From Transparency to Accountability and Back

From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing

Sarah H. Cen
Stanford University
United States of America
shcen@stanford.edu

Rohan Alur
Massachusetts Institute of Technology
United States of America
ralur@mit.edu

Abstract

Artificial intelligence (AI) is increasingly intervening in our lives, raising widespread concern about its unintended and undeclared side effects. These developments have brought attention to the problem of *AI auditing*: the systematic evaluation and analysis of an AI system, its development, and its behavior relative to a set of predetermined criteria. Auditing can take many forms, including pre-deployment risk assessments, ongoing monitoring, and compliance testing. It plays a critical role in providing assurances to various AI stakeholders, from developers to end users. Audits may, for instance, be used to verify that an algorithm complies with the law, is consistent with industry standards, and meets the developer's claimed specifications. However, AI developers and companies will rarely grant auditors unfettered access to their systems.

In this work, we examine a key consideration in AI auditing: what type of access to an AI system is needed to perform a meaningful audit? Addressing this question has direct policy relevance, as it can inform AI audit guidelines and requirements. We begin by discussing the factors that auditors balance when determining the appropriate type of access, and unpack the benefits and drawbacks of four types of access. We conclude that, at *minimum*, black-box access—providing query access to a model without exposing its internal implementation—should be granted to auditors. In particular, we argue that black-box access effectively balances concerns related to proprietary technology, data privacy, audit standardization, and audit efficiency. We then suggest a framework for determining how much *further* access (on top of black-box access) to provide to auditors. We show that auditing can be cast as a natural hypothesis test and argue that this framing provides clear and interpretable guidance on the implementation of AI audits. In particular, we draw parallels between aspects of hypothesis testing and those of legal procedure, such as legal presumption and burden of proof. As a result, hypothesis testing provides an approach to AI auditing that is both interpretable and effective, offering a potential path forward despite the challenges posed by AI's opacity.

EAAMO '24, October 29–31, 2024, San Luis Potosi, Mexico

ACM ISBN 979-8-4007-1222-7/24/10
https://doi.org/10.1145/3689904.3694711

CCS Concepts

• **Computing methodologies** → **Artificial intelligence**; **Machine learning**.

Keywords

AI auditing, monitoring and evaluation, AI policy, black-box auditing, hypothesis testing

ACM Reference Format:
Sarah H. Cen and Rohan Alur. 2024. From Transparency to Accountability and Back: A Discussion of Access and Evidence in AI Auditing. In *Equity and Access in Algorithms, Mechanisms, and Optimization (EAAMO '24), October 29–31, 2024, San Luis Potosi, Mexico.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3689904.3694711

1 Introduction

Auditing is the systematic evaluation of a system, often to determine whether it satisfies a predetermined set of criteria. With the proliferation of artificial intelligence (AI), auditing will serve as a vital tool for AI oversight and accountability. For example, without the ability to systematically test and assess—or *audit*—for compliance, AI regulations are impossible to enforce. Beyond compliance testing, auditing also plays several important roles. Perhaps most fundamentally, it allows for the independent evaluation of developer claims that would otherwise go unverified. It can also be used to certify whether an AI technology meets industry standards (e.g., privacy standards) that matter to downstream users (e.g., customers) even when they are not legally required. In this way, auditing not only plays an important role in AI accountability, but also takes an important step toward developing trustworthy AI.

Consider, by analogy, the U.S. car industry, in which auditing has three important functions. In the U.S., vehicles must adhere to a variety of federal standards and regulations related to safety and emissions, which are enforced through audits conducted by the National Highway Traffic Safety Administration (NHTSA). Beyond compliance testing, car manufacturers are required to disclose information about their vehicles, such as their fuel economy (i.e., mileage per gallon), which are both internally verified and subject to external audits by the Environmental Protection Agency (EPA). To gain an edge over competitors, car manufacturers also make claims about their vehicles; external and third-party audits validate these claims, legitimizing them and establishing trust between consumers and manufacturers.

