Foundations and Trends® in Human-Computer Interaction

Auditing Algorithms

Understanding Algorithmic Systems from the Outside In

Suggested Citation: Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock and Christian Sandvig (2021), "Auditing Algorithms", Foundations and Trends® in Human-Computer Interaction: Vol. 14, No. 4, pp 272–344. DOI: 10.1561/1100000083.

Danaë Metaxa
Stanford University

Joon Sung Park
Stanford University

Ronald E. Robertson
Northeastern University

Karrie Karahalios
University of Illinois at Urbana-Champaign

Christo Wilson
Northeastern University

Jeff Hancock
Stanford University

Christian Sandvig
University of Michigan

This article may be used only for the purpose of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval.

now
the essence of knowledge
Boston — Delft

Contents

Auditing Algorithms

Danaë Metaxa[1], Joon Sung Park[2], Ronald E. Robertson[3], Karrie Karahalios[4], Christo Wilson[5], Jeff Hancock[6] and Christian Sandvig[7]

[1] *Stanford University; metaxa@cs.stanford.edu*
[2] *Stanford University; joonspk@stanford.edu*
[3] *Northeastern University; robertson.ron@northeastern.edu*
[4] *University of Illinois at Urbana-Champaign; kkarahal@illinois.edu*
[5] *Northeastern University; cbw@ccs.neu.edu*
[6] *Stanford University; hancockj@stanford.edu*
[7] *University of Michigan; csandvig@umich.edu*

ABSTRACT

Algorithms are ubiquitous and critical sources of information online, increasingly acting as gatekeepers for users accessing or sharing information about virtually any topic, including their personal lives and those of friends and family, news and politics, entertainment, and even information about health and well-being. As a result, algorithmically-curated content is drawing increased attention and scrutiny from users, the media, and lawmakers alike. However, studying such content poses considerable challenges, as it is both dynamic and ephemeral: these algorithms are constantly changing, and frequently changing silently, with no record of the content to which users have been exposed over time. One strategy that has proven effective is the *algorithm audit*: a method of repeatedly querying an algorithm and observing its output in order to draw conclusions about the algorithm's opaque

Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock and Christian Sandvig (2021), "Auditing Algorithms", Foundations and Trends® in Human-Computer Interaction: Vol. 14, No. 4, pp 272–344. DOI: 10.1561/1100000083.
©2021 D. Metaxa et al.

inner workings and possible external impact. In this work, we present an overview of the algorithm audit methodology, including the history of audit studies in the social sciences from which this method is derived; a summary of key algorithm audits over the last two decades in a variety of domains, including health, politics, discrimination, and others; and a set of best practices for conducting algorithm audits today, contextualizing these practices using search engine audits as a case study. Finally, we conclude by discussing the social, ethical, and political dimensions of auditing algorithms, and propose normative standards for the use of this method.

1

An Introduction to Auditing

In 2012, Harvard professor Latanya Sweeney and her colleague found themselves Googling Dr. Sweeney's name, searching the web for a copy of a paper she had written. Instead, at the top of the search page they found an advertisement with the headline, "Latanya Sweeney. Arrested?" (Sweeney, 2013a). With no arrest record to speak of, Dr. Sweeney was shocked. After paying a fee to access the company's supposed information, she confirmed that the company's records did not contain any criminal information under her name. Investigating further, Dr. Sweeney and her colleague searched for his name, and found an advertisement from the same company—but this one simply offering information about people with that name, with no mention of an arrest record or anything of the sort. Searching for more and more names, Dr. Sweeney and her colleague were forced to conclude that it seemed like the advertisements Google was serving were racially biased, suggestive of arrest records more often for Black-sounding names like Dr. Sweeney's than white-sounding names like her colleague's. Well-equipped to study this phenomenon rigorously, Dr. Sweeney undertook a study collecting the ads served by Google for over 2,000 names of real people, using one set of names likely to belong to someone Black and another likely to

belong to someone white. She found that Google's advertisements were up to 25% more likely to suggest an arrest record for a Black name than a white one, a discrepancy that was statistically significant and large enough that, were an employer disparately treating employees by race to this degree, the employer could potentially be charged with violating U.S. labor discrimination laws.

The reason for this racist discrepancy in ads being shown by Google is hard to identify conclusively; at worst, companies buying advertisements from Google could be purposefully targeting minority-sounding names. But the same outcome could result if companies provided Google with several versions of ad copy for the algorithm to automatically choose to maximize clicks, and people searching for Black-sounding names were for some reason more likely to click ads mentioning arrest, while people searching for white-sounding names were more likely to click on neutrally-worded ads. In any case, the implications are obviously serious. Imagine your potential employers, university admissions officers, or even your new partner's parents searching for your name on Google and finding ads that suggest an arrest record. The negative impact of such ads could be severe and immediate, and in this case, as Dr. Sweeney showed, it disproportionately affected Black people.

This kind of discrimination, apparent only in aggregate, is especially challenging to study in the context of computer systems whose exact workings are opaque to an outside observer. Sweeney's strategy, systematically querying the Google Search algorithm with a wide range of inputs and statistically comparing the results, is one of the most effective ways to study bias in algorithms. It is known as the *algorithm audit*. In this monograph, we present an overview of this powerful method including what it is, how it is used, and why it matters. We discuss the history of the audit method, its use in algorithm contexts, and best practices for researchers conducting algorithm audits in their own work. Our team of researchers has extensive experience conducting algorithm audits, and in this work we seek to answer such questions by drawing from the history of auditing in the social sciences as well as exemplary work auditing sociotechnical systems in recent decades.

1.1 What is an Audit?

Algorithm audits, our focus in most of this monograph, are a specific sub-type of a broader method, the audit study. Before we delve into the specifics of what makes a good audit and how auditing is applied to different social and sociotechnical contexts, we must define this method. Developed originally as a type of experiment used by social scientists, auditing is a methodology used to deploy randomized controlled experiments in a field setting (i.e., outside the lab) (Gaddis, 2018). Auditors conducting such a study must probe a process (e.g., a company's hiring process; a professor's process of responding to student emails; an algorithm providing users search results) by providing it with one or more inputs, while changing some attributes of that input, such as e.g., the race of the applicant (Bertrand and Mullainathan, 2004); the gender of the student (Milkman *et al.*, 2012); or the search history or date of search (Robertson *et al.*, 2018b; Metaxa *et al.*, 2019). Many governments, including that of the United States, conduct audits routinely, as a part of civic infrastructure. In the U.S., for instance, the Government Accountability Office conducts audits at the specific request of Congress or as mandated by law, and investigates the allocation of federal funds, allegations of illegal activity, the success of policies enacted, and other aspects of government function *U.S. Government Accountability Office (U.S. GAO)* 2021.

Bertrand and Mullainathan (2004) is a classic example of a (non-algorithmic) audit, one that inspired Latanya Sweeney's later work online. In that study, the authors sought to test whether there was racial bias in hiring, specifically in the resume reviewing stage, across a wide range of companies and industries. To do so, they constructed and sent fictitious resumes with white-sounding or Black-sounding names in response to job postings, and measured the rate at which those fictional job applicants got callbacks for interviews. They found that overall, applicants with white-sounding names received 50% more callbacks than those with Black-sounding names, and that the amount of discrimination was uniform across the industries they studied, concluding that racial discrimination was still widely prevalent in the labor market.

Algorithm audits are a specific subset of audit studies focused on studying algorithmic systems and content (Sandvig *et al.*, 2014). Rather than studying racial bias in human resume reviewing, then, an algorithm audit might investigate potential bias in an automated, algorithmically-powered resume screening process. Challenges specific to studying algorithms also lead algorithm audits to use different strategies and techniques—while Dr. Sweeney was able to manually search for Black- and white-sounding names and examine the search results displayed, algorithm auditors often need to build a software apparatus to amass large quantities of data from their platform of interest.

1.2 Differentiating Algorithm Audits from Other Testing

As evidenced by the examples we have already discussed, audit studies often—but not always—have an end goal of determining whether a system is biased or discriminatory. What all algorithm audits do have in common is an aim to test whether some deficiency (discrimination, bias, or something else) exists in an algorithmic system or not, without direct access to the internals of that system. In pursuit of this goal, there are several key features of audits that differentiate them from other types of testing, including the focus of study, scope of the conclusions drawn, and the position of the investigator while auditing.

Unlike other forms of testing such as A/B tests, the audit's subject of study is the system itself, not any particular component or a user's response to it. In an A/B test, for instance, the subject of study is the user, with the investigator seeking to understand the user's change in behavior while interacting with a system. Auditors may also be interested in a system's effect on people, but the angle of an audit is different, focused on the system itself. For auditors, studying the user is neither necessary nor sufficient; while some audit studies may include a component of user testing, audits more often measure the raw output of a system and rely on theory to infer what these outputs mean for a system's users. In the rare case that an audit does experiment on users, they are usually paid and consenting participants, rather than unknowing users of a system. This is often the case because measuring user behavior would be impossible (as when auditing a system one to

which one does not have internal direct access), or unethical (we further discuss the ethics of auditing in Section 4).

Algorithm audits are also differentiated from other types of system testing by their scope. Most other forms of testing, including test suites, result in binary pass/fail conclusions at the level of individual test cases. An audit, on the other hand, has a broader scope and, it follows, must be systematic. It results in a declaration about the system as a whole; while auditors may conduct tests as part of their auditing, the overall finding of an audit is not merely to conclude that a given system is "right" or "wrong"—the results can only be discerned in aggregate. In this sense, an audit is a method of inspection or analysis more than of testing.

Finally, a third key difference is the role and position of the investigator conducting an audit study. A distinguishing feature of an audit study, unlike other forms of testing, is that an audit may be conducted with varying levels of participation or consent from the entity being audited—including partial or none at all. Audits are purposefully intended to be external evaluations, based only on outward-facing aspects, not insider knowledge on the process being studied. Most other testing is conducted internally, at the explicit direction of the proprietors of the system. This point raises interesting questions around the cost accrued when conducting an audit (for example, in system resources). Sending fake resumes to job postings costs companies employee time; auditing ads served by the Google search engine by repeatedly querying it uses Google's servers' resources. While most other forms of testing are conducted internally by a willing entity who bears the full cost, audits are conducted externally on an entity that is not necessarily willing or even informed of the ongoing audit, but the cost of the audit is shared between the investigators and the entity itself.

Before returning to algorithm-specific audits, in the the next section we will delve into the history of audit studies in the social sciences, establishing how the method was developed, what kinds of social systems it has been used to study, and what impacts these studies have had on the world.

1.3 Positionality Statement

As academic researchers in the United States with experience conducting search audits, we write primarily for fellow researchers interested in conducting them, with a secondary goal of speaking to an audience of academics, journalists, and others interested in interpreting and evaluating such research. Our team of authors has combined experience performing over 35 audits, covering areas including web search, social media, ridesharing, online marketplaces, online dating, and advertising.

As social computing researchers, in relation to the positionality of this work, we find it important to draw attention to the way the artifacts we study are usually specific to a time and place, rather than being universal or permanent. This influences our work in three important ways.

First, our own experience is necessarily limited by the contexts in which we have gained that experience. While we seek to provide a broad range of examples in this work, we focus many of those examples in Sections 3 and 4 on audits of search engines, where we have a particular depth of expertise. Further, many of the articles we reference come from the U.S. context; auditing itself is a broadly applicable practice, but the systems being audited and legal contexts surrounding audits vary widely, and the U.S. context is the one with which we are most familiar.

Second, the context dependence of social computing research impacts the goals of this article and its contributions. Since we expect these systems to develop and change over time, we seek to strike a balance between providing enough concrete details that other researchers in this domain can draw practical guidance from this work, while also focusing at a sufficiently high-level such that future researchers can understand the current moment from which we write—the motivations and considerations currently entailed in studying search after the specific details are deprecated.

Finally, as social computing researchers we also wish to draw attention to the potential for algorithm auditing to have significant political implications, a position we elaborate upon in Section 5. The algorithms that researchers such as ourselves audit are neither inevitable nor unchanging; rather, they are constantly in flux, and both constructed

and used by people, and our work as auditors has the potential to change them, and in doing so to change the society in which they exist. As has been argued by scholars from the related field of Science and Technology Studies, ownership over algorithmic tools and data, along with the ability to monitor and understand them, increasingly yields power in our society (Milan and Van Der Velden, 2016; Chun, 2011). The possibility for direct change precipitated by an audit presents great opportunity as well as risk, and we hope this work will help researchers consider the politically weighty and socially important aspect of the work at hand as deeply as the technical advice we can provide.

1.4 Road Map

In the sections that follow, we aim to provide readers with an understanding of the algorithm auditing method, including its history and best practices. To do so, in Section 2 we begin by describing the auditing method's roots in the social sciences, prior to its use in the digital realm. Next, in Section 3, we move our focus to algorithm auditing, describing the method itself and summarizing key domains in which it is applied along with notable algorithm audits. In Section 4, we decompose algorithm audits into nine key dimensions, describing the choices available to auditors and providing recommended best practices within each. Before concluding, in Section 5, we further discuss the social implications of conducting audits and advocate for auditors to view this work through the lens of its broader social impacts.

2

The Audit Study: Social Science

To give context to today's algorithm audits, we first provide an overview of audit studies in the social sciences, from which the technique of algorithm audits was drawn. In this section we will cover the basics of different types of audits as they were originally conceived, along with delving into a few examples of high-profile audits, like the one conducted by Bertrand and Mullainathan (2004) studying racial bias in hiring that we discussed in the previous section. We will contextualize this work with background on nondiscrimination legislation in the U.S. (including disparate treatment and disparate impact liability) that makes discrimination-focused audits particularly impactful in the real world.

Audits emerged as a method in the 1940s and 1950s, though they were mostly small-scale. It was not until the 1960s in England that auditing for racial discrimination became part of a Parliament-mandated effort and audits began being conducted at scale. Gaddis (2018) gives an excellent summary of the history auditing, which we encourage those interested in a detailed historical account to read. Such auditing has continued through the present day, helping governments, researchers, activists, and private entities identify phenomena, especially discrimina-

tion and unequal treatment, that are otherwise difficult to examine and often only emerge when analyzing in aggregate.

