them cited the "politics of search engines" [84] and/or the "search engine manipulation effect" (SEME) [53] as motivation. The SEME phenomenon involves a search engine shifting voting behavior in accordance with presented search results. For example, a large-scale experiment found that a search engine led to a 39% increase in the number of subjects "who indicated that they would vote for the candidate who was favored by their search rankings" [53]. SEME suggests that search engines exert some degree of influence over voting behavior and democratic outcomes, which appears to have guided search audits to focus on partisanship. After all, experimental evidence for the SEME prompts an urgent question: do real-world search engine results favor certain political candidates?

The audits reviewed occasionally surfaced evidence for small but statistically significant partisan leanings. One of the earliest studies showing such evidence was a set of case studies by Diakopoulos et al. [42], which found that Google searches for Republican candidates yielded a higher proportion of negative articles compared to searches for Democratic candidates, and overall the partisanship of search results leaned left. Several studies have replicated this small but significant left-leaning

Reference	Year	Method	Domain	Organization
Hannák et al. [73]	2013	Crowdsourcing Sock puppets	Search	Google
Kliman-Silver et al. [93]	2015	Sock puppets	Search	Google
Soeller et al. [144]	2016	Sock puppets	Mapping	Google
Eslami et al. [55]	2017	Direct scrape	Search	Booking.com
Kulshrestha et al. [94]	2017	Direct scrape	Search	Twitter Google
Snickars [143]	2017	Sock puppets	Recommendation	Spotify
Robertson et al. [136]	2018	Crowdsourcing	Search	Google
Andreou et al. [6]	2018	Crowdsourcing	Advertising	Facebook
Bechmann and Nielbo [16]	2018	Crowdsourcing	Recommendation	Facebook
Puschmann [129]	2018	Crowdsourcing	Search	Google
Chakraborty and Ganguly [25]	2018	Sock puppets	Recommendation	New York Times
Courtois et al. [32]	2018	Crowdsourcing	Search	Google
Weber and Kosterich [158]	2018	Code	Recommendation	Not specified
Kulshrestha et al. [95]	2018	Direct scrape	Search	Google Twitter
Rieder et al. [135]	2018	Direct scrape	Search	YouTube
Lurie and Mustafaraj [103]	2019	Direct scrape	Recommendation	Google
Trielli and Diakopoulos [149]	2019	Direct scrape	Search	Google
Moe [115]	2019	Direct scrape	Search	YouTube
Metaxa et al. [110]	2019	Direct scrape	Search	Google
Jiang et al. [86]	2019	Direct scrape	Language Processing	YouTube
Hu et al. [80]	2019	Direct scrape	Search	Google
Robertson et al. [137]	2019	Direct scrape	Search	Google Bing
Ali et al. [5]	2019	Carrier puppet	Advertising	Facebook
Cano-Orón [24]	2019	Direct scrape	Search	Google
Lai and Luczak-Roesch [97]	2019	Crowdsourcing	Search	Google
Hussein et al. [82]	2020	Sock puppets	Search	YouTube
Ribeiro et al. [134]	2020	Direct scrape	Recommendation	YouTube
Bandy and Diakopoulos [11]	2020	Direct scrape Crowdsourcing	Recommendation	Apple
Fischer et al. [59]	2020	Direct scrape	Search	Google

Table 5. Papers included in the review that audited for distortion. Sorted by year.

partisanship on Google, especially in terms of news items [136, 149]. However, the result is open to many different interpretations. Also, some studies have found conflicting results. For example, while Robertson et al. [136] found news items exhibited a slight left-leaning partisanship, they also found that "Google's ranking algorithm shifted the average lean of [Search Engine Results Pages] slightly to the right." Furthermore, Metaxa et al. [110] observed "virtually no difference" in the distribution of source partisanship between queries for Democrats versus Republicans. Shifting

the focus beyond just the "blue links," Hu et al. [80] audited the "snippet" text that appears below the links on search result pages. They found that partisan cues in the snippet text tend to amplify the partisanship of the original web page, an effect that was consistent across query topics, left- and right-leaning queries, and different types of web pages (e.g. social media, news, sports).

Even in the absence of partisanship, search algorithms can distort media by limiting exposure diversity, which runs counter to the notion of "the Web as a public good," as argued by Introna and Nissenbaum [84]. This can occur through overall source concentration, and through hyper-personalization that may create "echo chambers." In terms of overall source concentration, an audit by Trielli and Diakopoulos [149] focused on the "Top Stories" carousel of Google's results page, finding a high concentration of sources (the top five sources of impressions per query were CNN, New York Times, The Verge, The Guardian, and Washington Post). The substantial dependence on mainstream media outlets has also been evidenced in studies of medical queries [24], political queries in Dutch-speaking regions [32], and political queries in the United States [95, 136]. An audit by Lurie and Mustafaraj [103] provides an exception to this trend in source concentration, suggesting that Google "uses the 3rd position for exploration, providing users with unfamiliar sources," thus promoting lesser-known publishers. More recently, Fischer et al. [59] found that national outlets still dominate Google search results overall, suggesting that the search algorithm "may be diverting web traffic and desperately needed advertising dollars away from local news" [59].

Exposure diversity can also be limited through "echo chambers" that may arise when a search algorithm provides hyper-personalized results to users [19]. The earliest personalization audit reviewed was by Hannák et al. [73] in 2013, which found that on average only 11.7% of results were personalized, although personalization occured more often in some query categories ("gadgets," "places," and "politics"). A later study by Kliman-Silver et al. [93] showed that geolocation had the greatest effect on personalized results.

Generally, studies have shown limited evidence for the "echo chambers" that some scholars have feared. A study of 350 Dutch users found no substantial differences due to personalization [32] across 27 socio-political queries. An audit by Robertson et al. [136] measured personalization on all different components of a search engine including the "people also ask for" box, "tweet cards," and "knowledge components." The full-page audit "found negligible or non-significant differences between the [Search Engine Results Pages]" of personalized (logged in to Google) and non-personalized (incognito) windows. These findings could be limited to socio-political queries – as the earliest personalization audit showed, the degree of personalization varies across query categories [73]. Rather than queries for political candidates and topics, one study in New Zealand by Lai and Luczak-Roesch [97] used queries that public officials "might perform in the course of their everyday work duties." Asking 30 participants to rate the relevance of personalized and non-personalized results, their findings suggest that up to 20% of relevant results were removed due to personalization.

Lastly, in addition to partisanship and hyper-personalization, a search algorithm may be culpable of distortion if it disseminates low-quality media such as false information, junk news, or other opportunistic content. Hussein et al. [82] explore this phenomenon on YouTube, finding that the search algorithm exhibited a "rabbit hole" effect: "people watching videos promoting misinformation are presented with more such videos in the search results." The effect was not present in search results for queries related to vaccines, which is notable given the high stakes of health information illustrated by De Choudhury et al. [38].

4.3.2 Distortion in Recommendation. Distortion in recommendation is similar to distortion in search, and presents a similar level of public impact. Examples of public-facing recommender

systems include news recommenders such as Google News, song recommenders such as Spotify radio, and product recommenders such as the "customers also bought" feature on Amazon. Audits of recommender algorithms have focused on hyper-personalization and echo chamber effects, as well as source concentration.

As with internet search engines, recommender systems have the potential to create echo chambers by recommending users a narrow, specific set of content. In a typical depiction of the phenomenon, a user reads recommended news articles with a specific viewpoint, forming a feedback loop with the recommender which continually reinforces that viewpoint. But as with search algorithms, the echo chamber phenomenon has failed to materialize in numerous studies examining Google News recommendations [70, 121], Facebook News Feed recommendations [10, 16], and "Trending Stories" recommendations in Apple News [11]. One study of the New York Times website [25] found evidence for a potential feedback loop on the "Recommendations for You" page. However, the study focused on a single, explicitly-personalized page of the website, and did not examine ideological or topical differences in the content. Lastly, personalization effects were also minimal in an audit of Spotify's radio recommendations – not even liking all songs or disliking all songs had a significant impact on recommended songs [143].

While audits have rarely surfaced significant echo chamber effects, they have found that recommendation algorithms often exhibit problematic source concentration. The recent study by Bandy and Diakopoulos [11] found that the top three sources of "Trending Stories" in Apple News accounted for 45% of all recommendations. Similar source concentration occurs on Google News. Echoing early hints from Haim et al. [70], Nechushtai and Lewis [121] found that 35% of all recommended Google News stories came from just three sources. As noted by Introna and Nissenbaum [84] "systematically giving prominence to some at the expense of others" in these contexts is often inherently inequitable and problematic.

4.4 Exploitation

This study considered *exploitation* to be the inappropriate use of content from other sources and/or sensitive personal information from people. While discrimination applies to people and distortion applies to media, algorithms can exploit either people or media. For example, an algorithm implicated in exploitation may infer sensitive personal information from users without proper consent, or feature content from an outside source without attribution. This review found five studies that focused on exploitation in algorithmic systems.

4.4.1 Exploitation in search. Search engines face heightened scrutiny and audit attention for their reliance on – and potential exploitation of – outside content. Indeed, almost all studies reviewed in this area present evidence that search engines depend heavily upon user-generated content (UGC) and journalism content, which presents a number of potential problems entwined with copyright laws and monopoly power [47].

Notably, audits show that user-generated content from websites like Wikipedia, Twitter, and StackOverflow improve click-through rates on Google [108]. Considering that UGC adds value to Google search in the form of improved click-through rates, Vincent et al. [156] conducted an audit to quantify the extent of Google's UGC reliance, finding that Wikipedia content, created through voluntary labor, appears in 81% of results pages for queries trending in Google Trends, and 90% of results pages for queries about controversial topics (e.g. "Death Penalty," "Under God in the Pledge," "Israeli-Palestinian Conflict"). The study also found that the "News Carousel" appeared in the majority popular and trending queries, often in the top three Google results. As noted in the distortion section (4.3), Google's reliance on news outlets is highly concentrated [95, 136, 149], and their results often exclude local news [59].