There is a growing consensus that the AI industry would benefit from similar auditing mechanisms [9, 107, 109, 130]. For instance, in the European Union (EU), the AI Act mandates a mix of internal and third-party audits in the form of "conformity assessments" before the release of an AI system or after a substantial modification [118];

the General Data Protection Regulation (GDPR) calls for internal audits in the form of impact assessments conducted before data processing [14]; and the Digital Services Act (DSA) requires annual internal and external audits of risks [124]. Moreover, to ensure compliance, regulatory bodies (e.g., under GDPR, data protection authorities in each member state) are granted broad authority to conduct compliance tests. There are even provisions (e.g., in the DSA) that grant researchers special access to data and systems so that they can audit for criteria that the regulatory bodies may not consider [124].

Creating a healthy AI auditing ecosystem includes various considerations, such as who conducts the audits, who audits the auditors, what auditors test for, whether audits are prospective or retrospective, how often audits are conducted, and more. For many of these considerations, we may be able to look to other industries in which auditing practices are well established for inspiration. However, one question is particularly salient for AI auditing:

What type of access and how much access is needed to conduct a meaningful AI audit?

All auditing procedures require some form of access to the AI system, but the particular form of this access must balance (at times, competing) interests: (1) the protection of intellectual property and proprietary information; (2) security and privacy concerns; and (3) resource constraints. First, the protection of proprietary technologies and data (in particular, of trade secrets) is a key concern for companies, and access granted to auditors is therefore typically limited and carefully controlled. For example, EU regulatory bodies must adhere to strict principles of "necessity" and "proportionality" when handling personal data that is protected by GDPR [14, 77]. Second, requiring that AI companies open up their technology and datasets can create security and privacy risks. Third, auditors operate with limited resources (e.g., in labor, budget, and technical expertise) and are therefore interested in the amount of access that allows them to efficiently conduct effective audits. Motivated by this question, this work makes the following contributions:

- **Landscape of AI audits.** In Section 2, we survey contexts in which AI audits arise, including recent regulatory requirements, and discuss related work. We then examine the various (at times, competing) interests that influence the design and execution of AI audits, motivating our attention to audit *access*. In Section 3, we ground our discussion of audit access and implementation through a case study on New York City Local Law 144.
- **Audit access.** Auditing AI necessarily requires granting the auditor some form of *access* to the underlying system. In Section 4, we discuss the relative merits and limitations of different types of access to an AI system. While the appropriate form of access is context-dependent, we conclude that black-box access—the ability to observe a system's behavior on queries of the auditor's choice without examining the system's inner workings—provides the greatest amount of flexibility with the least amount of ambiguity.[1] We consequently argue that, at minimum, black-box should be granted to auditors. In particular, evidence obtained through black-box access can be definitively attributed to the AI system whereas evidence obtained through other forms of access is less conclusive. Black-box access offers several additional benefits: (1) it does not require direct access to proprietary algorithms or data, i.e., it does not "open the box;" (2) it is agnostic to the underlying AI mechanisms, meaning that the audit does not need to be updated even if the underlying training pipeline changes, i.e., a *standardized* audit can be used across different algorithms; (3) it is less resource-intensive than alternate auditing options, allowing continual, comprehensive, and scalable auditing; and (4) it can be run prospectively.[2]
- **Connections between auditing and hypothesis testing.** While we argue that black-box access is minimally necessary to conduct meaningful audits, it is not always sufficient. In Section 5, we discuss how auditing can be operationalized using hypothesis testing—a well studied statistical framework—and how this perspective provides clarity on two key auditing challenges. First, casting an audit as a hypothesis test provides clear guidance on how much access (*in addition* to black-box access) is needed to obtain the evidence required for a meaningful audit. That is, once the parameters of the hypothesis test are set, determining what evidence is needed for a meaningful test (i.e., audit) becomes a statistical exercise. Second, hypothesis testing has clear parallels to legal procedure and thus provides a way to map between complex statistical auditing methods and the law. In particular, we discuss how the null hypothesis can be viewed as legal presumption, placing the burden of proof on the party that wishes to falsify this presumption. Moreover, the "threshold" that an auditor selects maps to the "false positive rate" or "false negative rate" in hypothesis testing.