Audit studies have historically fallen into one of two main categories: field audits, and correspondence audits. Field audits are conducted in person (in the field), by sending a researcher or a researcher's accomplice to collect data. For instance, in one of the first audits, conducted in England and published in 1968, one of three trained assistants—one white and British, the second a white immigrant, and the third belonging to a racial minority group—was sent to apply for housing (Daniel, 1968). This experiment served to quantify the discrimination immigrants and racial minorities faced in housing, and led to the passage of updated anti-discrimination legislation in England (Gaddis, 2018). While effective, field audits are difficult to conduct at scale, and may present additional issues controlling for all differences between the associates sent to conduct the audit.

Since the 1980s, correspondence audits have become more common. In these audits, materials (e.g., hypothetical resumes created by researchers) are sent out (e.g., to employers), rather than dispatching actual people to collect data. Similar outcomes are measured, also received by correspondence (e.g., phone calls or postal mail inviting a candidate to interview). This strategy can allow tighter experimental control, as researchers have total command over the materials being sent, and can ensure (with some effort) that materials sent only differ along the researcher's axis of interest (Siegelman and Heckman, 1993).

The availability of new technologies have led to substantial changes in the way audits are conducted. In addition to later creating the possibility of auditing algorithms, software technologies have allowed researchers to, for instance, computationally and automatically produce thousands of distinct resumes, be which can then be electronically delivered to as many employers (Oreopoulos, 2011). Between the greater capacity for experimental control and increase in scalability, correspondence audits in conjunction with computational methods have allowed researchers in recent decades to overcome many of the largest challenges that faced the first auditors.

2.1 Common Auditing Domains

Separate from the type of audit method being deployed, there are two main axes to an audit study. The first is the category of discrepancy being investigated—often audits focus on detecting discrimination on the basis of a protected class like race, sex, nationality or ability. This axis is the researcher's independent variable, which is varied in each condition of the study. The second important axis is the domain of study, which is invariable within the confines of the study—for example, housing, employment, and healthcare are popular domains in which audits are deployed. For maximum impact, these dimensions are usually directly interrelated; the audit is targeted at a specific type of discrepancy situated within a specific domain, based on inequities observed or hypothesized to exist in the real world.

For historical context and inspiration for today's algorithm auditors, we discuss the two most common domains for traditional audits in more detail: housing and employment audits. These domains have the longest history of auditing due to their social importance, since social equity and justice have been the primary motivation of audit studies in the social sciences, and the existence of international legislation against discrimination in these domains.

Following in the footsteps of the first audit studies on racial discrimination in England, auditing for discrimination in housing has been very influential. In the U.S., the Department of Housing and Urban Development commissioned audits throughout the 1970s and 1980s to study race-based discrimination in housing markets around the country, including commissioning a large-scale audit of forty metropolitan areas in 1977 (Gaddis, 2018). These studies have been influential both because of the importance of secure housing to citizens' well-being, as well as the substantial history of racial and ethnic discrimination in housing and its the long-term effects, such as redlining—federal, state, and local policies that prevented Black Americans from accessing home ownership aid, while offering loans and subsidized housing to white Americans (Gross, 2017).

Another major domain for audits has been employment. High-profile audits in the past have identified, for instance, that employers' hiring

processes discriminate against African-American job applicants relative to white ones, a finding that holds for both men and women (Bertrand and Mullainathan, 2004). A similar method has been used to show employment bias along other axes, for instance against mothers relative to fathers and childless job seekers, comparing responses to resumes for women with and without children against men with and without children (Correll *et al.*, 2007). Note that both of these audits, and many others, study discrimination intersectionally—varying more than one identity characteristic in order to understand the interacting and compounding effects of different identity aspects, rather than generalizing about people's experiences or treatment across only one aspect (Crenshaw, 1990).

While this pair of areas is the focus of many audits, there are many more domains of focus as well. Healthcare is one of those, with researchers studying the likelihood of healthcare professionals from primary care doctors and psychotherapists to accept new clients by race or socioeconomic status (Sharma *et al.*, 2015; Kugelmass, 2016). Other work has studied market transactions, finding discrimination against disabled people (Gneezy and List, 2004) and racial/ethnic minorities (Yinger, 1998). Other consumer-facing services like financial advice have also been audited for bias (Mullainathan *et al.*, 2012). Recent work of particular interest to algorithm auditors by Hutchinson and Mitchell (2019) has investigated the concept of fairness through a review of work in areas like educational testing. For a detailed list of domains of audits since the mid-twentieth century and the types of discrimination they studied, see Gaddis (2018).

Driven by researchers' own interests and expertise, virtually any domain of social importance can be the focus of an audit study. The consequences of these audits also depend on the domain and degree of discrimination found. This brings us to our next area of focus: the legal context surrounding audits. While not all audits will find significant discrimination and not all types of discrimination are illegal, in some cases there are important legal ramifications to an audit study.

2.2 Legal Context and Impact

We will briefly discuss the main legal framework surrounding audits, nondiscrimination law, as well as criticism and limitations of audit studies over the last several decades.

Each country has its own legal system and context; in many, nondiscrimination laws enforce limits on the ways different types of people can be treated. As we mentioned above, however, these laws generally only apply in certain domains: commonly, employment and housing. Without getting into international nondiscrimination law in much detail, we will briefly cover two important concepts: disparate treatment and disparate impact discrimination. Disparate treatment refers to the act of explicitly (intentionally) treating different classes of people differently, such as a company policy of only hiring employees who are bilingual for a certain role. Disparate impact, meanwhile, describes practices that result in significantly different outcomes for members of different groups in the absence of explicit intent. For instance, if men are promoted more frequently than women at a certain company due to managers' own unconscious biases or other unintentional aspects of the promotion process, this may constitute disparate impact.

Notably, the examples given above are both in the context of employment. In the United States, these two concepts are codified by Title VII of the Civil Rights Act of 1964, which prohibits disparate treatment (barring a "business necessity") and disparate impact in employment practices. These legal standards do not necessarily apply in other domains or in other countries, so care must be taken to understand the specific legal context that might bear upon audit study and its consequences. Regardless of legal standing, audit studies are a powerful tool for establishing evidence of discrimination, especially disparate impact, since this form of discrimination is by definition not explicit and often only becomes visible in aggregate.

For all their uses, there are some limitations to traditional audit studies, however. In the past, and even today, audit studies are challenging to scale; this is especially true of field audits where someone must personally execute each data point. This is becoming less of an issue with computational methods, as we described above with the advent of

tools for the automated generation of materials used in correspondence audits. Additionally, care must be taken when conducting such audits not to mix signals; it can be difficult to control the materials used in an audit for correlated and indirect signals (e.g., studying race-based discrimination by sending resumes with names that may inadvertently communicate class status as well as race).

These and other challenges aside, audit studies have been used to successfully identify discrimination in a wide range of domains for over half a century. They remain powerful and relevant for their strength in uncovering implicit and important biases across our societies.

3

Algorithm Audits

Intersecting the rise in the popularity of auditing, the turn of the century led to the advent of systems powered by algorithms, like search engines and social networks. The audit method, designed for drawing inferences about the workings of complex and opaque systems, is naturally a strong fit for studying such systems, and in recent years this has culminated in the development of a modern class of audit study focused on auditing algorithms: the aptly-named algorithm audit.

The term "algorithm audit" (and not "*algorithmic* audit," which might refer to a traditional audit study done with some automated components) was proposed in 2014 by Sandvig *et al.* (2014) and provides a unifying label for prior and current research conducted in this vein. Similar to audit studies from the social sciences, these studies often involve investigations of discrimination and causality, but instead of investigating human inputs (e.g., resumes) and their corresponding responses (e.g., a call back), they study those of automated systems, powered by algorithms. In this section, we describe the algorithm audit method and also detail, as we did in the previous section for traditional audits, some domains in which it has been successfully employed. While algorithm audits can be targeted at a broad range of systems, we focus mostly on audits of search engines.

3.1 What is an Algorithm Audit?

Defining this method in more detail, an algorithm audit is a method of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs in order to draw inferences about its opaque inner workings. The techniques used in algorithm audits vary widely because they are often conducted in distinct, often online, environments that cover a wide range of domains, including traditional domains like housing, employment, product pricing, and health. For instance, as with social science audits of redlining in real estate, researchers have studied how algorithms contribute to similar biases online.

To give potential auditors context on the wide range of algorithm audits that have been conducted, we next provide some information and relevant citations for algorithm audits across domains from housing through healthcare. We do, however, stop short of a systematic review; for interested readers, we recommend Bandy (2021) for a systematic review on audits, Barocas *et al.* (2017) on fairness in automated systems more broadly, and Koshiyama *et al.* (2021) for a perspective on algorithm audits that parallels financial auditing.

3.2 Algorithm Auditing Domains

Beginning with the first auditing domain that began offline, algorithm auditors have examined "digital redlining", racial housing discrimination online, on the Airbnb apartment rental platform with respect to both the platforms' hosts (Edelman and Luca, 2014) and visitors (Edelman *et al.*, 2017). Others have expanded their scope to include real estate sites and ads more broadly, continuing to make housing audits as popular and impactful a domain for auditing online as it is offline (Asplund *et al.*, 2020b).

Employment is another classic domain for audit studies, and with employers and hiring companies increasingly conducting business online, algorithm audits provide tools to evaluate how they identify, rank, and present candidates. Audits have, for instance, identified gender discrimination in employment ads (Speicher *et al.*, 2018) and on hiring websites (Chen *et al.*, 2018) among others.

The migration of many social functions online has also opened new domains entirely for auditors. One major such area is media consumption. For instance, the role of digital spaces like social media sites and online news aggregators, in people's news access and consumption has led to a number of audits examining content exposure through such platforms (Bandy and Diakopoulos, 2020; Bechmann and Nielbo, 2018). Other work has audited music streaming sites to understand the impact of such recommendation systems on users' exposure to musical artists by gender (Eriksson and Johansson, 2017). Twitter recently crowdsourced audits of its platform, as part of their approach to identifying bias on the platform (Chowdhury and Williams, 2021).

Another new area for auditing is the so-called "sharing economy," a term describing peer-to-peer sharing of resources, with transactions often managed by corporate platforms. These companies—including ride-sharing firms like Uber and Lyft, housing companies like Airbnb, crowd funding platforms, and others—typically connect their contractors and customers through mobile and web applications using algorithms to perform the matching. As a result, they comprise a new online-only domain for algorithm audits. In one example, researchers built infrastructure to emulate dozens of Uber accounts in order to collect data about the function of surge pricing from the notoriously tight-lipped ridesharing company (Chen *et al.*, 2015).

As with traditional audits, one of the most important domains for auditing is healthcare, in particular as algorithmic systems are increasingly used by both patients seeking health information and doctors using technologies to diagnose and care for patients. On the patient side, one recent article audited the spread of COVID-19-related information online (Makhortykh *et al.*, 2020), and another audited the YouTube algorithm for misinformation on a variety of health topics (Hussein *et al.*, 2020). Other work has audited user queries to better understand people's health information needs (Abebe *et al.*, 2019). On the doctor's side, one highly publicized audit found that algorithms used for assessing health risks systematically underestimating Black patients' healthcare needs (Obermeyer *et al.*, 2019).

Consumer markets have also been studied using algorithm audits, generally focusing on online marketplaces and product pricing. Sit-

ting between this domain and the previous, healthcare, Juneja and Mitra (2021) have recently audited health misinformation appearing on e-commerce platforms. Other such investigations stem from concerns about price steering, the practice of charging consumers different amounts of the same items by personalizing according to consumer attributes. Algorithm audits conducted by academic researchers (Mikians *et al.*, 2012; Hannak *et al.*, 2014) as well as journalists (Angwin and Mattu, 2016) have identified and measured such bias on platforms like Amazon.

Finally, auditors have also targeted other forms of commercial software and decision-making systems for audits. Facial recognition systems are one very high-profile example after researchers found that software performance correlated with race and gender, performing best on lighter-skinned male faces and worst on darker female faces (Buolamwini and Gebru, 2018). A second famous example comes from ProPublica, where journalists analyzed data from an algorithm used in the legal system for evidence of racial bias against Black defendents (Angwin *et al.*, 2016). The job of uncovering flaws, biases, or other issues in non-user-facing systems merges algorithm auditing with the related topic of studying fairness in machine learning. Without going into this large field with much depth, a couple examples include work identfiying anti-Muslim sentiment in language models (Abid *et al.*, 2021) and age-related bias in sentiment analysis tools (Díaz *et al.*, 2018).

Above we've categorized algorithm audits by the domain in which they are conducted, while the platforms on which they are conducted vary widely. As a case study that will later provide motivating examples in Section 4, we next summarize the extensive literature of audits on one specific class of platform: search engines.

3.3 Search Algorithms: An Important Subclass of Algorithm Audits

There has been a wealth of research for more than two decades that specifically focused on audits of search engines due to their wide-spread use and ability to influence users, but perhaps also due to the relative ease of auditing such systems. Search engines are some of the most widely used systems, with over 90% of online adults using them as of

2012, and widely trusted, with the majority believing them to be "fair and unbiased," at least as of that year (Purcell and Brenner, 2012). Search engines are also some of the most widely audited algorithms, as found by a recent systematic review of algorithm auditing (Bandy, 2021). It bears note, however, that search systems may in some ways be more easily audited than other systems (e.g., social media sites, healthcare systems, or government systems), leading to a convenience sample in the literature. For instance, in contrast with many of the studies we describe below, recent attempts at auditing political ads served to users on Facebook by external academics were shut down by the company revoking researcher access to the platform (Bobrowsky, 2021).

Early studies of search engine coverage and bias date as far back as 1999 and continued as search engines, especially those for web search like Google, gained influence (Lawrence and Giles, 1999; Mowshowitz and Kawaguchi, 2002b; Mowshowitz and Kawaguchi, 2005; Fortunato *et al.*, 2006; Goldman, 2008). Sometimes these efforts uncovered bias with financial motives, either on the part of the search engines or bad actors external to the companies; in 2010, for instance, Google came under scrutiny for allegedly adding hard-coded rules in its algorithm to put its own products at the top of the page (Sandvig *et al.*, 2014).