Reference	Year	Method	Domain	Organization
Datta et al. [36]	2015	Sock puppets	Advertising	Google
Mähler and Vonderau [104]	2017	Direct scrape	Advertising	Spotify
Mcmahon et al. [108]	2017	Crowdsourcing	Search	Google
Cabañas et al. [23]	2018	Crowdsourcing	Advertising	Facebook
Vincent et al. [156]	2019	Direct scrape	Search	Google

Table 6. Papers included in the review that audited for exploitation. Sorted by year.

4.4.2 Exploitation in advertising. Advertising algorithms use sophisticated experimentation and optimization techniques to deliver personalized advertisements, in some cases leading them to exploit sensitive personal information. Distinct from algorithm audits for discrimination, audits for exploitation focus on the appropriation of sensitive personal information, especially without users' consent. A clear example of this was surfaced in the audit by Datta et al. [36]: agents with browsing history related to substance abuse were shown a different distribution of ads, but this trait was not shown in Google's Ad Settings. In other words, beyond discriminatory targeting (male agents were more likely than female agents to see ads for high paying jobs), Google's advertising algorithms also tracked and exploited private, sensitive browsing behavior to target users with ads.

With an eye toward the General Data Protection Regulation in the European Union, Cabañas et al. [23] noted that Facebook "should obtain explicit permission to process and exploit sensitive personal data" for commercial gain. Yet their audit showed that Facebook was exploiting sensitive information in their advertising algorithms, allowing advertisers to target people in categories such as "interested in homosexuality," "interested in Judaism," "interested in tobacco," and more. Again, these categories will likely lead to discrimination and other harms, though these studies highlight the problematic behavior of non-consensually inferring and exploiting these sensitive attributes for targeted advertising, as warned about by Zuboff [165], Couldry and Mejias [31], West [159], and other scholars.

4.5 Misjudgement

Misjudgement was defined generally as an algorithm making incorrect predictions or classifications. As noted earlier, misjudgement can often lead to discrimination, distortion, and/or exploitation, but some studies in our review focused on this initial error of misjudgement without exploring second-order problematic effects. Seven studies in the review focused on misjudgement, and were clustered in two areas: criminal justice and advertising.

4.5.1 Misjudgement in Criminal Justice. The use of algorithms in criminal justice presents alarming potential for harm. After ProPublica's 2016 "Machine Bias" report [91] suggested that a publicly-used recidivism prediction instrument (RPI) discriminated on the basis of race, researchers discussed and explored appropriate fairness criteria and methods for balancing error rates [29, 61]. While these theoretical and methodological explorations are not included in our review, a handful of related papers met the inclusion criteria. Two studies [48, 49] focused on recidivism prediction systems in the state of Minnesota, and detailed methods to improving prediction accuracy. Tolan et al. [148] provide a similar case study, showing that machine learning is more accurate but less fair (in terms of demographic parity and error rate balance) compared to statistical models. Also, in a code audit of a forensic DNA analysis system in New York City, Matthews et al. [106] showed how an allegedly "minor" change to the algorithm resulted in substantial data-dropping that led to more

Reference	Year	Method	Domain	Organization
Duwe and Kim [49]	2017	Direct scrape	Criminal Justice	Minnesota
Tschantz et al. [150]	2018	Crowdsourcing	Advertising	Google
Venkatadri et al. [154]	2019	Crowdsourcing	Advertising	Facebook Acxiom Epsiolon Experian Oracle (Datalogix)
Duwe [48]	2019	Direct scrape	Criminal Justice	Minnesota
Bashir et al. [15]	2019	Crowdsourcing	Advertising	Google Facebook Oracle Neilsen
Matthews et al. [106]	2019	Code	Criminal Justice	New York City
Silva et al. [142]	2020	Crowdsourcing	Advertising	Facebook

Table 7. Papers included in the review that audited for misjudgement. Sorted by year.

inaccurate results: a small change to removal criteria ended up excluding more true contributors and included more non-contributors for the DNA samples analyzed, which led to misjudgments and inaccuracies.

As we discuss later in section 5.1.1, algorithms in the context of criminal justice do not need audits and incremental improvements as much as they need holistic reform and abolition, as suggested by Benjamin [17] and others.

4.5.2 Misjudgement in Advertising. While advertising systems promise fine-grained targeting, audits show that targeted advertising algorithms often make misjudgements when inferring information about users, including demographic attributes and interest-based attributes. Tschantz et al. [150] found that among logged out users across the web, Google correctly inferred age for just 17% of females and 6% of males. Furthermore, a 2019 study by Venkatadri et al. [154] showed a substantial amount of third-party advertisers' data about users is "not at all accurate" (40% of attributes among the 183 users surveyed) – some users labeled as corporate executives were, in fact, unemployed. These inaccuracies also apply to interest-based attributes, as Bashir et al. [15] show in their audit of Facebook, Google, Oracle, and Neilsen. Asking users to review their "interest profiles" on these platforms, participants only rated 27% of the interests in their profiles as strongly relevant. Taken in concert, these audits suggest that advertisers who purchase targeted advertisements may often fail to reach their intended audience on platforms like Facebook and Google. While advertisers would find these misjudgements undesirable, notably, the same behavior may be desirable to users, who find it invasive when algorithms infer sensitive attributes without their consent [152, 164].

5 RESULTS FOR RQ2: FUTURE ALGORITHM AUDITS

While the previous section addressed how algorithm audits have diagnosed problematic machine behavior in some areas, this section points out areas that require further research attention. These areas were selected based on public impact as well as relative coverage reflected in Table 3, which tabulates studies by domain and problematic behavior. Importantly, this means that research areas in this section have received *less* attention, rather than *no* attention. This section thus includes studies that address or begin to address the topics at hand, as a starting point for future audits. As

with any literature review, these findings are subject to potential coverage biases. Some studies may have been errantly excluded due to the keyword search, mistakes in the coding process, and/or inaccessible publication.

5.1 Remaining Work: Discrimination

The audits reviewed in this paper shed significant light on the path forward for studying discrimination in algorithmic systems. Previous audits provide a number of helpful precedents for future research, and hint at potential issues in under-explored areas, which deserve further research attention.

The areas reported in Section 4.2 (discriminatory advertising, discriminatory pricing, discriminatory search, and discriminatory vision) received substantial audit attention, while still pointing to a need for further audits. As the economic engine behind large technology platforms, advertising in particular deserves further scrutiny, especially as companies like Facebook espouse changes intended to mitigate discriminatory harms (e.g. by removing the "multicultural affinity" targeting option in August 2020 [83]). For example, future algorithm audits might ask: do Facebook's measures mitigate discrimination, or obfuscate it? Even without overtly discriminatory targeting options, algorithmic targeting options that rely on income, location, and/or "lookalike audiences" may insidiously reproduce discriminatory behavior.

Similarly, price discrimination algorithms deserve further audit attention beyond merely *detecting* dynamic pricing. In their piece describing the study of "machine behavior," Rahwan et al. [131] explicitly suggest price discrimination as a research area. Following through on early evidence that pricing algorithms discriminate on the basis of location and class [111], future audits should seek evidence for more specific discrimination on the basis of race, age, sex, gender, and/or intersectional identities.

Accounting for intersectional identities is particularly important for future audits, as others such as Hoffmann [79] have pointed out, especially to avoid "fairness gerrymandering" [89]. In the words of Ogbonnaya-Ogburu et al. [125], HCI research aimed at reducing inequality (including algorithm audits) "must be anti-essentialist and incorporate an understanding that these intersecting identities create unique contexts." Ignoring intersectional identities means painting with too broad a brush, often aggregating and thus obfuscating the harmful disparities that affect people in their real-world, multifaceted identities. In other words, audit studies must center intersectionality in order to work toward algorithmic systems that "explicitly dismantle structural inequalities" [56].

This review also suggests two primary *domains* which have received negligible audit attention with respect to discrimination: language processing algorithms and recommendation algorithms. Just one study in the review (by Sap et al. [141]) audited a public-facing language processing algorithm, finding that the Perspective API tool from Jigsaw/Alphabet exhibited racial discrimination. A number of related studies find racial discrimination in *non-public* language algorithms, which were thus not included in the review. Future language processing audits may benefit from taking a similar approach to computer vision audits, targeting corporate APIs rather than generic algorithms. For example, amidst the COVID-19 pandemic, some platforms increased their reliance on automated moderation, leading to inconsistent policies and frustration for many users [105]. As demonstrated by Raji and Buolamwini [132], "publicly naming and disclosing performance results of biased AI systems" can directly improve real-world systems and benefit the public, in addition to serving the academic community.

Recommendation algorithms were the second domain that received relatively little audit attention with respect to discrimination, and they would likely benefit from following precedents in other domains. One exception was a study of discriminatory "gendered streams" on Spotify, in which Eriksson and Johansson [54] found that approximately 8 out of 10 recommended artists from Spotify

were male (the authors identified gender presentation for each artist based on pronouns, names, and images). This kind of discrimination may exist in a variety of recommendation algorithms, such as social media algorithms that recommend who to follow, apps that recommend restaurants [101], business reviews [55], and more. Algorithm audits targeting recommender systems will benefit from the rich literature in search discrimination, on platforms like LinkedIn, Monster, and CareerBuilder (e.g., [26, 64]). They will also become increasingly important as recommender systems become more common and influential in the public digital ecosystem.

5.1.1 Discriminatory Risk Assessment. Algorithmic systems used in criminal justice have been scrutinized since ProPublica's 2016 report entitled "Machine Bias" [91], which suggested that a publicly-used recidivism prediction instrument (RPI) exhibited racial discrimination. The report prompted many discussions and explorations of fairness criteria and methods for balancing error rates (e.g., [29, 61]), but surprisingly, this literature review found just one study that audited risk assessment algorithms for discrimination, authored by Tolan et al. [148]. The audit found that machine learning models "tend to discriminate against male defendants, foreigners, or people of specific national groups." For example, the study showed that machine learning models were twice as likely to incorrectly classify people who were not from Spain (the dataset came from the Spanish community of Catalonia). Notably, three other studies in the review ([48, 49, 106]) audited algorithmic systems used in criminal justice, but focused on general misjudgement rather than discrimination.