Remark. Black-box access is not adequate for every context. As we discuss in Section 4, black-box audits have their blind spot (e.g., they cannot speak to the intentions of AI developers or their data hygiene), and we argue that it is often necessary to complement black-box access with additional access, though full white-box access is often unnecessary.

2 Background and terminology

In this section, we briefly review auditing and its relation to AI. Of particular note is Table 1, which summarizes recent AI audit requirements. We note that AI auditing is a broad topic spanning computer science, law, and the social sciences, and we provide a more extensive discussion of related work in Appendices A and B.

2.1 A brief introduction to auditing

An audit is a systematic assessment of an organization, system, and/or process. Audits are conducted for many reasons, including (i) testing for *compliance with regulations*, (ii) determining whether a technology meets *certification standards*, (iii) validating *claims made by system designers*, (iv) monitoring an organization's *internal practices*, and (v) uncovering *vulnerabilities*.

As examples, the 2002 Sarbanes-Oxley Act made financial auditing commonplace as a tool to detect fraud and confirm the accuracy

[1] A simple example of black-box access is how most users interact with Google Search and ChatGPT. In both cases, a user submits a query or prompt, to which they receive a response in the form of search rankings for Google and an answer for ChatGPT. This constitutes black-box access in that one can collect input-output pairs without knowledge of the underlying mechanism.

[2] By "prospective," we mean that (i) the audit can be run before an AI system is deployed and (ii) the AI system can be tested on hypothetical examples (e.g., extreme contexts) that need not exist.

and completeness of financial reports; to earn a fair trade certification, vendors regularly undergo audits to ensure they uphold fair trade practices; many organizations audit themselves to detect, e.g., financial waste; and software designers often undertake audits to discover security vulnerabilities.

Generally, there are three types of auditors: internal, external, and third-party auditors. An **internal** auditor is selected from within an organization that seeks to audit itself. An **external** auditor is an independent party that is hired by the organization (e.g., its board of directors) to perform an audit. External auditors typically enter into a contractual agreement with the organization that outlines, for instance, the scope of the audit and level of engagement. Like an external auditor, a **third-party** auditor is not a part of the organization being audited. Unlike external auditors, third-party auditors are not hired by the organization that they audit. For example, a third-party auditor may be a regulatory agency or be employed by a regulatory agency. Third-party auditors also include journalists and non-profits (as long as they are not hired by the organization being audited). In order to ensure that auditors provide objective and high-quality reports, *auditors are also subject to audits*. It is customary for auditors to audit one another and, in certain industries, auditors are overseen by government agencies, such as the Public Company Accounting Oversight Board (PCAOB) in the US and the Financial Reporting Council (FRC) in the UK.

Audits can be initiated at different times and run with varying frequency. **Retrospective audits** evaluate a system's past behavior. For example, audits that examine financial records or system performance are retrospective. These retrospective audits can be ad hoc (in response to specific events), continuous, or periodic. In contrast, **prospective audits** are proactive. They can either (i) be performed before deployment or (ii) be tested on examples/contexts that have not yet arisen. As an example of (i), prospective audits can be run prior to an AI system's release or after major modifications to evaluate potential risk. As an example of (ii), prospective audits can also assess how a system *would* behave under conditions that have not yet occurred, e.g., in extreme scenarios.

Although out of the scope of this work, a final consideration for auditing is the development of relevant **metrics**, **measurement methods**, and **standards**, which has become a central focus of the U.S. National Institute for Standards and Technology (NIST) and European Commission. These efforts center around *what* AI audits should test for, which we take to be predetermined in this work in order to focus on AI auditing's implementation challenges.

2.2 AI auditing

Legally required AI audits. While some organizations audit themselves and some organizations are audited by journalists, researchers, and more, there have historically been few instances of legally mandated AI audits. That is beginning to change; Table 1 provides a (non-exhaustive) list of auditing requirements for the European Union (EU) Artificial Intelligence (AI) Act, the EU General Data Protection Regulation (GDPR), the California Consumer Privacy Act (CCPA), the US Algorithmic Accountability Act (AAA), the New York City (NYC) Local Law 144, Canada's Directive on Automated Decision-Making (CDADM), and the EU Digital Services Act (DSA).