In recent years, some of the most prominent topics of interest for search engine audits have focused on broader social issues. One line of such works investigated how search engines could be used to influence the political process of the societies in which they are situated. Studies suggested that web spammers with political motives have been trying to game search engines since at least 2006 to surface content that benefits them (Metaxas and Mustafaraj, 2009). Theoretical work since the early 2000s has also examined the implications of biased search results for democracy (Introna and Nissenbaum, 2000; Granka, 2010). This work includes behavioral experiments demonstrating how search results' rankings can influence user preferences (Epstein and Robertson, 2015; Epstein *et al.*, 2017), as well as audits focusing on representation of news in search results (Trielli and Diakopoulos, 2019; Kawakami *et al.*, 2020), political candidates (Metaxa *et al.*, 2019; Diakopoulos *et al.*, 2018; Robertson *et al.*, 2019), presence or absence of political content (Hu

et al., 2019), and users' political query formulation (Mustafaraj *et al.*, 2020; Trielli and Diakopoulos, 2020).

In addition to their prominent roles in political processes, search engines play a central role in mediating people's access to high-stakes information, including medical and health-related content. Work in this domain has included online behavioral experiments (Allam *et al.*, 2014) and work focusing on users' experiences with search when looking for health information (White and Horvitz, 2009), while other research has taken a data science approach, including research studying whether real-world phenomena like influenza outbreaks can be identified by collecting user search behavior in aggregate (with mixed results) (Lazer *et al.*, 2014).

Finally, leading researchers in this space have critically examined search algorithms' interactions with race and gender, including depictions of people of color in web search (Sweeney, 2013b; Noble, 2018a; Noble, 2013) and auditing for whether sites ranking resumes (Chen *et al.*, 2018). Other researchers have studied whether search algorithms serve their users equally across demographic categories (Mehrotra *et al.*, 2017), and have developed metrics that could help practitioners quantify and correct such unfairness (e.g., Speicher *et al.*, 2018; Kulshrestha *et al.*, 2019).

3.4 Legal Context

Though it is still developing, we will next address the legal landscape surrounding algorithm audits. In Section 2, we discussed the way legislation paved the way for the first audits, which were done at the behest of various governments seeking to verify that nondiscrimination laws were being followed. Nondiscrimination law motivating audit studies can still be important in the context of algorithm audits. The more important legislation pertaining to algorithm audits, however, addresses hacking and computer fraud, and can potentially put algorithm auditors—even academic researchers—at risk.

Algorithm audits began to be used before there existed laws and regulations pertaining to the technique, a result of the fact that internet technologies have been created and developed so quickly in recent years.

Here we will focus mainly on law and policy in the United States, since most of the largest corporate players in this domain and much case law has been within that jurisdiction. However, readers should remain attuned to legislation coming from the European Union, China, and other areas where technology companies are gaining influence and different legislatorial regimes hold.

As a result of this legal uncertainty along with hostility from companies fearing public reveal of their products' flaws, researchers have been vulnerable to lawsuits for conducting algorithm audits. For context, in the United States, beginning in 2010 companies began filing lawsuits for some Terms of Service (TOS) violations under the Computer Fraud and Abuse Act (CFAA) (Zetter, 2010). The intention of the CFAA was to prevent anyone from accessing another's computer "without authorization," making such access a criminal offense. Legal disputes debated what constituted "authorization" and with precedent it came to include TOS-violating behaviors like scraping a website in addition to behavior more conditionally considered hacking (Robertson, 2019). Such a reading of the CFAA has important implications for algorithm auditors, since such audits usually involve some behavior that would also violate a platform's TOS, like scraping webpages' content, or making multiple accounts.

This risk of legal liability in reaction to algorithm auditing, even when done by researchers in an academic, not-for-profit context, led some researchers (among them some of the authors of this monograph) to file a lawsuit against the U.S. Justice Department with the help of the American Civil Liberties Union (ALCU) challenging the CFAA (Union, 2019). That case, known as Sandvig v. Barr, led to a judgement by the U.S. 9th Circuit Court of Appeals in September of 2019 concluding that violating a website's TOS does *not* violate the CFAA (Robertson, 2019). This finding was confirmed on appeal by a federal judge in Washington, D.C., in mid-2020. A related case, Van Buren v. United States, affirmed this ruling at the federal level in June of 2021 (Thomas, 2021).

Where does this leave algorithm auditors? The latest word in the United States legal system affirms that algorithm auditing research in violation of a site's TOS is legal. However, the legal landscape is sure to continue developing over time, and these rulings do not apply in

non-U.S. jurisdictions. In conclusion, as we will discuss at more length in the next section regarding best practices for algorithm audits, algorithm auditors should continue to closely monitor this aspect of the work.

4

Best Practices

Based on prior work and discussions among the authors, we lay out nine main dimensions for consideration when auditing algorithmic systems. To make our discussion more concrete, we primarily focus on search engines—some of the most prevalent and highly-studied algorithmic systems with the immense power to influence people's preferences and behaviors (Introna and Nissenbaum, 2000; Pan *et al.*, 2007; Joachims *et al.*, 2007; Epstein and Robertson, 2015)—as a case study and explore how these key areas of consideration can be applied to auditing them. Supporting with relevant examples from prior literature, we provide recommended best practices in each key area, and discuss how our recommendations can inform researchers auditing a range of algorithms. We also summarize all nine best practices in Table 4.1 for ease of reference at the end of the section.

These areas of considerations and the choices within each were identified through our own extensive experience conducting such research, and by systematically examining prior work. We first compiled a list of works in this field that we consider particularly strong, either based on measures such as citation counts or our own interactions with those works. Next, using the citations referenced in those papers, we

expanded to a list (not intended to be exhaustive) of over three dozen published papers. Three authors then annotated each of these papers for key decision points (e.g., choice of algorithmic systems, choice of input for the algorithmic system, etc.). The research team then iteratively brainstormed other possible decision points, and agreed on a grouping into the nine key areas presented here. Finally, during a series of group discussions, we produced recommendations for each area based on our own experiences and opinions as algorithm audit researchers.

4.1 Legal and Ethical Considerations

The central importance of legal and ethical considerations in this space, especially when seeking to identify and call attention to problematic, irresponsible, or harmful algorithms, leads us to begin our guidelines with a discussion of this aspect of the work. Given the enormous potential of algorithmic systems to influence users' preferences, beliefs, and behaviors, researchers should be aware of relevant laws; also respect services that are being audited and their users; and make informed decisions on whether and what to audit based on risk and personal ethics.

In this section, we first discuss some key issues of which auditors should be aware, including the many potential costs of auditing, recent legal action in the U.S. context affecting auditors, and a discussion of instances in which it may actually be harmful to conduct an audit. As we will show throughout those discussions, most of these choices must be made at the discretion of individual researchers; to close this section, we will describe some alternate ideas, including recent proposals for more formalized ethical review processes in the academic setting.

4.1.1 Costs of Auditing

Barring cases in which data is, for example, provided by a company, auditing algorithmic systems almost always involves interacting with users and/or the systems to collect data. It is of ethical as well as practical importance that we as researchers are mindful of our impact on the ecosystem we study and behave responsibly. These costs can

include those related to human attention (especially the users', though also including the researchers' and system designers'), computational resources (the researchers' and/or the system's), financial and monetary resources (the researchers' and/or the system's), environmental resources (incurred, for example, by energy use in running audits), and others. Below we describe the first two, human impacts and computational resources, in more detail.

Regarding the impact on users, in addition to standard human-subjects research, studying algorithmic systems present some new risks. For instance, when auditing a search engine, recruiting participants for the sake of collecting search engine data on users' local machines is among existing strategies (Robertson *et al.*, 2018a). When using such strategies, we must take care to respect query rate limits imposed by search engines lest our participants experience a loss of service due to their participation. It is similarly important to consider the nature of the queries being conducted, as personalization based on a user's search history may affect their experience across the web after the fact.

It is also important to consider the stress placed on the system being studied. For example, conducting many queries on a niche search engine in quick succession or in parallel may overburden the search engine and slow the service or even unintentionally bring it down, negatively impacting the search engine and its other users. Although it is unlikely that audit studies conducted in an academic setting may bring noticeable burdens on, for instance, giant commercial search engines like Google and Bing, this may not be the case for smaller algorithmic systems. The appropriate choices will depend on each case's specifics, but we encourage researchers to consider the size of the service being studied to estimate an appropriate load.

4.1.2 Regulatory Violations

Almost all widely used algorithmic systems today are proprietary and governed by terms of service (TOS) agreements and other legislation limiting what users, including auditors, can do with these systems. Violations of relevant legal standards poses a very real risk to many auditors, who also be aware of the ever-changing landscape of relevant

laws. Below we detail that landscape surrounding one particular legal liability in the United States, terms of service violations Subsequently, we briefly discuss non-governmental regulations, such as those enforced by professional organizations. At the outset, we note that our review in this subsection cannot be considered as a comprehensive summary that covers every jurisdiction; we stress that the researchers must be aware of applicable laws and policies, and be judicious when conducting an audit to make sure that the risks conducting the audit do not outweigh the gains.

In the past, anyone in the United States violating a platform's Terms of Service—from users registering under a fake name to researchers scraping website content to audit for discrimination—could be charged with violating the Computer Fraud and Abuse Act (CFAA) (Zetter, 2015). Enacted in 1984, the CFAA forbids users' access to a computer in excess of authorization. For example, in the context of search engines many general-purpose web search engines forbid actions such as accessing the service through means other than the interface that is provided (*Google Terms of Service* 2017), or monitoring and storing the content of their results (Karahalios, 2018). Based on those TOS agreements, the CFAA has been interpreted as limiting researchers' ability to legally collect the search engine outputs (Zetter, 2016), and terms of service violations have been pursued within the CFAA (Zetter, 2016; Zetter, 2015; Zetter, 2010).

In response, there have been multiple calls from government and academia advocating researchers' right to study algorithmic systems. For instance, in a May 2016 White House report covering five national priorities that are "essential" for developing big data technologies, the Executive Office of the President underlined the importance of researchers' ability to "investigate normatively significant instances of discrimination involving computer algorithms" through the process of algorithm audits (President, 2016). The following month, a group of U.S. researchers filed a federal lawsuit, Sandvig v. Barr, asking that the government not criminally prosecute researchers conducting such studies (Zetter, 2016). The American Civil Liberties Union, which argued the case on behalf of Sandvig, argued that auditing practices are not illegal offline, and therefore should not be illegal online either (Thomas, 2021).

As discussed in the previous section, another related lawsuit, Van Buren v. United States, recently resulted in a Supreme Court victory for auditors, with the US federal court determining that "research aimed at uncovering whether online algorithms result in racial, gender, or other discrimination does not violate the Computer Fraud and Abuse Act" (ACLU, 2020). This victory is significant for algorithm auditors, a long-awaited affirmation of the legality of violating TOS while conducting such research. However, it is only one example of regulatory action of which auditors need to be aware. Researchers should become familiar with the legal standards of the jurisdiction from which they plan to conduct the audit. EFF's Coders' Rights Project documentation is a good resource that summarizes many of the key legal concerns related to this current discussion (*Coders' Rights Project* 2020).

In addition to governmental regulations on this topic, researchers must also be aware of their own professional organizations' stances. In 2018, the Association for Computing Machinery updated its Code of Ethics and Professional Conduct to specify that, "computing professionals must abide by [laws and other regulations] unless there is a compelling ethical justification to do otherwise" (Computing Machinery, 1992). Prior to this update, research involving a TOS violation was also in violation of the world's largest computing society. Such restrictions from professional organizations could prevent researchers from publishing their work in some venues, and warrant awareness by auditors prior to beginning an audit.

4.1.3 To Audit or Not to Audit

Earlier in this section, we have laid out potential costs and some legal considerations surrounding the conducting of audits. However, critics have rightfully pointed out that processes like audits can potentially serve to improve inherently problematic systems on metrics that are immaterial to the ethics of those systems, or even legitimize systems that are inherently harmful and should instead be wholesale opposed.

Illustrating this risk, Keyes *et al.* (2019) published a provocation proposing a review of the fairness, accountability, and transparency of

a system designed for "mulching" human beings into food products, humorously demonstrating the real possibility that audits might ensure equitable, accountable, and transparent enforcement of a system that should never exist in the first place. As this work suggests, audits are not a guarantee for ensuring ethical or pro-social outcomes, especially if auditors restrict their focus with the assumption that the existence and use of the system they are studying is a given.

In a more serious tone, Sloane (2021) also argues that algorithm audits risk use for legitimizing problematic systems, in the context of recent legislation in New York City that calls for "bias audits" of automated hiring and employment tools. One issue Sloane points out is the lack of clear guidance on how to conduct such an audit—an explicit aim of this work. Sloane also points out, though, that companies may use audits as smokescreens behind which to hide their problematic technologies, and that independent researchers can become complicit when agreeing to conduct audits on behalf of such groups.

In addition to considering whether the audit itself should be conducted, there is also the matter of who should conduct it. We recommend that auditors, whether from academia or industry, tend towards taking an impartial, third-party role—and in doing so, consider their funding sources and disclosures and other ties to the industry or company being audited. In short, auditors evaluating the ethics of a potential audit need to consider whether the audit should be done at all, under what circumstances, and by whom. We engage with these issues in some greater depth in Section 5, which focuses on the social impacts, good and bad, that audits can bring.

4.1.4 Formalizing Reviews of Research Ethics

As we have described, there are various forms of regulatory guidance and restrictions that pertain to the legality of conducting algorithm audits—but the ethical aspect of auditing is largely a consideration made at the discretion of individual auditors. Recently, some researchers have begun calling for a more formalized ethical review of research projects, at least in the academic setting.

One proposal, from Prosperi and Bian (2019), points out that academic research involving human subjects' testing must be reviewed by a university Institutional Review Board, or IRB. IRB review, however, is not meant to holistically evaluate the ethics of a project; rather, its aim is to protect the "rights and welfare" of human subjects involved in research. Moreover, in projects collecting and analyzing publicly available data, which includes some algorithm audits, the public nature of the data used means that the individuals whose data is collected may not be considered human subjects by the IRB, and such work would not require IRB review. Prosperi and Bian (2019) and others have, therefore, proposed that IRBs should adapt to this new research strategy and expand the purview of their review.

A second idea, from Bernstein *et al.* (2021), proposes an alternative to IRB review, "Ethics and Society Review (ESR)". Under this strategy, all applications for a specific major source of artificial intelligence-related grant funding at Stanford University were required to submit proposals evaluating the social risks, harms, and their planned mitigations. The proposals were then reviewed by an interdisciplinary panel of researchers at the institution, who provided feedback for grant applicants to iterate on. In its first implementation, nearly 60% of the researchers involved indicated that the ESR influenced the design of their project. At a high level, ESR represents one of many possible alternatives to IRB review that might explicitly consider the ethical dimensions of academic research.