In the context of criminal justice, researchers might benefit from an abolitionist rather than a reformist approach, as articulated by Benjamin [17] and others. Auditing for statistical disparities may be important in some cases, but often, the very use of algorithms in this context is at odds with any notion of justice or fairness. This is especially true in the United States, given the historic racial discrimination in many facets of the criminal justice system [4]. For example, if an algorithm is being used to inform an already oppressive process, the focus should be on abolishing that algorithm rather than auditing and improving it. Keyes et al. [90] poignantly illustrate this point through a satirical proposal for making a "fair, accountable, and transparent" algorithm to determine "which elderly people are rendered down into a fine nutrient slurry."

5.2 Remaining Work: Distortion

Audits reviewed in this study provided important insights about the ways that search and recommendation algorithms can distort digital media. Many studies focus on partisanship, echo chamber effects, and source concentration. These audits should continue, as they provide key insights about how algorithms exercise power in the media ecosystem. Future audits may also benefit from exploring other use cases, systems, and domains.

This review corroborates the suggestion by Mustafaraj et al. [120] that existing audit literature has only accounted for a narrow set of use cases. Future audits should strive for user-centered audits, and may benefit from scoping to specific use cases. For example, such audits can build on the work of Lai and Luczak-Roesch [97] which scoped to public officials' use of search algorithms, and the work of Mustafaraj et al. [120] which outlines methods for *voter*-centric audits. Also, Fischer et al. [59] scoped their audit to local news, a topic which deserves more audit attention given the current crisis in local journalism and its threat to impacted communities (e.g. [3, 34]).

The algorithm audit literature should also expand their efforts to different systems that may exhibit distortion. Again, it is important to conduct ongoing audits of popular systems and platforms such as Google search, but researchers must also be vigilant in directing their attention to those that are new and upcoming. For example, the Google Discover recommendation algorithm has driven significant web traffic since its introduction in 2018 Willens [161], but no audits in this

review examined the system. Similarly, TikTok's rapid growth around the world [145] presents new high-impact algorithms that call for audit attention. Lastly, personalization mechanisms in digital maps present a number of problematic potential distortions, such as the misrepresentation of international borders explored by Soeller et al. [144].

Distortion should also be audited in different domains, such as advertising and language. While echo chambers have been studied in the context of search and recommendation algorithms, advertising algorithms were the subject of just one echo chamber audit in this review. In the study, Ali et al. [5] found evidence that Facebook's advertising algorithms "preferentially exposes users to political advertising." In some cases, even when advertisers targeted users with opposing viewpoints, Facebook preferred to show advertisements that aligned with the users' viewpoints. Given this initial evidence, as well as the pervasiveness of targeted advertising algorithms on today's platforms, future audits should further scrutinize these algorithms. In fact, more so than search and recommendation algorithms, advertising algorithms are *premised* on hyper-personalized content that could create echo chambers, limiting the diversity of sources and content that users encounter. Advertising algorithms may also distort media by disseminating problematic content, such as false information or prohibited advertisements. As early work in this area, Silva et al. [142] found that Facebook's advertising algorithm sometimes misjudged whether an ad was political, and showed how the system would violate political advertising laws in some jurisdictions.

Language algorithms present another important domain, with some early work signaling the need for more attention. Hu et al. [80] showed that Google's algorithm for selecting search snippet text (which shows below the blue links) can distort the page it represents, for example, 54–58% of snippets amplified partisanship, and 19–24% of snippets exhibited inverted partisanship compared to the corresponding web page. With a similar focus on partisanship, an audit of comment moderation on YouTube by Jiang et al. [86] dispelled a perception that YouTube's comment moderation practices were politically biased (rather, comments were more likely to be moderated if they were extremist, contained false content, or were posted after a video was fact-checked). But partisanship is not the only potentially distorting behavior, as some language systems can present misleading information. In an audit of Facebook's "why am I seeing this ad?" feature, Andreou et al. [6] found that Facebook's linguistic explanations for targeted ads were often incomplete and sometimes misleading (similar to Google's lack of transparency demonstrated by Datta et al. [36]). Future audits should continue exploring these phenomena – partisan language and misleading language – in algorithms for various types of text summarization and language generation.

5.3 Remaining Work: Exploitation

Given the urgent concerns around exploitation on algorithmic systems, especially in terms of personal information, further audits of algorithmic exploitation could provide important clarifications. Considering the arguments and concerns articulated in concepts such as "surveillance capitalism" (by Zuboff [165]), "data colonialism" (by Couldry and Mejias [31]), "data capitalism" (by West [159]), and others, researchers may benefit from focused audits that characterize specific types of exploitation and consequent harms inflicted by algorithmic systems. Given the methodological challenges associated with auditing exploitation, future audits in this area should look to successful early work by Datta et al. [36] and related projects in advertising (e.g., [23, 154]).

One question in particular looms over this topic: how much money do companies make from exploitation? Some early work may help guide future audits, addressing the value of personal information and other types of content like news media.

In terms of personal information, Cabañas et al. [22] explored how Facebook users perceived the economic value of their data. The study found that only 23% of participants provided a close answer to the actual value of their personal data (about $1 per month, estimated using Facebook's

quarterly ad revenue and monthly active users in Q2 2016). Users' personal data also adds value to recommender systems. Vincent et al. [155] show that removing users' data "can bring recommender accuracy down to the levels of early recommender systems from 1999," thus, recommendation algorithms "are in fact a highly cooperative project between the public and companies." Even if the "data labor" used to power recommendation algorithms provides only a marginal increase in performance, it provides a massive increase in value. Netflix estimates a 2-4x increase in engagement, and about $1 billion in revenue per year [68]. Future algorithm audits should seek further clarification as to how platforms profit from exploiting users' personal information.

Similar clarifications are needed as to how platforms profit from news media and other content from journalism organizations. The News Media Alliance addressed this question in 2019, but extrapolated revenue share from a 2008 statistic, making the estimate of Google's revenue from news publishers ($4.7 billion) questionable. Also, Google announced in 2020 it will pay publishers more than $1 billion over the next three years to license news content. Moving forward, the ongoing debates about platforms and news media may benefit from audit studies that analyze the distribution of these funds, how equitable the distributions are, and how publishers are impacted – especially struggling local publishers which may be particularly disadvantaged by Google [59].

5.4 Remaining Work: Misjudgement

Future audits exploring misjudgement should generally opt to address more specific harms. Errant algorithms are often problematic merely by virtue of being errant, however, as detailed in the sections about discrimination, distortion, and exploitation, there is often a more specific problematic behavior associated with a simple error or misjudgement.

For example, a number of studies in our review addressed misjudgements in targeted advertising platforms, suggesting that advertising algorithms hold inaccurate targeting information about many users. In future audits, researchers may explore how these misjudgements specifically harm users and advertisers. Users may be unsettled when encountering a distorted characterization of their interests in the form of inaccurately targeted advertisements (as evidenced by Ur et al. [152]) – in fact, a Pew Research Study [128] found that "about half of Facebook users say they are not comfortable when they see how the platform categorizes them." Also, advertisers may be alarmed to know they spent money on inaccurately targeted advertisements, for example, aiming to reach corporate executives but in actuality reaching people experiencing unemployment (an error surfaced twice by Venkatadri et al. [154]). In addition to simply identifying the initial error, future audits should focus on these potential harms to users, advertisers, and other potential stakeholders.

5.5 Methods and Organizations for Future Audits

5.5.1 Methods. While this review is primarily organized by behavior and domain, it is also important to note some promising areas for future work based on methods used and organizations audited. In terms of methods, Table 8a shows that direct scraping was the most-used audit method (N=30) in the papers we reviewed (N=62). As noted by Sandvig et al. [140], a key limitation with scraping audits is that there is no randomization or manipulation, making them a primarily useful for descriptive tasks. The audits that used crowdsourcing in our review (N=16) did not suffer from this limitation, since crowdsourced audits provide real-world data and can even allow for some causal inferences. While the carrier puppet method was used less often (N=3), this may be desirable given that such audits may disrupt real-world systems in the process of auditing.

Code auditing appears to be under-explored. This may be due to the fact that many algorithms of interest to researchers are proprietary and opaque, with no code to audit in the first place. However, code audits may be useful for some high-impact and open-source systems, such as applications by

Method	Papers Reviewed
Direct scrape	30
Crowdsourcing	16
Sock puppets	12
Code audit	4
Carrier puppet	3

Organization	Papers Reviewed
Google	30
Facebook	8
Amazon	5
YouTube	5
State/Government	4
Twitter	3
Spotify	3
LinkedIn	1
Instagram	1
Apple	1

Table 8. Studies reviewed, tabulated by method (left) and organization (right).

Wikipedia and DuckDuckGo.[3] Generally, however, researchers can expect to continue using other methods when auditing systems from Facebook, Google, Amazon, and other corporations, since these systems are often proprietary and are unlikely to be open-sourced.

5.5.2 Organizations. Based on our review, the organizations audited have received skewed attention. Google has been the subject of the most audits (N=30), followed by Facebook (N=8), Amazon (N=5), and YouTube (N=5), as shown in Table 8. The focus on Google is warranted given the influence of its search algorithms. At the same time, there also appears to be a gap in audit attention toward other influential organizations, especially social media platforms such as Facebook, Twitter, Instagram, and LinkedIn. This may be due in part to the legal challenges with scraping these platforms, noted by Sandvig et al. [140]. With these challenges in mind, future audits may benefit from leveraging real-world data from end users, namely through crowdsourcing. Audits may also benefit from new data sources, such as Twitter's new academic access[4] and Facebook's dataset of all URLs shared more than 100 times [109].