Laws that indirectly mandate AI audits. In addition to those detailed in Table 1, there are several domains in which AI audits are indirectly required. For example, the Dodd-Frank Wall Street Reform and Consumer Protection Act and Sarbanes-Oxley Act (SOX) both require audits to check for compliance with financial regulations. Although neither explicitly mention AI or algorithms, they have both become common tools in organizations that are required to comply with Dodd-Frank and SOX. As a result, they indirectly fall within the scope of these required audits.

Other contexts. Audits are also a common tool for ensuring compliance with the law, even when the audits are not explicitly mentioned. Regulatory bodies and third-parties will often audit in order to hold organizations legally or publicly accountable, e.g., the Children's Online Privacy Protection Act (COPPA) does not explicitly mention audits, but the Federal Trade Commission (FTC) conducts investigations to verify compliance with COPPA's mandates [50].

Academic researchers, companies, and journalists may also conduct audits that fall outside the scope of compliance auditing. These audits are often used to verify that AI satisfies *normative* standards and meets developers' claimed specifications, even if these standards are not (or not yet) legally mandated. Some of these audits are an extension of longstanding approaches to AI and ML *evaluation*, which has historically focused on measures of performance (e.g., popular benchmark datasets like those in the UCI Machine Learning Repository [73] or ImageNet [42], and standardized evaluation suites like the Holistic Evaluation of Language Models (HELM) [82], among many others). Others have advocated for audits that look beyond performance, toward considerations such as bias, discrimination, and equity [114]. Several well known audit studies, such as those that brought light to potential discrimination in criminal justice algorithms [7] and facial recognition algorithms [21], underscore the need for such audits.

Efforts to audit for these kinds of harms are rapidly evolving, particularly in the context of newer generative AI technologies, but some common practices have started to emerge. These include "sock puppet audits," in which auditors programmatically impersonate users of an online platform to investigate the platform's algorithmic behavior [58, 60, 63, 65, 114], and "sociotechnical audits," in which auditors explicitly study interactions between humans and algorithms to contextualize an algorithm's behavior relative to the social environment in which it is deployed [80]. Raji et al. [108] propose an end-to-end framework for internal audits, e.g., those which are voluntarily conducted by a company or model provider, which nevertheless provides guidance on audits more broadly.

Auditing frameworks and methodology have also emerged for more specific contexts, including the auditing of social media platforms [29, 31], auditing for (a particular notion of) fairness via active learning [128], and auditing for differential privacy [69, 74, 98]. Red teaming, a classical approach to ensuring the security of software systems, has emerged as a popular method for evaluating modern generative AI systems (e.g., large language models) [54, 103, 111, 131], though questions remain about its efficacy in this context [51]. We provide additional background on other forms of AI auditing in Appendix A, and a discussion of other related work in Appendix B.

Table 1: Examples of legislation that require audits of data-driven or AI algorithms