Without advocating for either of these specific ideas (or any number of others that might be proposed), we think it is important that readers be aware of the movement towards more formalized ethical reviews of academic research, as resulting structures and restrictions will undoubtedly impact the future of algorithm audits.

Recommendation 1: We encourage researchers to be aware of relevant laws, their comfort with legal risk, and their own ethics when choosing to conduct search audits. Beyond this, researchers should respect algorithmic services and users, being mindful of the impact of their research on the same.

4.2 Selecting a Research Topic

As algorithmic systems are increasingly deployed across so many online domains, a plethora of topics stand to benefit from studies of these systems. Although the choice of topic is largely up to researchers' own interest, it is worth noting that topics related to identifying inequality and discrimination are particularly integral to the tradition of audit studies (Sandvig *et al.*, 2014). Here, we outline existing work in several domains that have been particularly fruitful directions for previous studies of search engines. It should be noted that this list is not meant to be exhaustive but meant to inform researchers interested in conducting audit studies of important existing threads of studies. Additionally, these topics can be adopted to study algorithmic systems beyond search engines.

4.2.1 Discrimination and Bias

One area of focus relates to discrimination and bias. While bias can broadly refer to any set of results presented which are not representative of the set from which they are drawn (also called "content bias" in some prior work (Pitoura *et al.*, 2018)), here we are referring to depictions of individuals or groups in search engines (particularly members of legally protected or socially marginalized categories), and equality of access to information for those individuals via search engines. While the language used surrounding this topic is still developing, some scholars have pointed out the potential for the term "bias" to obscure the role of systemic power in favor of a fallacious individualist framing (Dave, 2019). Other proposed terminology includes "algorithmic harms" (Dave, 2019) or "algorithmic oppression" (Noble, 2018a).

Leading researchers in this space have critically examined search algorithms' interactions with race and gender, including depictions of people of color in web search (Noble, 2018a; Sweeney, 2013b), auditing for whether sites ranking resumes (Chen *et al.*, 2018) or scoring workers in the gig economy (Hannak *et al.*, 2017) are biased, or whether image search algorithms present biased reflections of common occupations by gender (Kay *et al.*, 2015). Still other researchers have studied whether

search algorithms serve their users equally across demographic categories (Mehrotra *et al.*, 2017), and have developed metrics that could help practitioners quantify and correct such unfairness (Sapiezynski *et al.*, 2019). Outside the Computer Supported Cooperative Work (CSCW) community, in which many of such works have taken place, considering work from marginalized people and from scholars in other disciplines (including Science and Technology Studies, law, and many social science disciplines) with a history of studying discrimination can also situate and provide theoretical depth to audit studies.

4.2.2 Politics

Another of the most notable areas of prior audit studies for search engines is the political sphere, including misinformation, disinformation, propaganda, partisanship, and political polarization research. This includes audits focusing on news (Trielli and Diakopoulos, 2019), political candidates (Metaxa *et al.*, 2019; Diakopoulos *et al.*, 2018; Metaxas and Pruksachatkun, 2017), presence or absence of political content (Robertson *et al.*, 2019; Hu *et al.*, 2019), and comparison of different search engines' results in the domain of politics (Kulshrestha *et al.*, 2019; Mowshowitz and Kawaguchi, 2002a; Mowshowitz and Kawaguchi, 2005). Aside from the content itself, the other branch of such work has focused on user impact, for instance by conducting online and laboratory-based behavioral experiments (Epstein and Robertson, 2015; Epstein *et al.*, 2017), qualitatively analyzing users' search patterns (Tripodi, 2018), investigating the possibility for filter bubbles or imbalanced partisan content due to personalization (DuckDuckGo, 2018; Hannak *et al.*, 2013; Robertson *et al.*, 2018b), or studying user web search behavior when peer-produced content is included or removed (McMahon *et al.*, 2017; Rothshild *et al.*, 2019).

4.2.3 Health

A third major area of study that has garnered much attention from researchers and the public is search in the context of health information. In this domain too, there have been online behavioral experiments (Allam *et al.*, 2014) and work focusing on users' experiences with search

when looking for health information (Cartright *et al.*, 2011; White and Horvitz, 2009; De Choudhury *et al.*, 2014). Meanwhile, other work has taken a data science approach, including research studying whether real-world phenomena like influenza outbreaks can be identified by collecting user search behavior in aggregate, with mixed results (Ginsberg *et al.*, 2009; Lampos *et al.*, 2015; Lazer *et al.*, 2014), or identifying health information needs by mining users' queries (Abebe *et al.*, 2019).

Recommendation 2: Regardless of the domain of study, we encourage researchers to choose areas with potential for social impact and, where necessary, involve and collaborate with domain experts and key stakeholders—including social scientists, law and policy experts, and users themselves.

4.3 Selecting an Algorithm to Audit

When deciding what algorithms to study (e.g., when studying search engines, which search engine(s) to audit), motivate this decision using metrics of real-world influence of the systems in question (e.g., by focusing on search engines with the largest market share or widespread use among the population of interest). Much recent work in auditing search engines focuses on Google for this reason, but it's worth considering other powerful search engines, including those used more in an international context, where market dynamics may be very different (e.g., China, where Baidu, rather than Google, holds a majority of the market share (*Search Engine Market Share China* 2019)). The same suggestion would apply when researchers are auditing algorithmic systems that are not search engines, such as social media news feed algorithms. As outlined by Sandvig et al., the purpose of audit studies, whether auditing algorithmic systems like search engines or other social processes, is generally to detect harmful discrimination that impacts the society in which the audited systems operate (Sandvig *et al.*, 2014). When choosing the subject of an audit, it is important to consider systems that are widely used and have the broadest societal impact on the population of interest.

4.3.1 International Differences

Consider, as one highly-studied example domain, audits of web search engines. In the United States and in many international markets, Google has emerged as the most dominant web search engine, and as a result many studies of search media focus on it (Robertson *et al.*, 2019; Metaxa *et al.*, 2019; Trielli and Diakopoulos, 2019; Diakopoulos *et al.*, 2018). As of 2019, the estimated international market share of Google is as high as over 90%, with Baidu coming in as a distant second at approximately 3% (*Search Engine Market Share Worldwide* 2019). In other contexts, however, other search engines may be more worthwhile subjects; since 2010, for instance, Google has been blocked by the Great Firewall in the People's Republic of China, limiting the ability for users in mainland China and Hong Kong to query that search engine (*Google China* 2021). As a result, other web search engines are important to study in the Chinese context, including Baidu and Sogou, whose combined market share reached nearly 90% in China as of 2019 (*Search Engine Market Share China* 2019).

Unfortunately, general purpose web search engines like Baidu, Naver, and Yandex, whose market share is largely based outside of the United States, are severely under-studied compared to Google; it is unclear to what extent the findings from studying Google search might generalize to these other services. It is of paramount importance to take into consideration users—their diverse options and behaviors, and their search media landscapes—when scoping these projects.

4.3.2 Comparative Studies

Rather than focusing on only one particular algorithmic system, researchers may also consider actively comparing two or more of the same class of algorithmic system. In the context of studying search engines, this can allow researchers to draw conclusions about the search engine itself, in addition to its output—for instance, recent work by Robertson et al. compared autocomplete suggestions by the two most widely used U.S. search engines, Google and Bing, to investigate whether Google's results favored its subsidiary company, YouTube (Robertson *et al.*, 2019).

Similarly, researchers may choose to focus on different types of search engines—for instance, comparing web search results to those found by searching social media platforms. For example, Kulshrestha et al. investigated Twitter's social media search engine, which provides searching users with suggestions ranging from trending topics to relevant hashtags (Kulshrestha *et al.*, 2019). Doing so allowed them to investigate the features of the social media search engine that are different from a more generic search engine; they found that the social media search engine was more temporally dynamic and more politically left-leaning, potentially due to the politics of that platform's content. In another example, Chen et al. studied the search engines of various hiring sites including Indeed, Monster, and CareerBuilder, finding that they favored men over women candidates to different degrees (Chen *et al.*, 2018). Beyond these examples, there are many other forms of search engines for study, including those used in streaming media (e.g., Youtube, Netflix), online maps (e.g., Google Maps), and e-commerce platforms (e.g., Amazon, Etsy). Studying domain-specific search engines beyond web search can yield insights into how search affects other communities and information-seeking processes.

Recommendation 3: When choosing which algorithm to audit, consider the impact of various relevant options in the space, for example in terms of market share among the population of interest and the type of information it helps users to access, in order to produce relevant and impactful findings.

4.4 Temporal Considerations

While much past work has been based on data collected at a single point in time, many algorithmic systems are becoming ever more responsive and dynamic. As a result, researchers studying algorithmic systems must handle the temporal dimension of their work carefully: when and how frequently is data collected? How do changes in the algorithm or current events affect data over the course of its collection?

Before answering these questions, researchers might question whether it is worth conducting repeated audits, and how such repetitions should be viewed. Beginning with the latter question, one possible lens through which to view repeated audits is that of replication. While replications have an important role to play in research, we suggest that conducting audits repeatedly with the expectation that an underlying algorithm *has changed* should not be considered replications. Unlike in the context of auditing, replication is a term which refers to repeating a procedure in order to provide diagnostic evidence for a claim presented in earlier research (Nosek and Errington, 2020), a framing which does not hold when repeating an audit after an algorithm is expected to have changed. We instead propose that auditors seeking to understand the current state of an algorithm with the expectation that it has changed since its last audit should be seen as conducting re-applications—research that can help shed light on the ways a system has changed over time. Given the potential for these systems to change dynamically, audit re-application is a practice we encourage.

Next, auditors might ask how often to conduct audit re-applications. One key factor in making these decisions is the researcher's estimate of how quickly and significantly the underlying algorithm is changing, as well as what real-world events or phenomena are of interest. For algorithms that change frequently, longitudinal data collection—running the same data collection pipeline weekly or even daily—can provide insight into those changes. For example, Google's search results change frequently in response to current events, so prior work studying Google search results has often involved collecting data daily for a month or more (e.g., Metaxa *et al.*, 2019; Metaxas and Pruksachatkun, 2017). The temporal dimension of such work is also important when researchers are interested in a particular event, such as an election, inauguration, or other political event. In these cases, daily data collection may still be valuable, but it may be sufficient to collect data at a smaller number of time points and comparatively analyze differences in the data (e.g., Robertson *et al.*, 2018b). In other cases—when the algorithm is not expected to change quickly or often, and when no specific events are being studied—collecting data at a single point in time is a common practice (e.g., Kay *et al.*, 2015). However, even in such cases, multiple

rounds of data collection (true replications) may increase the reliability of the data, as changes in the algorithm or notable events are often unpredictable.

Recommendation 4: Consider collecting data at more than one time; while this may add complexity to the data collection process, it can give valuable and unexpected insights into the algorithm and external factors or current events affecting it, or else serve as a check on data robustness.

4.5 Collecting Data

In early efforts to collect search results, some search engines offered an API through which researchers could submit queries (Metaxas and Mustafaraj, 2009), but such APIs are largely unavailable and, when available, may prompt some concerns over the validity of data retrieved that way. In some cases, results returned by the API were shown to differ from what users would have seen when accessing results through the standard interface (McCown and Nelson, 2007). In general, search audits may find more success collecting data without relying on an API.

The simplest approach to collecting search results involves researchers, volunteers, or crowdworkers submitting queries manually and saving the results. The primary limitation of this process is its scalability; even with a large number of volunteers or paid workers, a single person can only submit so many queries for so long. As the list of queries a researcher wishes to conduct grows, this strategy quickly becomes infeasible. Another limitation lies in the validity of comparisons between the queries searched since, for instance, the time at which a search is conducted can have a large impact on the results returned.

4.5.1 Automated Approaches

There are a number of automated approaches to collecting data when auditing algorithmic systems, each with advantages, disadvantages, and varying levels of technical difficulty. We separate these by placing them on a continuum from controlled to ecological (Figure 4.1). Toward the

controlled end, algorithmic responses are instantiated automatically by or on behalf of real or fabricated users by software. On the ecological side, recruited users transmit the algorithm's output from their day-to-day usage back to the researcher. Once again, to make the matter more concrete, we discuss how these approaches get implemented when collecting data from an online search engine with some of the technical details, along with their strengths and weaknesses.

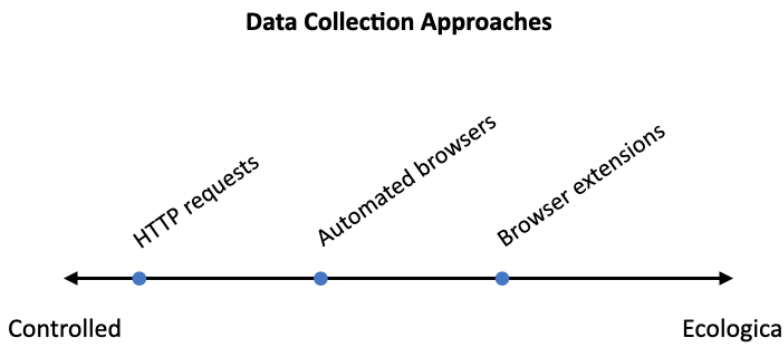

Figure 4.1: Illustration depicting approximately where various data collection approaches lie on the continuum from controlled to ecological data collection.

Due to these constraints, automated methods provide the best option for scalability and replicability. Moreover, with carefully documented data collection processes, and the open-sourcing of the data and/or data collection pipeline, this can aid in repeated audits to monitor the same service over time. However, given the personalized and interactive nature of some algorithms, it is important to be mindful of user privacy when collecting or open-sourcing data.

Automated methods for actively collecting search media include request-, automation-, and extension-based methods. Toward the controlled end of data collection approaches are *request-based methods*, which can be implemented in virtually any programming language, and involve submitting queries to the search engine of interest via an HTTP request from a computer or computers controlled by the researcher.

Automation-based methods are also available in multiple programming languages, and involve using custom software to drive the behavior of a real web browser. Compared to request-based methods, automation-based methods offer two main advantages: (1) the ability to collect and programmatically interact with web pages that include JavaScript-powered elements, like dynamically loading content; (2) the ability to avoid some methods of crawler blocking employed by websites. This latter advantage stems from the fidelity offered by browser automation: unlike request-based programs that simply send HTTP requests but do not parse or execute HTML page content, automated web browsers offer all the features a user would experience in their own web browser. Finally, *extension-based methods* lean closer to the ecological side, and involve the development of a custom browser add-on that must be implemented within the constraints of the browser(s) of interest. This extension must then be installed by a willing pool of participants, at which point it either interacts with an algorithm on their behalf, or monitors those interactions without intervention.

