Stanford University
Human-Centered
Artificial Intelligence

POLICY BRIEF
OCTOBER 2022

Using Algorithm Audits to Understand AI

Danaë Metaxa, Jeff Hancock

ARTIFICIAL INTELLIGENCE CONTINUES TO PROLIFERATE, FROM GOVERNMENT SERVICES AND ACADEMIC RESEARCH TO THE TRANSPORTATION, ENERGY, AND HEALTHCARE SECTORS. Yet one of the greatest challenges in using, understanding, and regulating AI persists: the black-box nature of many algorithms.

Dr. Latanya Sweeney's 2013 paper, "Discrimination in Online Ad Delivery," speaks to this very point. Sweeney, a professor at Harvard, surveyed 2,184 racially associated names in relation to searches tied to Google AdSense, Google's service for placing ads at the top of users' search results pages. All told, she found that ads placed on the page were far more likely to suggest an arrest record under queries for Black-sounding names than white-sounding ones—"raising questions as to whether Google's advertising technology exposes racial bias in society and how ad and search technology can help develop to assure racial fairness."

This question of racist or otherwise discriminatory AI is not just a widespread problem—as much other research has uncovered—it is also an issue of black-box decision-making. With respect to Sweeney's findings, one possibility is that Google deliberately targeted minority-sounding names with racist suggestions for "arrest records." It is also possible, however, that internet users were more likely to search Black names and then click on websites mentioning arrest. The harms and the dangers of this algorithmic discrimination are clear, but understanding

KEY TAKEAWAYS

- We identified nine considerations for algorithm auditing, including legal and ethical risks, factors of discrimination and bias, and conducting audits continuously so as to not capture just one moment in time.
- We found that researchers are activists—working on topics with social and political impacts, and behaving as actors with sociopolitical effects—and must factor the social impact of algorithmic development into their work.
- Algorithm auditors must collaborate with other experts and stakeholders, including social scientists, lawyers, ethicists, and the users of algorithmic systems to more comprehensively and ethically understand the impacts of those systems on individuals and society at large.

an algorithm's decision-making process can be far more difficult. Doing so matters greatly for researchers, policymakers, and the public.

In our paper, titled "Auditing Algorithms: Understanding Algorithmic Systems from the Outside In," we examine how algorithm audits—like the input- and output-testing Sweeney did for her research—are a powerful technique for understanding AI. In collaboration with researchers from Northeastern University, University of Illinois at Urbana-Champaign, and University of Michigan, we provide an overview of methodologies for algorithm audits, recount two decades of algorithm audits across numerous domains (from health to politics), and propose a set of best practices for conducting algorithm audits. We conclude with a discussion of algorithm audits and their social, ethical, and political implications.

The majority of algorithmic tests are pass/fail; they produce binary conclusions about an algorithm's operation. Audits of algorithms are more concerned with understanding a system in aggregate over time, even if they use tests along the way.

Introduction

Artificial intelligence applications are frequently used without any mechanism for external testing or evaluation. Simultaneously, many AI systems present black-box decision-making challenges. Modern machine learning systems are opaque to outside stakeholders, including researchers, who can only probe the system by providing inputs and measuring outputs. Researchers, users, and regulators alike are thus forced to grapple with using, being impacted by, or regulating algorithms they cannot fully observe.

Our paper reviews the history of algorithm auditing, describes its current state, and offers best practices for conducting algorithm audits today. Going beyond computer science, the concept of an audit refers to methodologically running randomized, controlled experiments in the field to evaluate a particular claim or requirement of the system. The U.S. Government Accountability Office conducts audits of government agencies and departments; independent consulting firms conduct audits of companies for tax liability, internal cybersecurity, and other risk- and compliance-related purposes. One of the most famous audit studies, conducted in 2004, highlights their potential impact: When Marianne Bertrand and Sendhil Mullainathan systematically developed and submitted fictitious résumés for help-wanted advertisements, they found that white names received 50% more interview callbacks—and that "differential treatment by race still appears to be prominent in the U.S. labor market."

Algorithm audits are different from traditional algorithmic testing. The majority of algorithmic tests are pass/fail; they produce binary conclusions about an algorithm's operation. Audits of algorithms are more concerned with understanding a system in aggregate over time, even if they use tests along the way. In general, audits are differentiated from other kinds of studies because the purpose is to understand the system itself (rather than a user's response to that system) and the auditor is generally positioned externally (or at least without insider knowledge into the system being probed).

We define algorithm audits as methods of repeatedly and systematically querying an algorithm with inputs and observing the corresponding outputs to draw inferences about its opaque inner workings. Researchers and investigators have successfully used audit studies for over half a century to expose implicit biases and discrimination across society. Audits are not always easy to conduct, and they do not always yield discrete conclusions; for instance, studying race-based discrimination through fictitious résumés raises the possibility that reviewers are drawing class-based conclusions from the names as well. In our paper, though, we find that algorithm audits hold great promise for better understanding the AI systems impacting our lives and the world around us.

Research Outcomes

In this work, we focus on search engines as an exemplar type of system, because they are prevalent and highly studied, with great power to shape people's behavior. In the service of new algorithm auditors, or anyone seeking to evaluate or understand the results of an audit, we identify nine key dimensions of algorithm audits: legal and ethical considerations; selecting a research topic; choosing an algorithm; temporal considerations; collecting data; measuring personalization; interface attributes; analyzing data; and communicating findings.