Law	Enforced by	Performed by	Audit frequency and requirements	Penalty
EU GDPR (2016)	Data Protection Authorities in EU member states	Data controllers (typically internal)	Data Protection Impact Assessments (DPIAs): Description of data processing, purposes, risks to rights & freedoms of subjects, measures to address risks. Conducted before high-risk data processing.	Up to €20M or 4% of annual worldwide turnover (whichever is higher) for severe violations.
EU AI Act (2023)	National authorities in EU member states	AI system providers (internal); must give national competent authorities & notified bodies access (third-party)	High-risk AI systems must undergo conformity assessments to ensure they meet requirements for safety, transparency, human oversight, data, and more (as laid out in Title III, Chapter 2). Conducted before system on market, ongoing post-market monitoring, and whenever system is substantially modified.	Determined by member states; Some infringements up to €30M or 6% of annual worldwide turnover, whichever is higher
CCPA (2018)	California Attorney General	Businesses whose data processing presents significant risks to consumer privacy or security	Cybersecurity Audit must assess effectiveness of business' cybersecurity measures in protecting consumer personal information. Risk Assessment weighs the benefits of processing personal information against potential risks to consumer rights. Cybersecurity Audit performed on annual basis; Risk Assessment performed on regular basis (unspecified)	Up to $7.5K per intentional violation; additional penalties given by California Privacy Protection Agency
US AAA (2023†)	Federal Trade Commission (FTC)	Covered entities (businesses using AI systems)	Evaluation of automated decision system's or augmented critical decision process' potential impacts on consumers, considering privacy, bias, fairness, transparency, and more. Conducted on an ongoing basis, with annual reports required.	Determined by the FTC
NYC 144 (2021)	NYC Dept. of Consumer & Worker Protection	Employer/agency using Automated Employment Decision Tool (internal); can use independent auditor (external)	Checks whether automated employment decision tools have disparate impact on persons of any "component 1 category"; summary must be made publicly available. Conducted prior to first use and annually.	Up to $1.5K per instance; others determined by enforcement body
CDADM (2019)	Treasury Board of Canada Secretariat	Federal institutions using automated decision systems	Assess effect of automated decision-making systems on individual/community rights, economic interests, sustainability, and more. Conducted early in development, before release, and after major changes.	Case-by-case, as determined by the Treasury Board
EU DSA (2022)	Digital Service Coordinators in each EU member state and the EC	Independent organizations with restrictions (e.g., cannot audit >10 consecutive years)	Tests compliance with the obligations set out in Chapter III of the DSA and voluntary commitments (e.g., in code of conduct or crisis protocol). Conducted annually.	Up to 6% of annual worldwide turnover for severe violations; ongoing penalties of up to 5% average daily turnover

† Proposed but not passed

Limitations on audit access. Auditors are typically limited in their ability to access the systems they audit. In some cases, limited access is mandated by the law, e.g., the CCPA states that "nothing in this section [on risk assessments] shall require a business to divulge trade secrets" and GDPR similarly states that data processing measures (including audits) "should be appropriate, necessary and proportionate in view of ensuring compliance with this Regulation." In addition to trade-secret and privacy concerns, auditor access may be limited for cybersecurity reasons. At times, third-party auditors do not notify (and therefore do not cooperate with) the organization being audited, and thus have minimal access to the AI system. This consideration of *access* will be central to our discussion.

3 Illustrating AI auditing challenges

To highlight the challenges of implementing AI audits, we will ground our discussion in a case study of New York City (NYC) Local Law 144, which we will return to throughout our discussion.

NYC Local Law 144. This law, which was enacted in 2021 and took effect in 2023, governs the use of automated employment decision tools (AEDTs). It "prohibits employers and employment agencies from using an automated employment decision tool unless the tool has been subject to a bias audit within one year of the use of the tool, information about the bias audit is publicly available, and certain notices have been provided to employees or job candidates" [1]. The law defines a bias audit as an "impartial evaluation by an independent auditor" that assesses "the tool's disparate impact" with respect to job category, sex, and race or ethnicity; and an AEDT is defined as any "computational process" that is "derived machine learning, statistical modeling, data analytics, or artificial intelligence" that substantially supports, assists, or replaces any "discretionary decision-making" task in hiring or promotion, e.g., by providing an assessment of a candidate's skills or screening candidates for interviews [99].

Although the administrative code enacted in 2021 does not specify implementation details, rules released by city agencies in 2023 provided more specific guidance that we discuss below [43].

Regulations providing guidelines for NYC Local Law 144. The rules released in 2023 provide greater details on what should be reported in the audits and what data should be used. It specifies that a bias audit must, at minimum, calculate selection rates, scoring rates, median scores, and impact ratios for race/ethnicity, sex, and intersectional categories. The audit may also interrogate other features of the AEDT, though no further assessment is required.

The rules also require that bias audits be conducted on "*historical data*" (data collected during the employer's or employment agency's use of the AEDT or, in some circumstances, others' use of an AEDT). This audit must be performed on sufficient historical data to be "*statistically significant,*" though the rules do not clarify key elements of this requirement, which can lead to two very different outcomes (see Section 5). The law also makes allowances for the use of "*test data*" (any data that is not historical data, such as synthetic inputs) only if sufficient historical data is unavailable to conduct a statistically significant audit, though it does not provide explicit requirements or guidance on how such data should be acquired or generated. However, the law does require that a "summary of results of the bias audit must explain why historical data was not used and describe how the test data used was generated and obtained" [43]. Similarly, the law does not restrict the ability to exclude certain historical data (e.g., specific time periods) from the audit, but requires that these choices be publicly disclosed.