5.5.3 Replication. Replication also presents an important area for future algorithm audits. While the findings in previous audit work are important and compelling, they may become quickly outdated given the frequent updates to algorithmic systems. Results may also change if researchers use different methods (e.g. crowdsourcing instead of sock puppets) or explore algorithms from different organizations (e.g. Bing's search algorithm instead of Google's search algorithm). These efforts would coincide with a broader push in computing research for replication studies [28], demonstrated clearly in the "repliCHI" workshop [162].

Fortunately, researchers have already demonstrated a commitment to open science practices that will help enable replication audits. To name a few examples, Buolamwini [21] released a facial analysis dataset balanced by gender and skin type, Robertson et al. [138] released scraping tools for recursive algorithm interrogation (RAI) and scraping web searches, and Barlas et al. [12] published their "SocialB(eye)as" dataset for exploring biases in computer vision algorithms.

[3] https://github.com/wikimedia, https://github.com/duckduckgo
[4] https://developer.twitter.com/en/solutions/academic-research

5.6 Important Ingredients for Future Audits

To help guide future work in algorithm auditing, here we note the common ingredients for a successful audit that clearly demonstrates how problematic machine behavior affects the public. While different systems, different contexts, and different research questions prompt different methods, we identify at least four important common ingredients for a successful algorithm audit: a public-facing algorithmic system, suspected problematic behavior(s), a compelling baseline, and metric(s) to quantify the problematic behavior(s).

5.6.1 Public-Facing Algorithmic System. Given the growing number of opaque algorithmic systems that are already deployed to the public, audits that focus on these public-facing systems are especially important. In the absence of "an FDA for algorithms" [151], researchers can continue conducting audits to clarify the safety of public-facing algorithms, and help determine if and how they serve the public good. In some cases, public-facing does not mean the system is publicly-deployed, only that it holds imminent or potential harms to the public.

5.6.2 Suspected Problematic Behavior(s). With a public-facing algorithmic system identified, an audit study should define suspected problematic behavior(s), asking "how [social problems] manifest in technical systems" [2] in the most specific terms possible. Exploratory analysis can be helpful in some cases, but algorithmic systems tend to lend themselves to specific behavioral patterns, which should be the focus of audit studies.

5.6.3 A Compelling Baseline. Audits should establish compelling baselines that convincingly answer the question: "compared to what?" For example, one of the earliest audits reviewed compared occupational gender representation compared to statistics from the U.S. Census Bureau [88]. In other cases, such as the LinkedIn Talent Search audit by Geyik et al. [64] used *equality of opportunity* as a guiding theoretical framework to define a compelling baseline. Thus, compelling baselines need not come from the real world, and in many cases they should come from theoretical frameworks such as decolonialism, reparatory justice, feminist ethics of care, and justice as fairness, among others. For example, Costanza-Chock [30] explores some of these frameworks through the lens of *Design Justice*, D'Ignazio and Klein [45] outline a theory of *Data Feminism*, and Mohamed et al. [116] introduce a framework for "Decolonial AI." All such frameworks and concepts may be useful to establishing compelling baselines beyond parity, group fairness, and other statistical metrics.

5.6.4 Metric(s) to Quantify Problematic Behavior(s). Finally, audits should present clear metrics that quantify the problematic behavior. Unfortunately, auditing presents the potential for a kind of "p-hacking" that has plagued other scientific disciplines [77, 78]. Algorithm auditors can almost always find some metric that suggests inequity, discrimination, or other problematic behavior. (On the flipside, organizations can perform "fairness gerrymandering" [89] and find some metric that suggests equity, equal treatment, etc. [61]) Thus, audit studies can benefit from compelling baselines and robust, transparent, meaningful metrics to quantify and compare against the baseline. These metrics should be crafted with great care, especially when accounting for "representational harms" that can manifest more subtly than "allocative harms" [13].

6 DISCUSSION

6.1 Limitations

As with any systematic literature review, this work exhibits some notable limitations related to inclusion and exclusion of relevant work. First, we used the Scopus database as our initial source. Studies have shown Scopus to be an expansive and inclusive source for literature reviews [58], however, a number of important audit studies did not show up in the search and were included

through other means. Keywords may have been another factor in these studies not showing up. With the recent growth of "algorithm auditing" as a more defined research area, some initial studies may have used other terms to describe the work, and may have not appeared in our review. Also, our review only included academic papers, thus excluding any audit work published as books (ex. [124]), journalism (ex. [91]), or other mediums.

The review may also have temporal and linguistic limitations. Many of the papers identified from additional sources (N=36) were published in or before 2015, suggesting that the initial corpus search suffered from a recency bias. This could have excluded older papers that did not self-identify as algorithm audit work at the time of publication. The literature review was also scoped to studies published in English. It may be true that most algorithm audits have come from English-speaking countries, and have focused on algorithmic systems in those countries. In any case, future studies would likely benefit from auditing algorithmic systems in non English-speaking countries.

6.1.1 Author Standpoint. Another limitation of this review is the limited perspective of the author. As with any qualitative work, the researcher's positionality can affect all aspects of a literature review, and it is therefore important to name some characteristics of this position [75]. The first author only speaks English and has never lived outside the United States. Furthermore, as a person who identifies and passes as a cisgender white male in the United States, he benefits from a number of social systems and structures, including capitalism, colonialism, cis-normativity, whiteness, and patriarchy. The author's privileged standpoint within these systems likely limits how the literature review addresses the oppression and violence inflicted by algorithmic systems.

6.2 Toward Algorithmic Justice

This review has shown how algorithm audits can clarify–and thus help improve–the relationship between technology and society. It is important to recognize how algorithm audits fit into broader work that addresses this relationship and aims to work toward "algorithmic justice," to use a term from Raji and Buolamwini [132]. In the words of Abebe et al. [2], "meaningful advancement toward social change is always the work of many hands." To situate algorithm auditing within a broader context, this subsection aims to identify some of the "many hands" working toward algorithmic justice within the realm of human-computer interaction research. Specifically, we note researchers working on algorithmic justice through user studies, development histories, non-public algorithm audits, and case studies.

6.2.1 User Studies. User studies can help uncover problematic algorithm behavior through indirect methods such as user surveys and interviews [51, 60, 100]. While these user studies were not included in the review, they provide an effective method for identifying potentially problematic algorithms and painting the "before-and-after picture," as Nissenbaum calls it [122], of the whole socio-technical system. For example, in interviews with Uber and Lyft drivers, Lee et al. [100] found that algorithmic assignments often did not make sense to the workers, a finding which could help guide audit studies of the assignment algorithms. In a similar vein, Díaz and Diakopoulos [43] interviewed residents about an algorithmic neighborhood walkability score, finding that many factors the residents named (ex. places of worship and transit access) were missing in the patented Walk Score algorithm.

6.2.2 Development Histories. Studies that explore the development history of an algorithmic system provide important narrative frameworks for algorithmic justice work, which can help inform algorithm audits. For example, Gillingham [67] demonstrates a muddled development of predictive tools for child protection services in New Zealand. DeVito [40] conducts a content analysis of patents, press releases, and other documents to show how values such as personal preferences

were considered in Facebook's design of the News Feed algorithm. Future algorithm audit studies may use these development histories to help select problematic behaviors or organizations worth auditing.

6.2.3 Non-public Audits. While this review did not include several audits of non-public algorithms (due to the focus on public accountability), these audits can help characterize potential and/or eminent problems with algorithmic systems. Some of these non-public audits found age-related discrimination in sentiment analysis models [44], gender discrimination in semantic representation algorithms that may be used in hiring [37], and age-related disparities in landmark detection algorithms often used for facial analysis [147]. Karakasidis and Pitoura [87] found evidence that name matching algorithms may exhibit racial discrimination, and Binns et al. [18] found that models used for language processing can exhibit performance disparities along gendered lines. This early evidence can help inform future audits of public-facing algorithmic systems that rely on or resemble these non-public algorithms.

6.2.4 Case Studies. Finally, small-scale audits can have a large impact, as demonstrated by some studies included in this review. As mentioned in the computer vision section, a single image-tagging result ("Google Mistakenly Tags Black People as 'Gorillas'" [14]) led to wide publicity and swift corporate response. Gillespie [66] authored one "small-scale" case study that vividly demonstrated the social and political complexity of internet search using just one search term ("santorum"). Another salient example is the paper by Noble [123], which exposed significant problems using a single screenshot of a Google search for "black girls." For these empirical examples, small scale is a strength rather than a weakness, as they provide extremely powerful evidence for diagnosing "how [social problems] manifest in technical systems" [2], especially in terms of representational harms [13]. As framed by Benjamin [17], case studies may surface a "glitch" pointing to a broader representational harm, which could then guide justice-oriented changes or inform larger-scale studies.

7 CONCLUSION

This systematic literature review focused on audit studies of public-facing algorithmic systems, synthesizing findings from 62 studies and identifying recurring types of problematic machine behavior. The review shows that algorithm audits have diagnosed a range of problematic machine behaviors, such as discrimination in advertising algorithms and distortion in search algorithms. These studies provide empirical evidence that the public harms of algorithmic systems are not theoretical conjectures, rather, they play out in real-world public systems affecting millions of people. The review also suggests that some areas are ripe for future algorithm audits, including addressing discrimination in terms of intersectional identities, further exploring advertising algorithms which are the economic backbone of large technology companies, and employing under-explored methods such as code auditing. Some organizations (e.g. Twitter, TikTok, LinkedIn) also deserve further attention. If future audits continue to examine public-facing algorithms, hone in on specific problematic behavior(s), and use compelling baselines, then the field of algorithm auditing can continue holding algorithmic systems accountable by diagnosing harms they pose to the public.