Algorithm audits' *legal and ethical considerations* include relevant laws, the terms of service of different platforms, users involved with or implicated by audits, and personal and institutional (e.g., university) ethical views and processes. For example, algorithm audits come with human attention costs (e.g., to do research, to interview individuals involved), computational costs, financial and monetary costs, and environmental costs (e.g., to run computers), among others, and these must be considered in the ethical calculus around conducting an audit. *Selecting a research topic* for an algorithm audit can include weighing discrimination and bias issues and political considerations (e.g., political polarization, a technology's political impacts). After that, *selecting the specific algorithm* to audit includes factoring in international considerations (e.g., which algorithms are popular where) and comparative factors (e.g., auditing one versus multiple algorithms and then comparing them).

We find that auditors must factor in *temporal considerations*—such as how often the algorithm is updated and how the data might change before, during, and after an audit is conducted. The point is not to conduct a single audit whose findings stand forever but to recognize that algorithm audits are meant to be continuous. In *collecting data*, researchers must consider the possible available data sources (e.g., APIs, manual collection, getting data sets directly from companies over email) and how analyzing the

data might scale. Auditors also need to *consider how personalization might change algorithms* from person to person and how that might impact audits; options include avoiding personalization in measurement, introducing experimental controls, and identifying personalization and making that a lens of analysis. Then, on the last three considerations: For *interfaces*, auditors must examine the relationship between interfaces and metadata (e.g., how searches are displayed on a webpage); for *analyzing data*, auditors must consider filtering the data, merging it with external data, and choosing points of comparison; and for *communicating findings*, auditors must consider the wider public discourse concerning the algorithms.

In terms of impact, we find that audits often have activist implications—and auditors should view their role through that lens. Science and technology scholars, as well as those from other fields, have increasingly underscored how algorithmic tools and data, as well as control over their design, use, and implementation, are a kind of societal power. Algorithms can also be products and amplifiers of societal power structures, like facial recognition that does not accurately recognize people of color. Algorithm audits are a way of interrogating and understanding those power dynamics and implications, and measured by their impact on the world, not all audits are created equal. Researchers should therefore give careful consideration to setting research priorities—for example, by choosing a topic or system to audit according to its potential for social impact.

Beyond the choice of *subject*, we also find that activism should inform researchers' *strategies*. Almost inevitably, algorithm auditors' findings can result in tension with some other party, whether a government agency or multinational corporation. Auditors must

In terms of impact, we find that audits often have activist implications—and auditors should view their role through that lens.

anticipate these scenarios—and may have to employ nontraditional research methods, like engaging with political actors and other activists, as part of their algorithm auditing process. The experiences of Dr. Safiya Noble, whose research on racist and sexist Google Search results put her in conflict with that company, and of researchers at New York University, whose Facebook Ad Observatory project led the company to cut off their data access and eventually to the involvement of U.S. senators, are but two examples of this reality. This potential for social impact makes interdisciplinary collaboration with social scientists and others outside of technical fields essential—their expertise is necessary for anticipating and addressing the social and political context surrounding an audit.

Policy Discussion

Algorithm audits are attracting more attention in Washington, D.C.—from congressional bills to conversations at the White House and the National Institute of Standards and Technology (NIST). The more

algorithms pervade our lives, the more policymakers will be forced to grapple with the implications of the fairness and black-box nature of their decision-making. Indeed, policymakers already have to confront this reality. Further, the more that researchers, outside auditing firms, and other stakeholders want to understand the field of algorithm auditing, the more important it will be for policymakers to monitor and help shape those developments.

Many of our recommendations are made to researchers and are relevant to policy discussions of algorithm audits:

- Researchers should be aware of relevant laws, their comfort with legal risk, and their own ethics, plus the impact of their audits on algorithmic services and their users.
- They should also focus on audit areas with the potential for positive social impacts, involving and collaborating with other domain experts and stakeholders across the social sciences, law, policy, and users.
- Auditors should also consider collecting data multiple times, which can provide unique and unexpected insights into algorithmic behavior and other external factors that might shape it.

These are all places where policymakers can be part of the discussion about algorithm audit best practices.

Further, we recommend that researchers clearly communicate their work to the public and to policy audiences. This includes providing technical details of their data collection process, open-sourcing their processes and/or data when appropriate (depending on privacy and security considerations), and proactively engaging in public discourse (by writing op-eds, discussing with journalists, writing blog posts, etc.) about their findings. We believe researchers have a responsibility to be aware of the wider public discussion around algorithms, audits, and societal impacts—and in kind, to contextualize and communicate their work with that in mind. Algorithm audits matter for policymaking, especially as government organizations, healthcare providers, law enforcement agencies, insurance firms, retail stores, and countless others use algorithms that shape public life.

We believe researchers have a responsibility to be aware of the wider public discussion around algorithms, audits, and societal impacts—and in kind, to contextualize and communicate their work with that in mind.

The original article is accessible at Danaë Metaxa et al., "**Auditing Algorithms: Understanding Algorithmic Systems from the Outside In**," *Now Foundations and Trends* (2021), https://ieeexplore.ieee.org/document/9627858.

Danaë Metaxa is the Raj and Neera Singh Term Assistant Professor of Computer and Information Science at University of Pennsylvania.

Jeff Hancock is the Harry and Norman Chandler Professor of Communication and the founding director of the Stanford Social Media Lab at Stanford University.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.

Stanford University
Human-Centered
Artificial Intelligence

Stanford HAI: Gates Computer Science Building, 353 Jane Stanford Way, Stanford University, Stanford, CA 94305

T 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**