Interestingly, *imputed* and *inferred* data cannot be used in a bias audit, implying that if race, ethnicity, or gender is not reported for an individual, then they would be excluded from the corresponding bias analysis, even if the AEDT imputes or infers race, ethnicity, or gender from the individual's other attributes. Thus, even if the historical dataset contains many samples, statistically significant results may only be possible with the help of synthetic data if candidates' demographics were not collected by the employer/employment agency. Several other elements of the bias audit remain unspecified, including whether the bias audit is performed over the AEDT alone, over the entire decision-making process that includes the AEDT, or over a subset of it.

Open implementation challenges. The rules discussed above provide clarifications on NYC Local Law 144 but also defer many implementation details to the auditor (or relevant standards-setting body), including the manner in which the auditor *accesses* the algorithm, *what* evidence they should collect, *how* one should interpret that evidence, and more.

For example, there are many cases in which test data may be required (e.g., if there is insufficient historical data to conduct a statistical test, or demographic information is missing), but there are multiple ways in which the auditor could construct the test data, potentially yielding very different results. As another example, the rules require that bias audits yield statistically significant results, but what model should be used to assess "statistical significance?"

Both questions are implicitly related to the "access question" that motivates our discussion: creating test data requires, at a minimum, query (or black-box) *access* to a system, and constructing a reasonable model similarly requires insights into the AEDT (e.g., a form of *access* to the training procedure), as we discuss extensively in Section 4. Finally, there are also challenges around audit reporting, such as how the auditor should communicate the results of (often highly technical) audits, which we discuss in Section 5.

Connections to this work. As the prior section highlights, these concerns are broadly relevant for AI audits, thus motivating our focus, in Section 4, on the *type* of access an auditor needs to audit AI systems and, in Section 5, on a principled and flexible framework for (1) determining what *additional* assumptions or information are necessary to conduct an AI audit, and (2) gathering and interpreting evidence gathered from the audit. In Section 5, we discuss what kind of access to an AI system is necessary to perform a meaningful audit. We survey possible options and argue that the ability to *query* an AI system—construct arbitrary inputs and observe the system's outputs—is almost always necessary (i.e., minimally required) to perform an informative audit. In other words, auditors should at least be granted *black-box* access to an AI system. We proceed to connect AI auditing to the field of hypothesis testing in Section 5, which we find provides a principled framework for collecting and interpreting evidence in the context of an audit. In particular, we map the regulatory process of gathering evidence in

service of an audit (e.g., to gather conclusive evidence of compliance or noncompliance) to well-known statistical frameworks, which serve to illustrate the assumptions—in addition to black-box query access—under which AI auditing is feasible. Finally, we show that hypothesis testing provides a rigorous and well-studied framework for auditing which remains interpretable.

4 Types of auditing access

In this section, we address our motivating question: what form of access does an auditor require to perform a meaningful audit? In practice, answering this question is not straightforward due to competing concerns: while the auditor necessarily requires some form of access to the underlying system to conduct an audit, it is often desirable to keep this access "minimal" to avoid unnecessarily exposing proprietary technologies, compromising private data, and introducing cybersecurity risks, among other concerns. To address this question, we discuss the relative merits of four types of auditor access: access to the training data, the training procedure, the model architecture, and (white- and black-box access) to the trained model. Our discussion below hews closely to that of Cen et al. [28].

4.1 Option 1: Access to training data

AI models learn patterns and relationships that are exhibited in their *training data*. Because this data is one of the key determinants of an AI system's behavior, an auditor may wish to examine it for suggestive evidence of potential harms or failure modes. For example, over- or under-representation of a population in the training data can lead to bias [35]. Similarly, differences between the test and training data distributions can lead to generalization failures [134]. Nonetheless, access to the training data alone is typically insufficient for a rigorous audit. The primary reason is that the same training data can induce many different downstream models, whose behavior ultimately depends on the entire training pipeline (e.g., the choice of hyperparameters, model architecture, and learning algorithm). Although they arise from the same training data, these models may differ substantially along nearly any dimension of interest, including accuracy, fairness, and robustness. This "model multiplicity" [17] or "underspecification" [40] is an unavoidable feature of modern AI systems. As such, it is not only difficult, but often impossible for an auditor to conclusively characterize a system's behavior from its training data alone.