ACKNOWLEDGMENTS

I greatly appreciate Dr. Michelle Shumate for guiding this literature review in her class, "The Practice of Scholarship," and in a subsequent independent study. Thanks also to Daniel Trielli, who helped immensely in synthesizing and coding the papers, and to Priyanka Nanayakkara, who generously reviewed a draft of this paper and provided insightful feedback.

REFERENCES

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. Association for Computing Machinery. https://doi.org/10.1145/3173574.3174156

[2] Rediet Abebe, Solon Barocas, Jon Kleinberg, Karen Levy, Manish Raghavan, and David G Robinson. 2020. Roles for Computing in Social Change. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM. https://doi.org/10.1145/3351095.3372871

[3] Penelope Muse Abernathy. 2018. *The Expanding News Desert*. University of North Carolina Press Chapel Hill. https://www.cislm.org/wp-content/uploads/2018/10/The-Expanding-News-Desert-10{_}14-Web.pdf

[4] Michelle Alexander. 2010. *The new Jim Crow: Mass incarceration in the age of colorblindness*. The New Press. https://doi.org/10.4324/9781912282586

[5] Muhammad Ali, Piotr Sapiezynski, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Ad Delivery Algorithms: The Hidden Arbiters of Political Messaging. (2019). arXiv:1912.04255 http://arxiv.org/abs/1912.04255

[6] Athanasios Andreou, Giridhari Venkatadri, Oana Goga, Krishna P Gummadi, Patrick Loiseau, and Alan Mislove. 2018. Investigating Ad Transparency Mechanisms in Social Media: A Case Study of Facebook's Explanations. In *NDSS 2018*. https://doi.org/10.14722/ndss.2018.23191

[7] Michael Armbrust, Ion Stoica, Matei Zaharia, Armando Fox, Rean Griffith, Anthony D. Joseph, Randy Katz, Andy Konwinski, Gunho Lee, David Patterson, and Ariel Rabkin. 2018. It's Time to Do Something: Mitigating the Negative Impacts of Computing Through a Change to the Peer Review Process. http://www.brenthecht.com/papers/FCADIscussions{_}NegativeImpactsPost{_}032918.pdfhttps://acm-fca.org/2018/03/29/negativeimpacts/

[8] Joshua Asplund, Motahare Eslami, Hari Sundaram, Christian Sandvig, and Karrie Karahalios. 2020. Auditing Race and Gender Discrimination in Online Housing Markets. In *ICWSM*, Vol. 2020. 24–35. www.aaai.org

[9] Mahsa Badami, Olfa Nasraoui, and Patrick Shafto. 2018. PrCP: Pre-recommendation counter-polarization. In *IC3K 2018 - Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Vol. 1. 282–289. https://doi.org/10.5220/0006938702820289

[10] Eytan Bakshy, Solomon Messing, and Lada A Adamic. 2015. Exposure to ideologically diverse news and opinion on Facebook. *Science* 348, 6239 (jun 2015), 1130–1132. https://doi.org/10.1126/science.aaa1160

[11] Jack Bandy and Nicholas Diakopoulos. 2020. Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News. *International conference on web and social media (ICWSM)* (2020).

[12] Pınar Barlas, Kyriakos Kyriakou, Styliani Kleanthous, and Jahna Otterbacher. 2019. Social B(eye)as: Human and machine descriptions of people images. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. 583–591. https://www.aaai.org/ojs/index.php/ICWSM/article/view/3255/3123

[13] Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. 2017. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*.

[14] Alistair Barr. 2015. Google Mistakenly Tags Black People as 'Gorillas,' Showing Limits of Algorithms. https://www.wsj.com/articles/BL-DGB-42522https://blogs.wsj.com/digits/2015/07/01/google-mistakenly-tags-black-people-as-gorillas-showing-limits-of-algorithms/

[15] Muhammad Ahmad Bashir, Umar Farooq, Maryam Shahid, Muhammad Fareed Zaffar, and Christo Wilson. 2019. Quantity vs. Quality: Evaluating User Interest Profiles Using Ad Preference Managers. In *Proceedings 2019 Network and Distributed System Security Symposium*. https://doi.org/10.14722/ndss.2019.23392

[16] Anja Bechmann and Kristoffer L. Nielbo. 2018. Are We Exposed to the Same "News" in the News Feed?: An empirical analysis of filter bubbles as information similarity for Danish Facebook users. *Digital Journalism* 6, 8 (2018), 990–1002. https://doi.org/10.1080/21670811.2018.1510741

[17] Ruha Benjamin. 2019. *Race After Technology*. Polity Press.

[18] Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. 2017. Like trainer, like bot? Inheritance of bias in algorithmic content moderation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 10540 LNCS. Springer Verlag, 405–415. https://doi.org/10.1007/978-3-319-67256-4_32

[19] Engin Bozdag and Jeroen van den Hoven. 2015. Breaking the filter bubble: democracy and design. *Ethics and Information Technology* 17, 4 (2015), 249–265. https://doi.org/10.1007/s10676-015-9380-y

[20] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (jan 2006), 77–101. https://doi.org/10.1191/1478088706qp063oa

[21] Joy Buolamwini. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of Machine Learning Research*, Vol. 81. 1–15. http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf

[22] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. 2017. FDVT: Data Valuation Tool for Facebook Users. In *CHI*. https://doi.org/10.1145/3025453.3025903

[23] José González Cabañas, Ángel Cuevas, and Rubén Cuevas. 2018. Unveiling and Quantifying Facebook Exploitation of Sensitive Personal Data for Advertising Purposes. In *{USENIX} Security Symposium*. 479–495. https://www.usenix.org/conference/usenixsecurity18/presentation/cabanas

[24] Lorena Cano-Orón. 2019. Dr. Google, what can you tell me about homeopathy? Comparative study of the top10 websites in the United States, United Kingdom, France, Mexico and Spain. *Profesional de la Informacion* 28, 2 (2019). https://doi.org/10.3145/epi.2019.mar.13

[25] Abhijnan Chakraborty and Niloy Ganguly. 2018. Analyzing the news coverage of personalized newspapers. In *Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, ASONAM 2018*. Institute of Electrical and Electronics Engineers Inc., 540–543. https://doi.org/10.1109/ASONAM.2018.8508812

[26] Le Chen, Ruijun Ma, Anikó Hannák, and Christo Wilson. 2018. Investigating the impact of gender on rank in resume search engines. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. https://doi.org/10.1145/3173574.3174225

[27] Le Chen, Alan Mislove, and Christo Wilson. 2016. An Empirical Analysis of Algorithmic Pricing on Amazon Marketplace. In *Proceedings of the 25th International Conference on World Wide Web - WWW '16*. 1339–1349. https://doi.org/10.1145/2872427.2883089

[28] Ed H. Chi. 2011. On the Importance of Replication in HCI and Social Computing Research. https://cacm.acm.org/blogs/blog-cacm/109916-on-the-importance-of-replication-in-hci-and-social-computing-research/fulltext

[29] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5 (2017). https://doi.org/10.1089/big.2016.0047

[30] Sasha Costanza-Chock. 2018. Design Justice: towards an intersectional feminist framework for design theory and practice. In *Proceedings of the Design Research Society*, Vol. 2. https://doi.org/10.21606/drs.2018.679

[31] Nick Couldry and Ulises A Mejias. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television and New Media* 20, 4 (2019), 336–349. https://doi.org/10.1177/1527476418796632

[32] Cédric Courtois, Laura Slechten, and Lennert Coenen. 2018. Challenging Google Search filter bubbles in social and political information: Disconforming evidence from a digital methods case study. *Telematics and Informatics* 35, 7 (oct 2018), 2006–2015. https://doi.org/10.1016/j.tele.2018.07.004

[33] Kimberlé W. Crenshaw. 1989. Demarginalising the intersection of race and sex: A black feminist critique of anti-discrimination doctrine, feminist theory, and anti-racist politics. *University of Chicago Legal Forum* (1989). https://doi.org/10.4324/9781315582924-10

[34] Joshua P Darr, Matthew P Hitt, and Johanna L Dunaway. 2018. Newspaper Closures Polarize Voting Behavior. *Journal of Communication* 68, 6 (2018), 1007–1028. https://doi.org/10.1093/joc/jqy051

[35] Amit Datta, Anupam Datta, Jael Makagon, Deirdre K Mulligan, and Michael Carl Tschantz. 2018. Discrimination in Online Advertising: A Multidisciplinary Inquiry. *Proceedings of the 1st Conference on Fairness, Accountability and Transparency* 81 (2018), 20–34.

[36] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings. *Proceedings on Privacy Enhancing Technologies* 2015, 1 (apr 2015), 92–112. https://doi.org/10.1515/popets-2015-0007

[37] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. 2019. Bias in BIOS: A case study of semantic representation bias in a high-stakes setting. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 120–128. https://doi.org/10.1145/3287560.3287572

[38] Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and sharing health information online: Comparing search engines and social media. In *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, New York, New York, USA, 1365–1375. https://doi.org/10.1145/2556288.2557214

[39] Nicola Dell and Neha Kumar. 2016. The ins and outs of HCI for development. In *Conference on Human Factors in Computing Systems - Proceedings*. 2220–2232. https://doi.org/10.1145/2858036.2858081

[40] Michael A. DeVito. 2017. From Editors to Algorithms: A values-based approach to understanding story selection in the Facebook news feed. *Digital Journalism* 5, 6 (jul 2017), 753–773. https://doi.org/10.1080/21670811.2016.1178592

[41] Terrance DeVries, Ishan Misra, Changhan Wang, and Laurens van der Maaten. 2019. Does Object Recognition Work for Everyone?. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*. 52—-59. arXiv:1906.02659

[42] Nicholas Diakopoulos, Daniel Trielli, Jennifer Stark, and Sean Mussenden. 2018. I Vote For-How Search Informs Our Choice of Candidate. *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple* (2018). http://www.nickdiakopoulos.com/wp-content/uploads/2018/05/i-vote-for-open-access.pdf

[43] Mark Díaz and Nicholas Diakopoulos. 2019. Whose walkability?: Challenges in algorithmically measuring subjective experience. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359228

[44] Mark Díaz, Isaac Johnson, Amanda Lazar, Anne Marie Piper, and Darren Gergle. 2018. Addressing age-related bias in sentiment analysis. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2018-April. https://doi.org/10.1145/3173574.3173986

[45] Catherine D'Ignazio and Lauren F Klein. 2020. *Data Feminism.* MIT Press.