These three methods all offer the advantage of being fully automated, enabling rapid data collection and a high degree of researcher control. However, search engines have rate limits—limits on the number of requests they will accept from a given server before blocking it—that can complicate these approaches. While there are methods for evading these rate limits for request and automation-based methods, they come with ethical and legal considerations of which researchers should be aware. Among these, extension-based methods get the researcher the closest to measuring a users' interactions with an algorithm under ecological conditions.

4.5.2 Ecological Data Collection

To collect data on what real users do, the researcher must have access to a user's actual search behavior, thus requiring an extension-based method. Although this approach offers greater ecological validity because it involves real users conducting real searches, this validity comes at the expense of researcher control and comparability of the results. For example, it is unlikely that participants will conduct the exact

same query at the exact same time, which complicates between-subjects comparisons. Even more challenging is the recruitment of a representative sample of participants. Not everyone is willing to install a browser extension that monitors their search behavior, and this could skew the sample of enrolled users in ways that are hard to predict.

The involvement of real users and real search queries in the extension-based method also raises serious concerns around user safety. Researchers taking this approach should seek approval from their institution's Institutional Review Board (IRB); obtain informed consent from all participants; securely transmit and store collected data; scrub personally identifying information from the collected data to the greatest extent possible; and restrict access to the data to those who were IRB-approved. Note that raw or semi-raw data collected from a browser extension generally cannot ever be released publicly, even with IRB approval, because it is impossible to ensure that all personally identifiable information is redacted or anonymized. As past researchers has shown, participants may search for their own name, their home address, or any number of free text strings that are privacy sensitive, yet impossible to exhaustively enumerate for the purposes of data cleaning (McCullagh, 2006).

Recommendation 5: Regardless of whether the data is collected by actively instantiating algorithmic systems or observationally from user behavior, automated methods provide valuable efficiency and scalability. When reporting on the results of any search audit, researchers should provide technical details of the data collection process and consider open-sourcing their data collection pipeline and/or data to facilitate further study.

4.6 Measuring Personalization

As previously mentioned, algorithmic systems are unique in that they can tailor the user experience for individual user based on the existing data about the user or the world at large. When auditing algorithms, it is important that the researchers take into consideration the role that

such personalization have on the outputs of the algorithms. In some audit studies, personalization could serve as a confound that needs to be controlled, while in other studies, it may be the element that the researchers are trying to explicitly measure.

In search, for example, researchers have tried to measure personalization; such work has consistently found that personalization is not a major source of content variation (Hannak *et al.*, 2013; Robertson *et al.*, 2018b; Robertson *et al.*, 2018a; Le *et al.*, 2019), while other factors that vary by user but do not constitute user-level personalization, such as geographic localization, can have a substantial impact (Kliman-Silver *et al.*, 2015; Xing *et al.*, 2014). Meanwhile, in light of the often-limited role of personalization in search engines, other researchers have instead tried to mitigate the effect of personalization to produce results that might be more generalizable to the broader user population (Metaxa *et al.*, 2019; Trielli and Diakopoulos, 2019).

Strategies for handling these confounds differ depending on the auditing approach being used, but generally involve controls at the point of data collection. Depending on the research questions being asked and the data collection approach being used, one might wish to avoid personalization, or to explicitly measure it. We summarize previously used approaches for both directions in the context of auditing search engines. It is worth noting that despite our focus on search engines, we would expect the methods discussed here to generalize across other forms of algorithmic systems that are deployed on the web.

4.6.1 Avoiding Personalization

If the goal is to measure general trends in search engines, then personalization can hamper that goal, as past work has found some evidence of content and rank personalization in many search algorithms (e.g., Hannak *et al.*, 2013; Kliman-Silver *et al.*, 2015). While some search engines provide a setting for disabling personalization, experiments have shown that these settings may be ignored by the system in some cases (Ballatore, 2015). Factors influencing personalization must therefore be controlled in other ways. Within a request or automation-based data collection framework, there are two primary strategies for con-

trolling personalization: (1) hold as many personalization factors as possible constant, and (2) randomize personalization factors.

For example, to implement the former strategy, researchers have submitted queries from the same computer, in the same location, with no web history, and with the same browser and operating system fingerprint (these parameters can be specified in the headers of the request sent to the search engine) (Trielli and Diakopoulos, 2019; The Economist, 2019). Using the latter strategy, researchers have, for instance, used the Tor browser to randomize the origin of their query requests (Ballatore, 2015), or have rotated through a list of a dozen or more browser fingerprints as a weaker proxy for randomization (Metaxa *et al.*, 2019).

4.6.2 Identifying Personalization

If the goal is instead to measure the extent to which searches conducted from different locations or by different people return different results, then personalization is the variable of interest and needs to be isolated. To accomplish this isolation, researchers need to: (1) collect search rankings that have been personalized to different people or locations, (2) collect a control set of non-personalized search rankings from the same person/location, and (3) control for time, carry-over effects, and other sources of noise that can create differences in search rankings. In order to induce a specific kind of personalization, several strategies are available. For example, to study *location-based personalization*, researchers have held web identity constant while manipulating the GPS coordinates of their requests (Kliman-Silver *et al.*, 2015).Alternatively, other researchers have submitted queries to different country versions of Google by manipulating the web suffix (e.g. 'google.uk', 'google.de') and holding all else constant (Ballatore *et al.*, 2017).

In studies of *user-based personalization*, researchers have taken several different approaches, including manufacturing accounts with web histories that resemble certain group characteristics, such as age or political leaning, and then conducted searches from those accounts (Hannak *et al.*, 2013; Le *et al.*, 2019). Although this approach offers greater control, it has ecological validity issues because researchers must create web histories, and their choices in that respect may not accurately repre-

sent real users. Extension-based data collection strategies can overcome this issue by observing query behavior of recruited participants or even actively conducting searches from participant computers. By conducting searches on users' local machines, researchers have been able to measure personalization by running simultaneous pairs of queries in standard and incognito (a.k.a. private) browser windows, as results in the former may be personalized while the latter are not (Robertson *et al.*, 2018b; Robertson *et al.*, 2018a).

When studying personalization, it is important to control for additional sources of noise, including carry-over effects, updates to the search engine's index or algorithms, and A/B tests (Hannak *et al.*, 2013). Carry-over effects, where a recently conducted search affects the results returned in a subsequent search, were previously identified as a source of noise on Google Search (Hannak *et al.*, 2013), but a recent study did not find evidence that this was still occurring (Robertson *et al.*, 2018b). One way to address the remaining sources of noise is to conduct the same search multiple times while holding personalization factors constant; the results from those searches can then be used to establish a noise floor, above which differences can be attributed to personalization (Hannak *et al.*, 2013; Kliman-Silver *et al.*, 2015).

Recommendation 6: The role of personalization depends greatly on the particular algorithmic system in question, in some cases significantly changing system outputs and in others having very little effect. Researchers should be aware of possible sources of personalization and noise we have mentioned here and, when appropriate, design controls appropriate for their approach and research questions.

4.7 Interface Attributes

When analyzing any algorithms, researchers must make decisions about the level of detail to extract and analyze. When scraping or using APIs to collect data, we recommend casting a wide net when possible to allow for later analysis of the interplay between different interface attributes

or relevant metadata. For example, in the context of auditing search engines, much of prior work has focused on webpages of search results, termed *search engine result pages*, or SERPs. Important parts of the SERP include the search bar for user input (and the dynamic autocomplete suggestions that it may provide as the user alters their query); the main column of search results; and any additional information on the page, such as a side bar that might appear alongside the search column. In this section we focus on the main column of results, including the way its design has evolved over time, and how to make decisions about what to extract and measure.

While early SERPs consisted of a single column, where each row contained as little as a hyperlink or a short webpage summary, modern search results contains a variety of *components* that incorporate internal or cross-platform data (see Figure 4.2 for a visual example). In 2012, *extended components* began to appear on Google Search, which we define as components that incorporate internally curated data and data from external partnerships. Among these are "Knowledge boxes" that attempt to provide direct answers, lists of related questions with drop downs containing answers, and recent tweets from a relevant Twitter account. Lastly, there are also a number of *seasonal components* on Google Search, such as those that appear during sporting events or political elections (Diakopoulos *et al.*, 2018). These components come in a variety of designs, and their presence varies depending on the query searched and the device (e.g., from a desktop computer or a mobile phone) it was sent from (Robertson *et al.*, 2018b; Tober *et al.*, 2016). All these components vary in position on the SERP, with some appearing only near the top of the SERP, and in orientation, with some containing sub-results (e.g., different videos) that are arranged horizontally (Robertson *et al.*, 2018b).

Prior SERP analyses have varied from considering only the main list of links (Mowshowitz and Kawaguchi, 2005; Le *et al.*, 2019; Hannak *et al.*, 2013), to a focus on specific component types or their text summaries (McMahon *et al.*, 2017; Kay *et al.*, 2015; Hu *et al.*, 2019; Trielli and Diakopoulos, 2019), to an analysis of multiple component types and their positions (Vincent *et al.*, 2019; Robertson *et al.*, 2018b; Robertson *et al.*, 2018a; Kliman-Silver *et al.*, 2015; Vincent and Hecht, 2021).

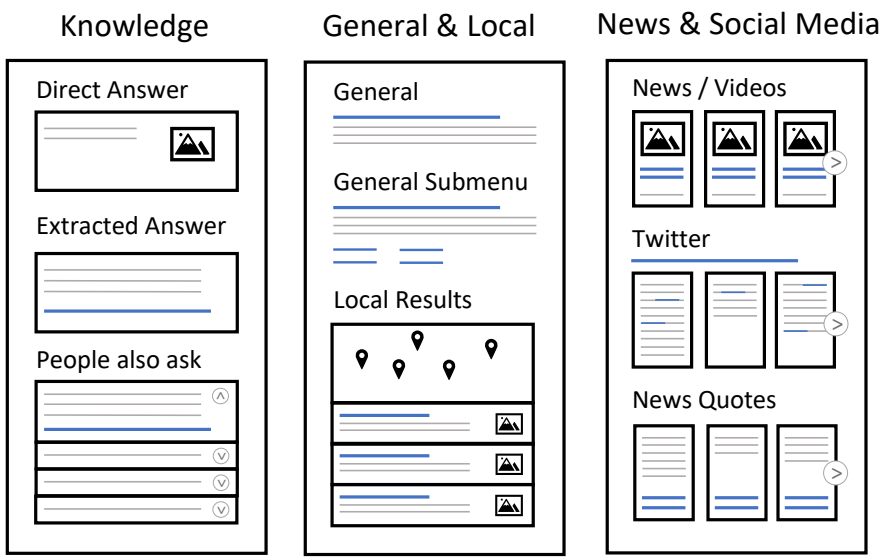

Figure 4.2: Examples highlighting the diversity of component types that can appear in modern search results.

Other research has shown that the design of a SERP affects what users pay attention to and how they evaluate it (Lurie and Mustafaraj, 2018; Granka, 2010). For example, the ranking and presence of component types can have a strong influence on a user's behavior and opinion formation (Epstein *et al.*, 2017). These findings emphasize the need to consider SERP composition in audits of search media, as the content presented in each component can vary systematically, altering the conclusions drawn (Robertson *et al.*, 2018a).

On a practical note, there are also technical factors to consider in the choice of page components to study, as the details of some components, such as those that require interaction (e.g., clicking on a Google SERP's "People also Ask" component), are harder to extract. If a researcher's question involves these components, they may need to use an automation-based method to collect data. Similarly, to account for biases in user attention, researchers commonly apply attention distributions or use rank-weighted metrics to more accurately quantify user exposure to links and content (Sapiezynski *et al.*, 2019; Robertson *et al.*, 2018a; Metaxa *et al.*, 2019).

Recommendation 7: When extracting data from an output page of an algorithmic system for analysis (e.g., links or text summaries on a search results page) the component type and its position should be recorded to provide a closer representation of the actual interface of the algorithm. These details can provide insights about what users paid attention to, and can be used to adjust estimates of user exposure and engagement with the results.

4.8 Analyzing Data

While analyses are very specific to each study, researchers should be explicit and conscientious about the choice of a baseline or comparison data set when drawing conclusions about algorithmic systems. For example, one influential audit of Google Image Search used queries for common occupations in order to compare the gender of people depicted therein to the representation of different genders in the workforce as a baseline (Kay *et al.*, 2015). In this penultimate section, we discuss some of the practical, high-level considerations for analysis. The guidelines laid out in this monograph so far have primarily covered methods for collecting algorithms' outputs for audit studies in ways that are ethical, and produce data that are representative of users' experiences. The process of analyzing data will vary greatly depending on the type of algorithm that is being audited, data that is collected, and questions of interest that are being raised. So here we focus on outlining three important areas that are commonly applicable to many analyses in the context of auditing search engines, but that can be generalized to other forms of algorithmic systems.

4.8.1 Data Filtering

When analyzing outputs of an algorithmic system, researchers may consider choosing a subset of the data to focus on in their analysis for the sake of a more narrowly-scoped question. For example, researchers studying politically-related search results might choose to focus on news

media content or content with a clear partisan leaning, as was done in a 2019 study on partisanship in Google search results by Metaxa *et al.* (2019). Data may also need filtering for the purpose of cleaning noise, as in another study that examined political partisanship on Google by Robertson *et al.* (2018a), who found that Wikipedia dominated the search results for many of the political search terms, possibly obscuring patterns that would be worth reporting. To this end, some of the analyses in the paper were replicated after removing the links to Wikipedia. In such cases, we recommend researchers clearly explain and justify what data was filtered and why, and clearly articulate the effect this had on analysis when describing the results.

4.8.2 Merging with External Data

Studies of algorithmic systems often seek to shed light on a phenomenon of social import (e.g., bias or misinformation), and as a result, it is often necessary to merge the outputs of these systems with data from external sources (e.g., to define the set of URLs considered misinformation when looking at the URLs returned by a search engine). The aforementioned prior work measuring political partisanship in web search, for example, employed existing partisanship scores established in prior work (Metaxa *et al.*, 2019; Robertson *et al.*, 2018a). We encourage researchers to draw upon existing data sets and research to enrich their analyses, but caution that doing so is not always straightforward, and the quality of the resulting analysis is only as robust as the weakest of the data sources used.