[46] Tawanna R Dillahunt, Xinyi Wang, Earnest Wheeler, Hao Fei Cheng, Brent Hecht, and Haiyi Zhu. 2017. The sharing economy in computing: A systematic literature review. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 38. https://doi.org/10.1145/3134673

[47] Emily Dreyfuss. 2019. What the Google-Genius Copyright Dispute Is Really About. https://www.wired.com/story/what-the-google-genius-copyright-dispute-is-really-about/

[48] Grant Duwe. 2019. Better Practices in the Development and Validation of Recidivism Risk Assessments: The Minnesota Sex Offender Screening Tool–4. *Criminal Justice Policy Review* 30, 4 (may 2019), 538–564. https://doi.org/10.1177/0887403417718608

[49] Grant Duwe and Ki Deuk Kim. 2017. Out With the Old and in With the New? An Empirical Comparison of Supervised Learning Algorithms to Predict Recidivism. *Criminal Justice Policy Review* 28, 6 (jul 2017), 570–600. https://doi.org/10.1177/0887403415604899

[50] Editors. 2020. Price Gouging Laws by State. https://consumer.findlaw.com/consumer-transactions/price-gouging-laws-by-state.html

[51] Malin Eiband, Sarah Theres Völkel, Daniel Buschek, Sophia Cook, and Heinrich Hussmann. 2019. When people and algorithms meet: User-reported problems in intelligent everyday applications. In *International Conference on Intelligent User Interfaces, Proceedings IUI*, Vol. Part F1476. 96–106. https://doi.org/10.1145/3301275.3302262

[52] B ELSEVIER. 2016. Scopus | The largest database of peer-reviewed literature.

[53] Robert Epstein, Ronald E Robertson, David Lazer, and Christo Wilson. 2017. Suppressing the Search Engine Manipulation Effect (SEME). *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–22. https://doi.org/10.1145/3134677

[54] Maria ; Eriksson and Anna Johansson. 2017. Tracking Gendered Streams. *Culture unbound : Journal of current cultural research* 9 (2017), 163–183. https://doi.org/10.25595/1449

[55] Motahhare Eslami, Kristen Vaccaro, Karrie Karahalios, and Kevin Hamilton. 2017. "Be careful; Things can be worse than they appear" - Understanding biased algorithms and users' behavior around them in rating platforms. In *Proceedings of the 11th International Conference on Web and Social Media, ICWSM 2017*. 62–71. www.aaai.org

[56] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press. https://doi.org/10.1080/10999922.2018.1511671

[57] Marc Faddoul, Guillaume Chaslot, and Hany Farid. 2020. *A longitudinal analysis of YouTube's promotion of conspiracy videos.* Technical Report. https://github.com/youtube-dataset/conspiracy

[58] Matthew E. Falagas, Eleni I. Pitsouni, George A. Malietzis, and Georgios Pappas. 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal* 22, 2 (feb 2008), 338–342. https://doi.org/10.1096/fj.07-9492LSF

[59] Sean Fischer, Kokil Jaidka, and Yphtach Lelkes. 2020. Auditing local news presence on Google News. *Nature Human Behaviour* (sep 2020). https://doi.org/10.1038/s41562-020-00954-0

[60] Richard Fletcher and Rasmus Kleis Nielsen. 2018. Automated Serendipity: The effect of using search engines on news repertoire balance and diversity. *Digital Journalism* 6, 8 (2018), 976–989. https://doi.org/10.1080/21670811.2018.1502045

[61] Anthony Flores, Christopher T. Lowenkamp, and Kristin Bechtel. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias". *HeinOnline* 80, September (2016), 38–46. https://heinonline.org/HOL/Page?handle=hein.journals/fedpro80{&}div=21{&}g{_}sent=1{&}casa{_}token=l0DKlZlxxegAAAAA:HK9QpZrQVh123JWE0JXWoJyFjiyKviZDZS6y7FmtXQE-ohuaqtfaR0P0dJX3T6LzOh91l3Yu{&}collection=journals

[62] Sarah Fox, Jill Dimond, Lilly Irani, Tad Hirsch, Michael Muller, and Shaowen Bardzell. 2017. Social justice and design: Power and oppression in collaborative systems. In *CSCW 2017 - Companion of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. Association for Computing Machinery, Inc, New York, New York, USA, 117–122. https://doi.org/10.1145/3022198.3022201

[63] Jon Froehlich, Leah Findlater, and James Landay. 2010. The Design of Eco-Feedback Technology. In *Proceedings of the SIGCHI conference on human factors in computing systems*.

[64] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2221–2231. https://doi.org/10.1145/3292500.3330691

[65] Tarleton Gillespie. 2014. The Relevance of Algorithms. *Media technologies: Essays on communication, materiality, and society* 167 (2014), 167–194. https://doi.org/10.7551/mitpress/9780262525374.003.0009

[66] Tarleton Gillespie. 2017. Algorithmically recognizable: Santorum's Google problem, and Google's Santorum problem. *Information Communication and Society* 20, 1 (2017), 63–80. https://doi.org/10.1080/1369118X.2016.1199721

[67] Philip Gillingham. 2016. Predictive Risk Modelling to Prevent Child Maltreatment and Other Adverse Outcomes for Service Users: Inside the 'Black Box' of Machine Learning. *British Journal of Social Work* 46, 4 (2016), 1044–1058. https://doi.org/10.1093/bjsw/bcv031

[68] Carlos A. Gomez-Uribe and Neil Hunt. 2015. The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems* 6, 4 (dec 2015), 1–19. https://doi.org/10.1145/2843948

[69] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of Sheffield SIGIR - Twenty-Seventh Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, New York, USA, 478–479. https://doi.org/10.1145/1008992.1009079

[70] Mario Haim, Andreas Graefe, and Hans Bernd Brosius. 2018. Burst of the Filter Bubble?: Effects of personalization on the diversity of Google News. *Digital Journalism* 6, 3 (2018), 330–343. https://doi.org/10.1080/21670811.2017.1338145

[71] Alex Hanna, Emily Denton, and Andrew Smart. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, New York, NY, USA. https://doi.org/10.1145/3351095.3372826

[72] Anikó Hannák, Alan Mislove, Claudia Wagner, Markus Strohmaier, David Garcia, and Christo Wilson. 2017. Bias in Online freelance marketplaces: Evidence from TaskRabbit and Fiverr. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 1914–1933. https://doi.org/10.1145/2998181.2998327

[73] Anikó Hannák, Piotr Sapieżyński, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2013. Measuring Personalization of Web Search. In *Proceedings of the 22nd international conference on World Wide Web*. ACM, 527—-538. https://doi.org/10.1145/2488388.2488435

[74] Aniko Hannak, Gary Soeller, David Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. In *Proceedings of the 2014 Conference on Internet Measurement Conference - IMC '14*. 305–318. https://doi.org/10.1145/2663716.2663744

[75] Sandra G Harding. 2004. *The feminist standpoint theory reader: Intellectual and political controversies*. Psychology Press. https://doi.org/10.4135/9781483346229.n142

[76] Alexa M Harris, Diego Gómez-Zará, Leslie A Dechurch, and Noshir S. Contractor. 2019. Joining together online: The trajectory of CSCW scholarship on group formation. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 27. https://doi.org/10.1145/3359250

[77] Megan L Head, Luke Holman, Rob Lanfear, Andrew T Kahn, and Michael D Jennions. 2015. The Extent and Consequences of P-Hacking in Science. *PLoS Biology* 13, 3 (2015). https://doi.org/10.1371/journal.pbio.1002106

[78] Felicitas Hesselmann, Verena Graf, Marion Schmidt, and Martin Reinhart. 2017. The visibility of scientific misconduct: A review of the literature on retracted journal articles. , 814–845 pages. https://doi.org/10.1177/0011392116663807

[79] Anna Lauren Hoffmann. 2019. Where fairness fails: data, algorithms, and the limits of antidiscrimination discourse. *Information Communication and Society* 22, 7 (jun 2019), 900–915. https://doi.org/10.1080/1369118X.2019.1573912

[80] Desheng Hu, Ronald E Robertson, Shan Jiang, and Christo Wilson. 2019. Auditing the partisanship of Google search snippets. In *The Web Conference 2019 - Proceedings of the World Wide Web Conference, WWW 2019*. 693–704. https://doi.org/10.1145/3308558.3313654

[81] Thomas Hupperich, Nicolai Wilkop, Dennis Tatang, and Thorsten Holz. 2018. An empirical study on online price differentiation. *CODASPY 2018 - Proceedings of the 8th ACM Conference on Data and Application Security and Privacy* 2018-Janua (2018), 76–83. https://doi.org/10.1145/3176258.3176338

[82] Eslam Hussein, Prerna Juneja, and Tanushree Mitra. 2020. Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW1 (may 2020), 1–27. https://doi.org/10.1145/3392854

[83] Andrew Hutchinson. 2020. Facebook Removes Over 1,000 Ad Targeting Options Due to Low Usage. https://www.socialmediatoday.com/news/facebook-removes-over-1000-ad-targeting-options-due-to-low-usage/583406/

[84] Lucas D. Introna and Helen Nissenbaum. 2000. Shaping the web: Why the politics of search engines matters. *Information Society* 16, 3 (2000), 169–185. https://doi.org/10.1080/01972240050133634

[85] Shan Jiang, John Martin, and Christo Wilson. 2019. Who's the Guinea pig? Investigating online A/B/N tests in-the-wild. In *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, New York, New York, USA, 201–210. https://doi.org/10.1145/3287560.3287565

[86] Shan Jiang, Ronald E. Robertson, and Christo Wilson. 2019. Bias misperceived: The role of partisanship and misinformation in YouTube comment moderation. In *Proceedings of the 13th International Conference on Web and Social Media,*

ICWSM 2019, Vol. 13(01). 278–289. https://wvvw.aaai.org/ojs/index.php/ICWSM/article/view/3229/3097

[87] Alexandros Karakasidis and Evaggelia Pitoura. 2019. Identifying bias in name matching tasks. In *Advances in Database Technology - EDBT*, Vol. 2019-March. 626–629. https://doi.org/10.5441/002/edbt.2019.72

[88] Matthew Kay, Cynthia Matuszek, and Sean A Munson. 2015. Unequal Representation and Gender Stereotypes in Image Search Results for Occupations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. 3819–3828. https://doi.org/10.1145/2702123.2702520

[89] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2017. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. , 2564–2572 pages. http://proceedings.mlr.press/v80/kearns18a.html

[90] Os Keyes, Meredith Durbin, and Jevan Hutson. 2019. A mulching proposal: Analysing and improving an algorithmic system for turning the elderly into high-nutrient slurry. In *Conference on Human Factors in Computing Systems - Proceedings*. https://doi.org/10.1145/3290607.3310433

[91] Lauren Kirchner, Surya Mattu, Jeff Larson, and Julia Angwin. 2016. Machine Bias — ProPublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[92] Rob Kitchin. 2017. Thinking critically about and researching algorithms. *Information Communication and Society* 20, 1 (2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[93] Chloe Kliman-Silver, Aniko Hannak, David Lazer, Christo Wilson, and Alan Mislove. 2015. Location, Location, Location: The Impact of Geolocation on Web Search Personalization. In *Proceedings of the 2015 Internet Measurement Conference*. ACM, 121—-127. https://doi.org/10.1145/2815675.2815714

[94] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P Gummadi, and Karrie Karahalios. 2017. Quantifying search bias: Investigating sources of bias for political searches in social media. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 417–432. https://doi.org/10.1145/2998181.2998321

[95] Juhi Kulshrestha, Motahhare Eslami, Johnnatan Messias, Muhammad Bilal Zafar, Saptarshi Ghosh, Krishna P. Gummadi, and Karrie Karahalios. 2019. Search bias quantification: investigating political bias in social media and web search. *Information Retrieval Journal* 22, 1-2 (apr 2019), 188–227. https://doi.org/10.1007/s10791-018-9341-2

[96] Kyriakos Kyriakou, Pınar Barlas, Styliani Kleanthous, and Jahna Otterbacher. 2019. Fairness in proprietary image tagging algorithms: A cross-platform audit on people images. In *Proceedings of the 13th International Conference on Web and Social Media, ICWSM 2019*. 313–322. https://www.aaai.org/ojs/index.php/ICWSM/article/view/3232/3100

[97] Cameron Lai and Markus Luczak-Roesch. 2019. You Can't See What You Can't See: Experimental Evidence for How Much Relevant Information May Be Missed Due to Google's Web Search Personalisation. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 11864 LNCS. 253–266. https://doi.org/10.1007/978-3-030-34971-4_17 arXiv:1904.13022

[98] Anja Lambrecht and Catherine Tucker. 2019. Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of stem career ads. *Management Science* 65, 7 (2019), 2966–2981. https://doi.org/10.1287/mnsc.2018.3093

[99] Nadia Latif. 2017. It's lit! How film finally learned to light black skin. https://www.theguardian.com/film/2017/sep/21/its-lit-how-film-finally-learned-how-to-light-black-skin

[100] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2015-April. 1603–1612. https://doi.org/10.1145/2702123.2702548

[101] Hanlin Li and Brent Hecht. 2020. *3 Stars on Yelp, 4 Stars on Google Maps: A Cross-Platform Examination of Restaurant Ratings*. Technical Report. https://doi.org/124564

[102] Gustavo López and Luis A Guerrero. 2017. Awareness supporting technologies used in collaborative systems - A systematic literature review. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. 808–820. https://doi.org/10.1145/2998181.2998281

[103] Emma Lurie and Eni Mustafaraj. 2019. Opening Up the Black Box: Auditing Google's Top Stories Algorithm. In *Proceedings of the International Florida Artificial Intelligence Research Society Conference*. https://en.wikipedia.org/wiki/Christinehttps://aaai.org/ocs/index.php/FLAIRS/FLAIRS19/paper/view/18316/17433

[104] Roger Mähler and Patrick Vonderau. 2017. Studying ad targeting with digital methods: The case of spotify. *Culture Unbound* 9, 2 (2017), 212–221. https://doi.org/10.3384/cu.2000.1525.1792212

[105] Louise Matsakis and Paris Martineau. [n.d.]. Coronavirus Disrupts Social Media's First Line of Defense. https://www.wired.com/story/coronavirus-social-media-automated-content-moderation/

[106] Jeanna Matthews, Marzieh Babaeianjelodar, Stephen Lorenz, Abigail Matthews, Dan Krane, Mariama Njie, Jessica Goldthwaite, Nathaniel Adams, and Clinton Hughes. 2019. The right to confront your accusers: Opening the black box of forensic DNA software. In *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 321–327. https://doi.org/10.1145/3306618.3314279

[107] Anna May, Johannes Wachs, and Anikó Hannák. 2019. Gender differences in participation and reward on Stack Overflow. *Empirical Software Engineering* 24, 4 (aug 2019), 1997–2019. https://doi.org/10.1007/s10664-019-09685-x

[108] Connor Mcmahon, Isaac Johnson, and Brent Hecht. 2017. The Substantial Interdependence of Wikipedia and Google: A Case Study on the Relationship Between Peer Production Communities and Information Technologies. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*. www.aaai.orghttp://www-users.cs.umn.edu/{~}joh12041/Publications/WikiGoogle{_}ICWSM17.pdf

[109] Solomon Messing, Christina DeGregorio, Bennett Hillenbrand, Gary King, Saurav Mahanti, Zagreb Mukerjee, Chaya Nayak, Nate Persily, Bogdan State, and Arjun Wilkins. 2020. *Facebook Privacy-Protected Full URLs Data Set*. Technical Report.

[110] Danaë Metaxa, Joon Sung Park, James A Landay, and Jeff Hancock. 2019. Search media and elections: A longitudinal investigation of political search results in the 2018 U.S. Elections. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019). https://doi.org/10.1145/3359231

[111] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2012. Detecting price and search discrimination on the internet. In *Proceedings of the 11th ACM Workshop on Hot Topics in Networks - HotNets-XI*. 79–84. https://doi.org/10.1145/2390231.2390245

[112] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2013. Crowd-assisted Search for Price Discrimination in E-Commerce: First results. (2013). https://doi.org/10.1145/2535372.2535415 arXiv:1307.4531

[113] Jakub Mikians, László Gyarmati, Vijay Erramilli, and Nikolaos Laoutaris. 2013. Crowd-assisted Search for Price Discrimination in E-commerce: First results. *CoNEXT 2013 - Proceedings of the 2013 ACM International Conference on Emerging Networking Experiments and Technologies* (2013), 1–6. https://doi.org/10.1145/2535372.2535415

[114] Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. 2016. The ethics of algorithms: Mapping the debate. *Big Data & Society* 3, 2 (2016), 205395171667967. https://doi.org/10.1177/2053951716679679

[115] Hallvard Moe. 2019. Comparing Platform "Ranking Cultures" Across Languages: The Case of Islam on YouTube in Scandinavia. *Social Media and Society* 5, 1 (2019). https://doi.org/10.1177/2056305118817038

[116] Shakir Mohamed, Marie Therese Png, and William Isaac. 2020. Decolonial AI: Decolonial Theory as Sociotechnical Foresight in Artificial Intelligence. *Philosophy and Technology* (jul 2020), 1–26. https://doi.org/10.1007/s13347-020-00405-8

[117] David Moher, Alessandro Liberati, Jennifer Tetzlaff, Douglas G. Altman, Doug Altman, Gerd Antes, David Atkins, Virginia Barbour, Nick Barrowman, Jesse A. Berlin, Jocalyn Clark, Mike Clarke, Deborah Cook, Roberto D'Amico, Jonathan J. Deeks, P. J. Devereaux, Kay Dickersin, Matthias Egger, Edzard Ernst, Peter C. Gøtzsche, Jeremy Grimshaw, Gordon Guyatt, Julian Higgins, John P.A. Ioannidis, Jos Kleijnen, Tom Lang, Nicola Magrini, David McNamee, Lorenzo Moja, Cynthia Mulrow, Maryann Napoli, Andy Oxman, Ba' Pham, Drummond Rennie, Margaret Sampson, Kenneth F. Schulz, Paul G. Shekelle, David Tovey, and Peter Tugwell. 2009. Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement. , 264–269 pages. https://doi.org/10.7326/0003-4819-151-4-200908180-00135

[118] David Moher, Lesley Stewart, and Paul Shekelle. 2015. All in the Family: Systematic reviews, rapid reviews, scoping reviews, realist reviews, and more. , 183 pages. https://doi.org/10.1186/s13643-015-0163-7

[119] Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information Communication and Society* 21, 7 (2018), 959–977. https://doi.org/10.1080/1369118X.2018.1444076

[120] Eni Mustafaraj, Emma Lurie, and Claire Devine. 2020. The case for voter-centered audits of search engines during political elections. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 559–569. https://doi.org/10.1145/3351095.3372835

[121] Efrat Nechushtai and Seth C. Lewis. 2019. What kind of news gatekeepers do we want machines to be? Filter bubbles, fragmentation, and the normative dimensions of algorithmic recommendations. *Computers in Human Behavior* 90 (jan 2019), 298–307. https://doi.org/10.1016/j.chb.2018.07.043

[122] Helen Nissenbaum. 2001. How Computer Systems Embody Values. *Computer* 34, 3 (2001), 119–120. https://doi.org/10.1109/2.910905

[123] Safiya Noble. 2013. Google search: Hyper-visibility as a means of rendering black women and girls invisible. *InVisible Culture* 19 (2013). https://urresearch.rochester.edu/institutionalPublicationPublicView.action?institutionalItemId=27584

[124] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

[125] Ihudiya Finda Ogbonnaya-Ogburu, Angela D.R. Smith, Alexandra To, and Kentaro Toyama. 2020. Critical Race Theory for HCI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM, New York, NY, USA, 1–16. https://doi.org/10.1145/3313831.3376392

[126] Cathy O'Neil. 2017. *Weapons of math destruction : how big data increases inequality and threatens democracy*. Broadway Books.