4.8.3 Choosing Baselines

When analyzing and drawing conclusions about an algorithmic system, researchers often need a well-justified baseline or point of comparison against which to interpret their data. The choice of comparison may be straightforward: consider, for example, a study auditing housing platforms like Zillow or Redfin. In the United States, the Fair Housing Act prohibits sellers of real estate from discriminating against citizens on the basis of the citizen's membership in a protected class such as race or religion (Asplund *et al.*, 2020a), so given this clear legal expectation,

such a study would have a compelling finding if results suggested that the search engines treated users differently by race.

In many cases, however, the choice of an appropriate baseline is challenging. For instance, for researchers studying news media in web search, it would be insufficient to solely report what percentage of the search results came from, say, left-leaning and right-leaning sources and assume that the relevant baseline would be equal representation of sources on both sides without controlling for other factors such as the number and quality of media outlets in each group.

Recommendation 8: The data analysis process is very study-specific, but at a high level researchers should pay close attention to their choices of baseline or comparison data sets, and clearly communicate any data filtering or cleaning that was done as part of the analysis.

4.9 Communicating Findings

Presenting findings, both to an audience of academic peers as well as the general public, is one of the final steps in research and one of the most important; it deserves attention from the outset of the research process. We believe that researchers have a responsibility to consider the wider public discourse surrounding our work and our potential to impact it—an argument we will advance further in Section 5. This is particularly true in audit studies as they may have immediate implications toward individual users and society as a whole, and has the potential to change society; as such they could even materialize into activism, an argument we expand upon in the final section. Towards this end, we suggest work be both peer-reviewed and also communicated to the public in a more accessible format.

When communicating work in writing, researchers can choose to produce one or more of several options including peer-reviewed publications, less formal white papers, news media editorials, self-published blog posts, and others. Each has a different purpose and is best-suited for communicating to a different audience. Peer-reviewed publication is normally

considered the standard for disseminating research, and one which we endorse. In high-stakes domains such as politics or health, researchers should also strongly consider the opportunity to make their findings more accessible (e.g., to the general public, journalists, and policy makers), as well as to avoid misinterpretation or misrepresentation, by also engaging explicitly in the public dialog around their topic of study and publishing accompanying, less formal pieces of writing. In addition to benefiting the wider public, researchers have demonstrated that public disclosure of algorithm audit results can prompt the specific companies audited to improve their algorithms (Raji and Buolamwini, 2019).

Being more concrete, notable and high-impact examples of research communication outside academia include the memorable ProPublica article by Angwin *et al.* (2016) that was very effective in bringing public attention to the algorithms used in the legal system to assign risk scores to defendants. Another such example is testimony by Buolamwini (2019) before United States Congress on the impact of facial recognition technologies on civil rights.

It is also imperative that any communication around findings is clear and forthcoming about the implications of the work and its shortcomings, as work of interest to the public is likely to be consumed by a wide audience not restricted to other academics well-versed in the nuances of the topic. While it is impossible for any researcher to fully foresee the future impacts of their work, we believe that researchers are, at least in part, responsible for the influence of our work in the world, and should therefore attempt to make our research carefully contextualized and interpretable by a wider audience, especially stakeholder communities.

Recommendation 9: We believe it is our responsibility as researchers to be cognizant of the wider public discourse surrounding our topic of study, to give careful consideration to the impact our work will have on this discourse, and to contextualize and communicate the implications and limitations of our work accordingly. More concretely, we recommend findings be peer-reviewed, and also encourage researchers to publish blogs or editorials for a more general audience.

As a team of researchers with experience conducting algorithm audits, we hope these guidelines will provide context and insight to other algorithm auditors. However, we hope that they also provide insights to those outside the academy and those indirectly involved with this method, including journalists and interested citizens, to better comprehend and evaluate algorithm audits. Finally, for ease of reference we have summarized the nine key dimensions and recommendations in Table 4.1.

Table 4.1: Summarizing the nine key dimensions and their best practices for algorithm audits.

Dimension	Recommended best practice
Legal and Ethical Considerations	Researchers should be aware of relevant laws; make informed decisions based on risk and personal ethics; and respect systems and their users.
Selecting a Research Topic	When choosing an area of study with potential for social impact, collaborate with domain experts from multiple disciplines to thoughtfully situate the work and its implications.
Choosing an Algorithm	Keep in mind the real-world influence of the system, for example by considering those with large user bases or widespread use among populations of interest.
Temporal Considerations	Consider collecting data at more than one time point, both for the purpose of comparative study and as a check on the robustness of the data.
Collecting Data	When possible, use automated methods both for scalability and replicability. Carefully document the data collection process and consider open-sourcing the data and/or data collection pipeline.
Measuring Personalization	Be aware of potential sources of personalization or other noise affecting the generalizability of the data, and design appropriate controls to account for (or explicitly measure) these when possible.
Interface Attributes	Consider collecting metadata about the presentation of elements on the page to allow for later analysis of the interplay between different interface attributes and to adjust analyses for hypothetical or actual user exposure.
Analyzing Data	While analyses are very specific to each study, researchers should being explicit and conscientious about the choice of a baseline or comparison data.
Communicating Findings	Consider the wider public discourse surrounding the work, since auditing can have significant potential to impact it. We suggest work be both peer-reviewed and also communicated to the public in other more accessible formats.

5

Audits as Activism

From the history of social science audits through many algorithm audits conducted in recent years, a key focus of auditing has always been the identification of discriminatory practices or outcomes in a wide range of socially important contexts.

While not all algorithm audits have social effects, and not all algorithm auditors are attuned to this dimension, in this section we advance a normative argument that much auditing work constitutes activism: it is a practice of direct action, often with the effect of drawing attention to an issue and bringing about political and social change.

5.1 Are Audits Activist?

In short, we argue, yes. We come to this conclusion by examining the real-world implications of such work, from which the work cannot be separated.

As has been argued by scholars in the related field of Science and Technology Studies, ownership over algorithmic tools and data, along with the ability to monitor and understand them, increasingly wields power in our society (Milan and Van Der Velden, 2016; Chun, 2011). Moreover, algorithm audits usually target sociotechnical systems: com-

putational artifacts that are socially situated, whose development and use are constantly *shaping and shaped by* the people who use them and the society in which they are used. Concrete examples of this abound, several of which we have mentioned in this monograph, such as changes to Google's ad delivery catalyzed by the discovery that searches for Black-sounding names suggested arrest records where similar searches for white-sounding names did not (Sweeney, 2013b). In another key example, researchers Buolamwini and Gebru (2018) identified that commercial facial recognition software under-performed on darker-skinned and women's faces, a finding which led to measurable changes in such systems (Raji and Buolamwini, 2019) and author Joy Buolamwini's invitation to testify before the U.S. Congress on the issue of facial recognition technology (Buolamwini, 2019).

There may well be exceptions—audits that do not bear on citizens or society in any significant way, audits identifying discrepancies so banal the findings do not have any meaningful effect on anyone's life, or audits garnering attention in the popular press nonetheless failing to translate to meaningful change. We suggest that cases of the former two categories are few and far between, with such audits likely conducted internally to organizations and are not in the domain of the independent researchers whom we envision as our primary audience in writing this piece. There are more examples of the latter, cases where audit results were dismissed, or led slowly to only partial changes—such as in 2004, when anti-Semitic search results appeared on Google, and the company's initial response was to tell users to use different search terms (Vaidhyanathan, 2011). In another example, research by Noble (2013) showed Google Search's racist and sexist bias in search results for the query "Black girls," but those findings took years to be addressed Noble, 2018b. In many cases, corporations' responses might more closely resemble band-aid solutions rather than prompt, deep, and meaningful change.

Still, despite the possibility of insufficient response by culpable companies, the outcomes of many audit studies have real-world consequences, helping give citizens, policy makers, consumers, and others evidence or impetus to create social change. While this closed loop between science and society applies to a plethora of epistemologies, algorithm auditors need to be cognizant of the possibility that their findings will have real

social effects, and conduct them accordingly, paying careful attention to their choice of topic, the rigor of their methods, and the communication of their findings.

This activist lens has very real and important implications for algorithm auditors. In addition to informing researchers' priorities in the choice of research topic or system of study, understanding that auditing work can precipitate social change is necessary for approaching such work with an eye towards the strategies used in this work. In one high-profile recent case, for instance, researchers at New York University were engaged in a project involving monitoring and auditing advertisements on Facebook. The NYU Ad Observatory was launched in September of 2020, and recruited volunteers to participate in a noninvasive user audit by installing an extension that would collect data about the political ads they saw on Facebook. In August of 2021, less than a year after its launch, Facebook disabled all of the Ad Observatory's platform access, citing "violation of our Terms of Service," a justification the project's researchers called a pretext (Bobrowsky, 2021). Within a week, this move had also drawn commentary from several United States Senators and other political actors. As the researchers in this study found, potential real-world social and policy implications can put algorithm auditors in direct conflict with platforms and other invested parties. Such consequences must be anticipated by auditors and may require them use strategies that go beyond traditional research methods, like engaging with political actors and other activists on behalf of their work.

The possibility for direct change precipitated by an algorithm audit presents great opportunities as well as risks, and we hope this framing will encourage researchers to consider the politically weighty and socially important aspect of the work at hand as deeply as any technical advice we can provide.

As algorithm auditors ourselves, we acknowledge that our skill set and expertise are often biased towards the technical; this is natural, given the computationally intensive and complex methods involved in conducting an algorithm audit. With this in mind, we implore collaboration with social scientists, policy experts and policymakers when identifying the need for an audit and conducting one.

5.2 The Importance of Impartiality

For technologists looking for a long-term vision, we see promise in moving algorithm auditing towards the goal of developing enduring infrastructure, such that neutral third-parties can more easily conduct and repeat audits. While some of this mission is being taken up by private ventures, such as Cathy O'Neil's consulting agency (ORCAA), we see value in this functionality falling under the umbrella of governmental agencies, not-for-profit companies, and other groups that will explicitly act in the public's best interest. Importantly, to facilitate this process and act upon their results, we also believe there is a need for savvy legislation and regulation to create real consequences and incentivize responsible, transparent auditing.

As evidence of this need, HireVue, a company selling software for analyzing job interview videos, has come under fire for the high likelihood of bias in its system. After the ORCAA consultancy conducted a recent audit of the product, HireVue announced in January 2021 that the audit found that its product was not biased—a framing journalists have called a mischaracterization (Engler, 2021). Unfortunately, both ORCAA's restrictions on the sharing of its report and the fully opt-in nature of such an audit make it likely that such audits will become widely conducted, or that their outcomes will not be widely and accurately circulated.

In another example, Wilson *et al.* (2021)—a team of academics collaborating with the company being audited—conducted a collaborative audit. Audits done in direct collaboration with industry, however, are not yet common practice, and some critics have pointed out the potential risks involved in close collaboration between auditors and companies (Sloane, 2021), and highlighted the value of research conducted financially (and otherwise) independently from the companies or services studied (Matias, 2020). Auditors must take care to consider the position from which they are auditing, including the composition of their team, their funding sources and disclosures, and other logistical aspects that may impede (or be seen to impede) the impartiality of their audit.

5.3 Future Frameworks for Auditing

Many possible frameworks have begun to be discussed for algorithm auditing. Some of these are inspired by existing infrastructure; governmental agencies' audit practices (in the United States, for instance, these include the Government Accountability Office and US Army Audit Agency) and other third-party auditors provide models for better algorithm auditing practices. One specific framework comes from the U.S. Government Accountability Office, developed after engaging with government, industry, and nonprofit experts (GAO), 2021). The framework advocates for four central pillars for accountability in artificial intelligence: data ("quality, reliability, and representativeness"), monitoring ("reliability and relevance over time"), governance (promoting accountability through organization- and system-level processes), and performance (measuring component- and system-level outcomes against program objectives) (GAO), 2021). Given the prominent role of the GAO in other forms of auditing, conducting similar oversight of algorithms could hold promise.

Two other related frameworks were recently proposed; first, Cobbe *et al.* (2021) introduce "reviewability" as a possibility for ensuring that automated decision-making involving machine learning technologies is more accountable, though here too, as critics have pointed, accountability as a standard has limitations in its ability to guarantee ethical algorithms, as we discussed in Section 4.1 (Keyes *et al.*, 2019). Similarly, Brown *et al.* (2021) suggest an audit instrument focused on stakeholder interests with the intent to more carefully measure the social context in which algorithms are deployed.

A final framework comes from academic researchers who have proposed using bottom-up user-driven processes to essentially crowdsource potential issues from users (Shen *et al.*, 2021). Unlike the previous frameworks proposed, which rely on experts—whether from academia, industry, or government—to identify issues, this framework proposes that everyday users could help identify problematic algorithmic content that experts might otherwise miss.

Many of these frameworks are very recent, a reflection that the future of algorithm auditing is currently and actively being developed.

Guided by such models, we will continue to encourage and work towards infrastructure and policy in support of third-party public interest algorithm auditing to ensure fairness and transparency in the algorithmic systems that impact us all.

6

Conclusion

In this work, we have sought to provide readers with the history and context to understand algorithm audits as they gain popularity and use in research. Beginning in the 1960s in the social sciences, auditing opaque processes in domains like housing and employment became a way to understand whether those services were equitable and serving protected classes of citizens fairly. The emergence of algorithmically-powered systems around the turn of the century created a new opportunity for this method to be expanded into the algorithm audit, with the same goal of understanding how (algorithmic) services were treating different categories of content and users. Drawing from our own experience, we have next outlined best practices in conducting algorithm audits seeking to cover the whole pipeline, from deciding to conduct an algorithm audit and selecting the domain through collecting and analyzing data, to eventually communicating findings. It is our goal that this information will help those new to algorithm auditing tackle the learning curve and begin conducting their own audits as smoothly and effectively as possible. Lastly, we concluded with Section 5 by discussing the stakes—the potential for algorithm audits to act as vehicles for meaningful social change, and the implications this has for those conducting them.

Algorithm audits are a promising area for researchers interested in a wide range of emerging sub-disciplines across many fields, including social justice informatics; human-centered artificial intelligence; fairness, accountability, transparency, and ethics in technology; and others. We hope to see these growing communities adopt and leverage the algorithm audit, and, paralleling the use of social science audits, use it as a tool for direct action and accountability towards more equitable and just technologies.