[127] Seeta Peña Gangadharan and Jędrzej Niklas. 2019. Decentering technology in discourse on discrimination *. *Information Communication and Society* 22, 7 (jun 2019), 882–899. https://doi.org/10.1080/1369118X.2019.1593484

[128] Pew Research Center. 2019. *Facebook Algorithms and Personal Data.* Technical Report. https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

[129] Cornelius Puschmann. 2018. Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism* (nov 2018), 1–20. https://doi.org/10.1080/21670811.2018.1539626

[130] Alexander J Quinn and Benjamin B Bederson. 2011. Human computation: A survey and taxonomy of a growing field. In *Conference on Human Factors in Computing Systems - Proceedings*. 1403–1412. https://doi.org/10.1145/1978942.1979148

[131] Iyad Rahwan, Manuel Cebrian, Nick Obradovich, Josh Bongard, Jean-François Bonnefon, Cynthia Breazeal, Jacob W. Crandall, Nicholas A. Christakis, Iain D. Couzin, Matthew O. Jackson, Nicholas R. Jennings, Ece Kamar, Isabel M. Kloumann, Hugo Larochelle, David Lazer, Richard McElreath, Alan Mislove, David C Parkes, Alex 'Sandy' Pentland, Margaret E. Roberts, Azim Shariff, Joshua B Tenenbaum, and Michael Wellman. 2019. Machine behaviour. *Nature* 568, 7753 (2019), 477–486. https://doi.org/10.1038/s41586-019-1138-y

[132] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial AI products. In *AIES 2019 - Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435. https://doi.org/10.1145/3306618.3314244

[133] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, New York, NY, USA, 33–44. https://doi.org/10.1145/3351095.3372873 arXiv:2001.00973

[134] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A.F. Almeida, and W. M. Wagner Meira. 2020. Auditing radicalization pathways on YouTube. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, New York, NY, USA, 131–141. https://doi.org/10.1145/3351095.3372879

[135] Bernhard Rieder, Ariadna Matamoros-Fernández, and Òscar Coromina. 2018. From ranking algorithms to 'ranking cultures': Investigating the modulation of visibility in YouTube search results. *Convergence* 24, 1 (2018), 50–68. https://doi.org/10.1177/1354856517736982

[136] Ronald E Robertson, Shan Jiang, Kenneth Joseph, Lisa Friedland, David Lazer, and Christo Wilson. 2018. Auditing partisan audience bias within Google search. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 22. https://doi.org/10.1145/3274417

[137] Ronald E Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing autocomplete: Suggestion networks and recursive algorithm interrogation. In *WebSci 2019 - Proceedings of the 11th ACM Conference on Web Science*, Vol. 10. ACM, 235–244. https://doi.org/10.1145/3292522.3326047

[138] Ronald E Robertson, Shan Jiang, David Lazer, and Christo Wilson. 2019. Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation. 10 (2019). https://doi.org/10.1145/3292522.3326047

[139] Jeffrey S. Saltz and Neil Dewar. 2019. Data science ethical considerations: a systematic literature review and proposed project framework. *Ethics and Information Technology* 21, 3 (sep 2019), 197–208. https://doi.org/10.1007/s10676-019-09502-5

[140] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cedric Langbort. 2014. Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms. In *Data and discrimination: converting critical concerns into productive inquiry*. 1—-23. https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf

[141] Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2020. The risk of racial bias in hate speech detection. In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference*. 1668–1678.

[142] Márcio Silva, Lucas Santos De Oliveira, Athanasios Andreou, Pedro Olmo Vaz De Melo, Oana Goga, and Fabricio Benevenuto. 2020. Facebook Ads Monitor: An Independent Auditing System for Political Ads on Facebook. In *The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020*. Association for Computing Machinery, Inc, New York, NY, USA, 224–234. https://doi.org/10.1145/3366423.3380109

[143] Pelle Snickars. 2017. More of the same - On spotify radio. *Culture Unbound* 9, 2 (2017), 184–211. https://doi.org/10.3384/cu.2000.1525.1792184

[144] Gary Soeller, Karrie Karahalios, Christian Sandvig, and Christo Wilson. 2016. MapWatch: Detecting and monitoring international border personalization on online maps. In *25th International World Wide Web Conference, WWW 2016*. International World Wide Web Conferences Steering Committee, New York, New York, USA, 867–878. https://doi.org/10.1145/2872427.2883016

[145] Leo Sun. 2020. TikTok Is Still Growing Faster than Facebook and Snapchat. https://www.nasdaq.com/articles/tiktok-is-still-growing-faster-than-facebook-and-snapchat-2020-09-14

[146] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. *Queue* 11, 3 (2013). https://doi.org/10.2139/ssrn.2208240
[147] Babak Taati, Shun Zhao, Ahmed B. Ashraf, Azin Asgarian, M. Erin Browne, Kenneth M. Prkachin, Alex Mihailidis, and Thomas Hadjistavropoulos. 2019. Algorithmic bias in clinical populations - Evaluating and improving facial analysis technology in older adults with dementia. *IEEE Access* 7 (2019), 25527–25534. https://doi.org/10.1109/ACCESS.2019.2900022
[148] Songül Tolan, Marius Miron, Emilia Gómez, and Carlos Castillo. 2019. Why machine learning may lead to unfairness: Evidence from risk assessment for juvenile justice in Catalonia. In *Proceedings of the 17th International Conference on Artificial Intelligence and Law, ICAIL 2019*. 83–92. https://doi.org/10.1145/3322640.3326705
[149] Daniel Trielli and Nicholas Diakopoulos. 2019. Search as News Curator: The Role of Google in Shaping Attention to News Information. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.
[150] Michael Carl Tschantz, Serge Egelman, Jaeyoung Choi, Nicholas Weaver, and Gerald Friedland. 2018. The Accuracy of the Demographic Inferences Shown on Google's Ad Settings. In *WPES*. 33–41. https://doi.org/10.1145/3267323.3268962
[151] Andrew Tutt. 2017. An FDA for algorithms. , 83–123 pages. https://doi.org/10.2139/ssrn.2747994
[152] Blase Ur, Pedro Giovanni Leon, Lorrie Faith Cranor, Richard Shay, and Yang Wang. 2012. Smart, useful, scary, creepy: Perceptions of Online Behavioral Advertising. In *Soups 2012, ACM Press*. 1. https://doi.org/10.1145/2335356.2335362
[153] Ester van Laar, Alexander J.A.M. van Deursen, Jan A.G.M. van Dijk, and Jos de Haan. 2017. The relation between 21st-century skills and digital skills: A systematic literature review. *Computers in Human Behavior* 72 (2017), 577–588. https://doi.org/10.1016/j.chb.2017.03.010
[154] Giridhari Venkatadri, Piotr Sapiezynski, Elissa M Redmiles, Alan Mislove, Oana Goga, Michelle Mazurek, and Krishna P Gummadi. 2019. Auditing Offline Data Brokers via Facebook's Advertising Platform. In *WWW '19*. https://doi.org/10.1145/3308558.3313666
[155] Nicholas Vincent, Brent Hecht, and Shilad Sen. 2019. "Data Strikes": Evaluating the Effectiveness of a New Form of Collective Action Against Technology Companies. In *The World Wide Web Conference on - WWW '19*. 1931–1943. https://doi.org/10.1145/3308558.3313742
[156] Nicholas Vincent, Isaac Johnson, Patrick Sheehan, and Brent Hecht. 2019. Measuring the Importance of User-Generated Content to Search Engines. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*.
[157] Sara Wachter-Boettcher. 2017. *Technically Wrong: Sexist Apps, Biased Algorithms, and Other Threats of Toxic Tech Sexism*. WW Norton & Company.
[158] Matthew S Weber and Allie Kosterich. 2018. Coding the News: The role of computer code in filtering and distributing news. *Digital Journalism* 6, 3 (2018), 310–329. https://doi.org/10.1080/21670811.2017.1366865
[159] Sarah Myers West. 2019. Data Capitalism: Redefining the Logics of Surveillance and Privacy. *Business and Society* 58, 1 (2019), 20–41. https://doi.org/10.1177/0007650317718185
[160] Maranke Wieringa. 2020. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Association for Computing Machinery, Inc, 1–18. https://doi.org/10.1145/3351095.3372833
[161] Max Willens. 2019. Publishers see mobile traffic spikes from Google Discover. https://digiday.com/media/publishers-see-mobile-traffic-spikes-google-discover/
[162] Max L. Wilson, Paul Resnick, David Coyle, and Ed H. Chi. 2013. RepliCHI – The Workshop. In *Conference on Human Factors in Computing Systems - Proceedings*, Vol. 2013-April. Association for Computing Machinery, New York, New York, USA, 3159–3162. https://doi.org/10.1145/2468356.2479636
[163] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (2016), 38–44. https://doi.org/10.1145/2907069
[164] Hui Zhang, Munmun De Choudhury, and Jonathan Grudin. 2014. Creepy but inevitable? the evolution of social networking. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*. ACM Press, New York, New York, USA, 368–378. https://doi.org/10.1145/2531602.2531643
[165] Shoshana Zuboff. 2015. Big other: Surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology* 30, 1 (2015), 75–89. https://doi.org/10.1057/jit.2015.5