References

Abebe, R., S. Hill, J. W. Vaughan, P. M. Small, and H. A. Schwartz. (2019). "Using Search Queries to Understand Health Information Needs in Africa." *Proceedings of the International AAAI Conference on Web and Social Media*. 13(01): 3–14. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/3360.

Abid, A., M. Farooqi, and J. Zou. (2021). "Large language models associate Muslims with violence." *Nature Machine Intelligence*. 3(6): 461–463.

ACLU. (2020). "Federal Court Rules 'Big Data' Discrimination Studies Do Not Violate Federal Anti-Hacking Law." Mar. URL: https://www.aclu.org/press-releases/federal-court-rules-big-data-discrimination-studies-do-not-violate-federal-anti (accessed on 08/26/2020).

Allam, A., P. J. Schulz, and K. Nakamoto. (2014). "The Impact of Search Engine Selection and Sorting Criteria on Vaccination Beliefs and Attitudes: Two Experiments Manipulating Google Output." *Journal of Medical Internet Research*. 16(4): e100. DOI: 10.2196/jmir.2642.

Angwin, J., J. Larson, S. Mattu, and L. Kirchner. (2016). "Machine Bias." *ProPublica*. May. URL: https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing (accessed on 01/05/2021).

Angwin, J. and S. Mattu. (2016). "Amazon Says It Puts Customers First. But Its Pricing Algorithm Doesn't." *ProPublica*. Sept. URL: https://www.propublica.org/article/amazon-says-it-puts-customers-first-but-its-pricing-algorithm-doesnt (accessed on 01/05/2021).

Asplund, J., M. Eslami, H. Sundaram, C. Sandvig, and K. Karahalios. (2020a). "Auditing Race and Gender Discrimination in Online Housing Markets." In: *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2020)*. AAAI. 10.

Asplund, J., M. Eslami, H. Sundaram, C. Sandvig, and K. Karahalios. (2020b). "Auditing race and gender discrimination in online housing markets." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 24–35.

Ballatore, A. (2015). "Google Chemtrails: A Methodology to Analyze Topic Representation in Search Engine Results." *First Monday*. 20(7). DOI: 10.5210/fm.v20i7.5597.

Ballatore, A., M. Graham, and S. Sen. (2017). "Digital Hegemonies: The Localness of Search Engine Results." *Annals of the American Association of Geographers*. 107(5): 1194–1215. DOI: 10.1080/24694452.2017.1308240.

Bandy, J. (2021). "Problematic Machine Behavior: A Systematic Literature Review of Algorithm Audits." arXiv: 2102.04256 [cs.CY].

Bandy, J. and N. Diakopoulos. (2020). "Auditing News Curation Systems: A Case Study Examining Algorithmic and Editorial Logic in Apple News." *Proceedings of the International AAAI Conference on Web and Social Media*. 14(1): 36–47. URL: https://ojs.aaai.org/index.php/ICWSM/article/view/7277.

Barocas, S., M. Hardt, and A. Narayanan. (2017). "Fairness in machine learning." *Nips tutorial*. 1: 2017.

Bechmann, A. and K. L. Nielbo. (2018). "Are We Exposed to the Same "News" in the News Feed?" *Digital Journalism*. 6(8): 990–1002. DOI: 10.1080/21670811.2018.1510741. eprint: https://doi.org/10.1080/21670811.2018.1510741.

Bernstein, M. S., M. Levi, D. Magnus, B. Rajala, D. Satz, and C. Waeiss. (2021). "ESR: Ethics and Society Review of Artificial Intelligence Research." arXiv: 2106.11521 [cs.CY].

Bertrand, M. and S. Mullainathan. (2004). "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American economic review*. 94(4): 991–1013.

Bobrowsky, M. (2021). "Facebook Disables Access for NYU Research Into Political-Ad Targeting." *The Wall Street Journal*. Aug.

Brown, S., J. Davidovic, and A. Hasan. (2021). "The algorithm audit: Scoring the algorithms that score us." *Big Data & Society*. 8(1): 2053951720983865. DOI: 10.1177/2053951720983865. eprint: https://doi.org/10.1177/2053951720983865.

Buolamwini, J. (2019). "Hearing on Facial Recognition Technology (Part 1): Its Impact on our Civil Rights and Liberties." URL: https://www.congress.gov/116/meeting/house/109521/witnesses/HHRG-116-GO00-Wstate-BuolamwiniJ-20190522.pdf.

Buolamwini, J. and T. Gebru. (2018). "Gender shades: Intersectional accuracy disparities in commercial gender classification." In: *Conference on fairness, accountability and transparency*. PMLR. 77–91.

Cartright, M.-A., R. W. White, and E. Horvitz. (2011). "Intentions and Attention in Exploratory Health Search." In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. ACM. 65–74.

Chen, L., R. Ma, A. Hannák, and C. Wilson. (2018). "Investigating the Impact of Gender on Rank in Resume Search Engines." In: *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems - CHI '18*. Montreal QC, Canada: ACM Press. 1–14. DOI: 10.1145/3173574.3174225.

Chen, L., A. Mislove, and C. Wilson. (2015). "Peeking beneath the hood of uber." In: *Proceedings of the 2015 internet measurement conference*. 495–508.

Chowdhury, R. and J. Williams. (2021). "Introducing Twitter's first algorithmic bias bounty challenge." URL: https://blog.twitter.com/engineering/en_us/topics/insights/2021/algorithmic-bias-bounty-challenge.

Chun, W. H. K. (2011). *Programmed visions: Software and memory*. MIT Press.

Cobbe, J., M. S. A. Lee, and J. Singh. (2021). "Reviewable Automated Decision-Making: A Framework for Accountable Algorithmic Systems." In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. FAccT '21.* Virtual Event, Canada: Association for Computing Machinery. 598–609. DOI: 10.1145/3442188.3445921.

Coders' Rights Project. (2020). URL: https://www.eff.org/issues/coders (accessed on 10/27/2020).

Computing Machinery, A. for. (1992). "ACM Code of Ethics and Professional Conduct." *Code of Ethics.*

Correll, S. J., S. Benard, and I. Paik. (2007). "Getting a job: Is there a motherhood penalty?" *American journal of sociology.* 112(5): 1297–1338.

Crenshaw, K. (1990). "Mapping the margins: Intersectionality, identity politics, and violence against women of color." *Stan. L. Rev.* 43: 1241.

Daniel, W. W. (1968). *Racial discrimination in England: based on the PEP report.* Vol. 1084. Penguin.

Dave, K. (2019). "Systemic Algorithmic Harms." URL: https://points.datasociety.net/systemic-algorithmic-harms-e00f99e72c42 (accessed on 08/01/2021).

De Choudhury, M., M. R. Morris, and R. W. White. (2014). "Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media." In: *Proceedings of the 32nd annual ACM conference on Human factors in computing systems.* ACM. 1365–1376.

Diakopoulos, N., D. Trielli, J. Stark, and S. Mussenden. (2018). "I vote for—how search informs our choice of candidate." *Digital Dominance: The Power of Google, Amazon, Facebook, and Apple, M. Moore and D. Tambini (Eds.)* 22.

Díaz, M., I. Johnson, A. Lazar, A. M. Piper, and D. Gergle. (2018). "Addressing age-related bias in sentiment analysis." In: *Proceedings of the 2018 chi conference on human factors in computing systems.* 1–14.

DuckDuckGo. (2018). "Measuring the "Filter Bubble": How Google Is Influencing What You Click." URL: https://spreadprivacy.com/google-filter-bubble-study/.

Edelman, B., M. Luca, and D. Svirsky. (2017). "Racial discrimination in the sharing economy: Evidence from a field experiment." *American Economic Journal: Applied Economics*. 9(2): 1–22.

Edelman, B. G. and M. Luca. (2014). "Digital discrimination: The case of Airbnb.com." *Harvard Business School NOM Unit Working Paper*. (14-054).

Engler, A. C. (2021). "Independent auditors are struggling to hold AI companies accountable." *Fast Company*. Jan. URL: https://www.fastcompany.com/90597594/ai-algorithm-auditing-hirevue (accessed on 01/05/2021).

Epstein, R. and R. E. Robertson. (2015). "The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections." *Proceedings of the National Academy of Sciences*. 112(33): E4512–E4521. DOI: 10.1073/pnas.1419828112.

Epstein, R., R. E. Robertson, D. Lazer, and C. Wilson. (2017). "Suppressing the Search Engine Manipulation Effect (SEME)." *Proceedings of the ACM on Human-Computer Interaction*. 1(CSCW): 1–22. DOI: 10.1145/3134677.

Eriksson, M. C. and A. Johansson. (2017). "Tracking gendered streams." *Culture Unbound*. 9(2): 163–183.

Fortunato, S., A. Flammini, F. Menczer, and A. Vespignani. (2006). "Topical Interests and the Mitigation of Search Engine Bias." *Proceedings of the National Academy of Sciences*. 103(34): 12684–12689. DOI: 10.1073/pnas.0605525103.

Gaddis, S. M. (2018). *Audit studies: Behind the scenes with theory, method, and nuance*. Vol. 14. Springer.

GAO), U. G. A. O. ((2021). "Artificial Intelligence: An Accountability Framework for Federal Agencies and Other Entities." URL: https://www.gao.gov/products/gao-21-519sp.

Ginsberg, J., M. H. Mohebbi, R. S. Patel, L. Brammer, M. S. Smolinski, and L. Brilliant. (2009). "Detecting Influenza Epidemics Using Search Engine Query Data." *Nature*. 457(7232): 1012.

Gneezy, U. and J. List. (2004). "Are the disabled discriminated against in product markets? Evidence from field experiments." In: *American Economic Association Annual Meeting*.

Goldman, E. (2008). "Search engine bias and the demise of search engine utopianism." In: *Web Search*. Springer. 121–133.

Google Terms of Service. (2017). URL: https://policies.google.com/terms?hl=en-US (accessed on 08/22/2019).

"Google China." (2021). https://en.wikipedia.org/wiki/Google_China. (Accessed on 08/01/2021).

Granka, L. A. (2010). "The Politics of Search: A Decade Retrospective." *The Information Society*. 26(5): 364–374. DOI: 10.1080/01972243.2010.511560.

Gross, T. (2017). "A 'Forgotten History' Of How The U.S. Government Segregated America." *NPR*. May. URL: https://www.npr.org/2017/05/03/526655831/a-forgotten-history-of-how-the-u-s-government-segregated-america (accessed on 01/05/2021).

Hannak, A., P. Sapiezynski, A. Molavi Kakhki, B. Krishnamurthy, D. Lazer, A. Mislove, and C. Wilson. (2013). "Measuring Personalization of Web Search." In: *Proceedings of the 22nd International Conference on World Wide Web - WWW '13*. Rio de Janeiro, Brazil: ACM Press. 527–538. DOI: 10.1145/2488388.2488435.

Hannak, A., G. Soeller, D. Lazer, A. Mislove, and C. Wilson. (2014). "Measuring price discrimination and steering on e-commerce web sites." In: *Proceedings of the 2014 conference on internet measurement conference*. 305–318.

Hannak, A., C. Wagner, D. Garcia, A. Mislove, M. Strohmaier, and C. Wilson. (2017). "Bias in Online Freelance Marketplaces: Evidence from TaskRabbit and Fiverr." In: *20th ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW 2017)*. Portland, OR.

Hu, D., S. Jiang, R. E. Robertson, and C. Wilson. (2019). "Auditing the Partisanship of Google Search Snippets." In: *The World Wide Web Conference on - WWW '19*. San Francisco, CA, USA: ACM Press. 693–704. DOI: 10.1145/3308558.3313654.

Hussein, E., P. Juneja, and T. Mitra. (2020). "Measuring Misinformation in Video Search Platforms: An Audit Study on YouTube." *Proceedings of the ACM on Human-Computer Interaction*. 4(CSCW1): 1–27.

Hutchinson, B. and M. Mitchell. (2019). "50 Years of Test (Un)Fairness: Lessons for Machine Learning." In: *Proceedings of the Conference on Fairness, Accountability, and Transparency. FAT* '19*. Atlanta, GA, USA: Association for Computing Machinery. 49–58. DOI: 10.1145/3287560.3287600.

Introna, L. D. and H. Nissenbaum. (2000). "Shaping the Web: Why the Politics of Search Engines Matters." *The Information Society*. 16(3): 169–185. DOI: 10.1080/01972240050133634.

Joachims, T., L. Granka, B. Pan, H. Hembrooke, F. Radlinski, and G. Gay. (2007). "Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search." *ACM Transactions on Information Systems*. 25(2): 7–es. DOI: 10.1145/1229179.1229181.

Juneja, P. and T. Mitra. (2021). "Auditing E-Commerce Platforms for Algorithmically Curated Vaccine Misinformation." In: *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. CHI '21*. Yokohama, Japan: Association for Computing Machinery. DOI: 10.1145/3411764.3445250.

Karahalios, K. (2018). "Discrimination Audits: Challenges to Discrimination Studies." URL: http://social.cs.illinois.edu/presentations/law-data-summit/karahalios.pdf.

Kawakami, A., K. Umarova, and E. Mustafaraj. (2020). "The Media Coverage of the 2020 US Presidential Election Candidates through the Lens of Google's Top Stories." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 14. 868–877.

Kay, M., C. Matuszek, and S. 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*. Seoul, Republic of Korea: ACM Press. 3819–3828. DOI: 10.1145/2702123.2702520.

Keyes, O., J. Hutson, and M. Durbin. (2019). "A Mulching Proposal: Analysing and Improving an Algorithmic System for Turning the Elderly into High-Nutrient Slurry." In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems. CHI EA '19*. Glasgow, Scotland Uk: Association for Computing Machinery. 1–11. DOI: 10.1145/3290607.3310433.

Kliman-Silver, C., A. Hannak, D. Lazer, C. Wilson, and A. Mislove. (2015). "Location, Location, Location: The Impact of Geolocation on Web Search Personalization." In: *Proceedings of the 2015 ACM Conference on Internet Measurement Conference - IMC '15*. Tokyo, Japan: ACM Press. 121–127. DOI: 10.1145/2815675.2815714.

Koshiyama, A., E. Kazim, P. Treleaven, P. Rai, L. Szpruch, G. Pavey, G. Ahamat, F. Leutner, R. Goebel, A. Knight, *et al.* (2021). "Towards Algorithm Auditing: A Survey on Managing Legal, Ethical and Technological Risks of AI, ML and Associated Algorithms."

Kugelmass, H. (2016). ""Sorry, I'm Not Accepting New Patients" an audit study of access to mental health care." *Journal of Health and Social Behavior*. 57(2): 168–183.

Kulshrestha, J., M. Eslami, J. Messias, M. B. Zafar, S. Ghosh, K. P. Gummadi, and K. Karahalios. (2019). "Search Bias Quantification: Investigating Political Bias in Social Media and Web Search." *Information Retrieval Journal*. 22(1-2): 188–227. DOI: 10.1007/s10791-018-9341-2.

Lampos, V., A. C. Miller, S. Crossan, and C. Stefansen. (2015). "Advances in Nowcasting Influenza-like Illness Rates Using Search Query Logs." *Scientific reports*. 5: 12760.

Lawrence, S. and C. L. Giles. (1999). "Accessibility of Information on the Web." *Nature*. 400(6740): 107–107. DOI: 10.1038/21987.

Lazer, D., R. Kennedy, G. King, and A. Vespignani. (2014). "The Parable of Google Flu: Traps in Big Data Analysis." *Science*. 343(6176): 1203–1205.

Le, H., R. Maragh, B. Ekdale, A. High, T. Havens, and Z. Shafiq. (2019). "Measuring Political Personalization of Google News Search." In: *The World Wide Web Conference on - WWW '19*. San Francisco, CA, USA: ACM Press. 2957–2963. DOI: 10.1145/3308558.3313682.

Lurie, E. and E. Mustafaraj. (2018). "Investigating the Effects of Google's Search Engine Result Page in Evaluating the Credibility of Online News Sources." In: *Proceedings of the 10th ACM Conference on Web Science - WebSci '18*. Amsterdam, Netherlands: ACM Press. 107–116. DOI: 10.1145/3201064.3201095.

Makhortykh, M., A. Urman, and R. Ulloa. (2020). "How Search Engines Disseminate Information about COVID-19 and Why They Should Do Better." *Harvard Kennedy School Misinformation Review*. May. DOI: 10.37016/mr-2020-017.

Matias, J. N. (2020). "Why We Need Industry-Independent Research on Tech & Society." URL: https://citizensandtech.org/2020/01/industry-independent-research/.

McCown, F. and M. L. Nelson. (2007). "Agreeing to Disagree: Search Engines and Their Public Interfaces." In: *Proceedings of the 2007 Conference on Digital Libraries - JCDL '07*. Vancouver, BC, Canada: ACM Press. 309. DOI: 10.1145/1255175.1255237.

McCullagh, D. (2006). "AOL's Disturbing Glimpse into Users' Lives." CNET. URL: https://www.cnet.com/news/aols-disturbing-glimpse-into-users-lives/.

McMahon, C., I. Johnson, and B. 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 Eleventh International AAAI Conference on Web and Social Media (ICWSM 2018)*. AAAI. 10.

Mehrotra, R., A. Anderson, F. Diaz, A. Sharma, H. Wallach, and E. Yilmaz. (2017). "Auditing Search Engines for Differential Satisfaction Across Demographics." In: *Proceedings of the 26th International Conference on World Wide Web Companion - WWW '17 Companion*. Perth, Australia: ACM Press. 626–633. DOI: 10.1145/3041021.3054197.

Metaxa, D., J. S. Park, J. A. Landay, and J. Hancock. (2019). "Search Media and Elections: A Longitudinal Investigation of Political Search Results in the 2018 U.S. Elections." In: *Proceedings of the 22nd ACM Conference on Computer-Supported Cooperative Work and Social Computing*. ACM.

Metaxas, P. T. and E. Mustafaraj. (2009). "The Battle for the 2008 US Congressional Elections on the Web." In: *Proceedings of the 2009 WebScience: Society On-Line Conference*.

Metaxas, P. T. and Y. Pruksachatkun. (2017). "Manipulation of Search Engine Results during the 2016 US Congressional Elections." In: *Proceedings of the ICIW 2017*. 6.

Mikians, J., L. Gyarmati, V. Erramilli, and N. Laoutaris. (2012). "Detecting price and search discrimination on the internet." In: *Proceedings of the 11th ACM workshop on hot topics in networks*. 79–84.

Milan, S. and L. Van Der Velden. (2016). "The alternative epistemologies of data activism." *Digital Culture & Society*. 2(2): 57–74.

Milkman, K. L., M. Akinola, and D. Chugh. (2012). "Temporal Distance and Discrimination: An Audit Study in Academia." *Psychological Science*. 23(7): 710–717. DOI: 10.1177/0956797611434539. eprint: https://doi.org/10.1177/0956797611434539.

Mowshowitz, A. and A. Kawaguchi. (2002a). "Assessing Bias in Search Engines." *Information Processing & Management*. 38(1): 141–156. DOI: 10.1016/S0306-4573(01)00020-6.

Mowshowitz, A. and A. Kawaguchi. (2002b). "Bias on the Web." *Communications of the ACM*. 45(9). DOI: 10.1145/567498.567527.

Mowshowitz, A. and A. Kawaguchi. (2005). "Measuring Search Engine Bias." *Information Processing & Management*. 41(5): 1193–1205. DOI: 10.1016/j.ipm.2004.05.005.

Mullainathan, S., M. Noeth, and A. Schoar. (2012). "The market for financial advice: An audit study." *Tech. rep*. National Bureau of Economic Research.

Mustafaraj, E., E. Lurie, and C. Devine. (2020). "The Case for Voter-Centered Audits of Search Engines During Political Elections." In: *Conference on Fairness, Accountability, and Transparency (FAT* '20)*. 11.

Noble, S. U. (2013). "Google search: Hyper-visibility as a means of rendering black women and girls invisible." *InVisible Culture*. (19).

Noble, S. U. (2018a). *Algorithms of oppression*. New York University Press.

Noble, S. U. (2018b). "Google Has a Striking History of Bias Against Black Girls." *Time Magazine*.

Nosek, B. A. and T. M. Errington. (2020). "What is replication?" *PLoS biology*. 18(3): e3000691.

Obermeyer, Z., B. Powers, C. Vogeli, and S. Mullainathan. (2019). "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science*. 366(6464): 447–453. DOI: 10.1126/science.aax2342.

Oreopoulos, P. (2011). "Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes." *American Economic Journal: Economic Policy*. 3(4): 148–71.

Pan, B., H. Hembrooke, T. Joachims, L. Lorigo, G. Gay, and L. Granka. (2007). "In Google We Trust: Users' Decisions on Rank, Position, and Relevance." *Journal of Computer-Mediated Communication*. 12(3): 801–823. DOI: 10.1111/j.1083-6101.2007.00351.x.

Pitoura, E., P. Tsaparas, G. Flouris, I. Fundulaki, P. Papadakos, S. Abiteboul, and G. Weikum. (2018). "On Measuring Bias in Online Information." *ACM SIGMOD Record*. 46(4): 16–21. DOI: 10.1145/3186549.3186553.

President, E. O. of the. (2016). "Big Data: A Report on Algorithmic Systems, Opportunity, and Civil Rights." URL: https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/2016_0504_data_discrimination.pdf.

Prosperi, M. and J. Bian. (2019). "Is it time to rethink institutional review boards for the era of big data?" *Nature Machine Intelligence*. 1(6): 260–260.

Purcell, K. and J. Brenner. (2012). "Search Engine Use 2012." URL: http://www.pewinternet.org/2012/03/09/search-engine-use-2012/.

Raji, I. D. and J. Buolamwini. (2019). "Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products." In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 429–435.

Robertson, A. (2019). "Scraping public data from a website probably isn't hacking, says court." The Verge. URL: https://www.theverge.com/2019/9/10/20859399/linkedin-hiq-data-scraping-cfaa-lawsuit-ninth-circuit-ruling.

Robertson, R. E., S. Jiang, K. Joseph, L. Friedland, D. Lazer, and C. Wilson. (2018a). "Auditing Partisan Audience Bias within Google Search." *Proceedings of the ACM on Human-Computer Interaction*. 2(CSCW): 1–22. DOI: 10.1145/3274417.

Robertson, R. E., S. Jiang, D. Lazer, and C. Wilson. (2019). "Auditing Autocomplete: Suggestion Networks and Recursive Algorithm Interrogation." In: *Proceedings of the 10th ACM Conference on Web Science - WebSci '19*. Boston, Massachusetts, USA: ACM Press. 235–244. DOI: 10.1145/3292522.3326047.

Robertson, R. E., D. Lazer, and C. Wilson. (2018b). "Auditing the Personalization and Composition of Politically-Related Search Engine Results Pages." In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web - WWW '18*. Lyon, France: ACM Press. 955–965. DOI: 10.1145/3178876.3186143.

Rothshild, A., E. Lurie, and E. Mustafaraj. (2019). "How the Interplay of Google and Wikipedia Affects Perceptions of Online News Sources." In: *Computation+ Journalism Symposium*.

Sandvig, C., K. Hamilton, K. Karahalios, and C. Langbort. (2014). "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." *Data and discrimination: converting critical concerns into productive inquiry*. 22.

Sapiezynski, P., W. Zeng, R. E. Robertson, A. Mislove, and C. Wilson. (2019). "Quantifying the Impact of User Attention on Fair Group Representation in Ranked Lists." In: *Companion Proceedings of The 2019 World Wide Web Conference on - WWW '19*. San Francisco, USA: ACM Press. 553–562. DOI: 10.1145/3308560.3317595.

Sharma, R., A. Mitra, and M. Stano. (2015). "Insurance, race/ethnicity, and sex in the search for a new physician." *Economics Letters*. 137: 150–153.

Shen, H., A. DeVos, M. Eslami, and K. Holstein. (2021). "Everyday algorithm auditing: Understanding the power of everyday users in surfacing harmful algorithmic behaviors." *arXiv preprint arXiv:2105.02980*.

Siegelman, P. and J. Heckman. (1993). "The Urban Institute audit studies: Their methods and findings." *Clear and Convincing Evidence: Measurement of Discrimination in America, Washington*. 187: 258.

Sloane, M. (2021). "The Algorithmic Auditing Trap." URL: https://onezero.medium.com/the-algorithmic-auditing-trap-9a6f2d4d461d.

Speicher, T., M. Ali, G. Venkatadri, F. Ribeiro, G. Arvanitakis, F. Benevenuto, K. Gummadi, P. Loiseau, and A. Mislove. (2018). "Potential for discrimination in online targeted advertising." In: *Conference on Fairness, Accountability, and Transparency (FAT '20.*

"Search Engine Market Share China." (2019). URL: https://gs.statcounter.com/search-engine-market-share/all/china (accessed on 08/22/2019).

"Search Engine Market Share Worldwide." (2019). URL: https://gs.statcounter.com/search-engine-market-share (accessed on 08/22/2019).

Sweeney, L. (2013a). "Discrimination in Online Ad Delivery." URL: https://dataprivacylab.org/projects/onlineads/index.html (accessed on 10/26/2020).

Sweeney, L. (2013b). "Discrimination in Online Ad Delivery." *Queue.* 11(3): 10:10–10:29. DOI: 10.1145/2460276.2460278.

The Economist. (2019). "Google's Algorithm – Seek and You Shall Find." *The Economist.* June: 81 (US).

Thomas, L. (2021). "Supreme Court ruling that limits hacking law supports U-M researcher."

Tober, M., J. Grundmann, and A. Thakur. (2016). "Universal & Extended Search 2016: Facts, Trends and Optimization Tips." *Tech. rep.* searchmetrics.

Trielli, D. and N. 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 - CHI '19.* Glasgow, Scotland Uk: ACM Press. 1–15. DOI: 10.1145/3290605.3300683.

Trielli, D. and N. Diakopoulos. (2020). "Partisan Search Behavior and Google Results in the 2018 U.S. Midterm Elections." *Information, Communication & Society.* May: 1–17. DOI: 10.1080/1369118X.2020.1764605.

Tripodi, F. (2018). "Searching for Alternative Facts." *Tech. rep.* Data&Society. 64.

"U.S. Government Accountability Office (U.S. GAO)." URL: https://www.gao.gov (accessed on 08/01/2021).

Union, A. C. L. (2019). "Sandvig v. Barr – Challenge to CFAA Prohibition on Uncovering Racial Discrimination Online." May. URL: https://www.aclu.org/cases/sandvig-v-barr-challenge-cfaa-prohibition-uncovering-racial-discrimination-online (accessed on 01/05/2021).

Vaidhyanathan, S. (2011). *The Googlization of everything*. University of California Press.

Vincent, N. and B. Hecht. (2021). "A Deeper Investigation of the Importance of Wikipedia Links to Search Engine Results." *Proc. ACM Hum.-Comput. Interact.* 5(CSCW1). DOI: 10.1145/3449078.

Vincent, N., I. Johnson, P. Sheehan, and B. Hecht. (2019). "Measuring the Importance of User-Generated Content to Search Engines." In: *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 13.

White, R. W. and E. Horvitz. (2009). "Cyberchondria: Studies of the Escalation of Medical Concerns in Web Search." *ACM Transactions on Information Systems (TOIS)*. 27(4): 23.

Wilson, C., A. Ghosh, S. Jiang, A. Mislove, L. Baker, J. Szary, K. Trindel, and F. Polli. (2021). "Building and Auditing Fair Algorithms: A Case Study in Candidate Screening." In: *Conference on Fairness, Accountability, and Transparency (FAccT '21)*. 12.

Xing, X., W. Meng, D. Doozan, N. Feamster, W. Lee, and A. C. Snoeren. (2014). "Exposing Inconsistent Web Search Results with Bobble." In: *Passive and Active Measurement*. Ed. by M. Faloutsos and A. Kuzmanovic. *Lecture Notes in Computer Science*. Cham: Springer International Publishing. 131–140. DOI: 10.1007/978-3-319-04918-2_13.

Yinger, J. (1998). "Evidence on discrimination in consumer markets." *Journal of Economic perspectives*. 12(2): 23–40.

Zetter, K. (2010). "Wiseguys Plead Guilty in Ticketmaster Captcha Case." Wired. URL: https://www.wired.com/2010/11/wiseguys-plead-guilty/.

Zetter, K. (2015). "Experts Say Myspace Suicide Indictment Sets 'Scary' Legal Precedent." Wired. URL: https://www.wired.com/2008/05/myspace-indictm/.

Zetter, K. (2016). “Researchers Sue the Government Over Computer Hacking Law.” Wired. URL: https://www.wired.com/2016/06/researchers-sue-government-computer-hacking-law/.