Nonetheless, examining the training data can play a key role in AI auditing. For instance, auditing a company's data acquisition, cleaning, balancing, privacy, and provenance practices can encourage good data hygiene. Although requiring that the training data meet a strict set of property requirements does not generally have the intended effect (most bright-line rules are easily circumvented due to the underspecification phenomenon mentioned above), *data disclosure audits* can encourage responsible developer practices and prevent downstream harms [55, 96, 106]. For example, in the context of modern large language models (LLMs), the common practice of publishing a "knowledge cutoff date"—roughly, the most recent date of the data on which the LLM was trained—may obscure important heterogeneity in the model's temporal understanding across different tasks; disclosure audits may therefore encourage developers to release finer grained or resource-specific knowledge cutoffs that could enable more informed use [32].

4.2 Option 2: Access to training procedures

Another option is to grant an auditor access to the *training procedure*, which includes details such as the general model class (e.g., decision trees, linear models, transformers), objective function used to optimize parameters, hyperparameter tuning methodology, quality checks, or model selection criteria. The training procedure can be viewed as a roadmap for how the AI system was developed. For example, around 2017, one of the objective functions in Facebook's "News Feed" algorithm placed five times more weight on "reactions" than it did on "likes" when ranking content in a user's feed [92]. This change had the unintended consequence of amplifying emotional content; indeed, the company's own data scientists found that posts that "sparked [the] angry reaction emoji were disproportionately likely to include misinformation, toxicity and low-quality news" [92]. These issues were not detected by Facebook until 2019 and only made public in 2021 by a whistleblower [92]. An earlier audit of the training procedure, including the objective functions, may have encouraged the company to scrutinize and justify (and, if appropriate, abandon) these design choices.

However, as with the training data, a particular training procedure can yield many downstream models, and there is no guarantee that these models behave similarly. The resulting model depends on various other factors, including the training data, model weights at initialization, and more. Therefore, while an auditor who is given access to a system's training procedure can perform sanity checks, they typically cannot draw precise conclusions about the system's ultimate behavior. Furthermore, since the training procedure lays out the steps taken to produce the AI system, access to training procedures is carefully guarded by companies; of the forms of access discussed, the training procedure is arguably the most valuable intellectual property associated with a commercial product.

4.3 Option 3: Access to the model skeleton

The next form of access we consider is access to the model "skeleton", which we also refer to as the "untrained" model. This "skeleton" refers to the specific model *class* that is used (e.g., the specific neural network architecture) and the way that different system components interact (e.g., how an AI model interacts with other models within the same system). Importantly, this skeleton is disclosed without the model parameters (i.e., weights).

One defining feature of this form of access is that it reveals the key interfaces of an AI system. From the model skeleton, an auditor can determine the expected inputs (e.g., types of features) and outputs (e.g., a floating point number between 0 and 1) of the model. The auditor can also ascertain how many components make up the AI system and the relationship between different components of the AI system. For example, suppose that a job applicant's information and resume are first sorted into one of several job categories, processed by appropriate algorithms, before being assigned a score between 0 and 1, which is finally thresholded to produce a hiring recommendation. Then, this entire "pipeline" would be captured by the model skeleton. In this way, access to the model skeleton provides perhaps the most *interpretable* view of an AI system. Indeed, when the social media platform X (formerly known as Twitter)

voluntarily released their recommendation model skeleton, [123], it revealed qualitative insights into X's content curation algorithm. For example, the public could glean that X "sources half of a user's content from in-network tweets (i.e., from accounts that the user follows) and the other half from out-of-network tweets" [28].

As with the training procedure, a model skeleton allows the auditor to perform sanity checks to flag obvious flaws. It can even be used to identify discrepancies between an AI company's claims and the deployed system. The model skeleton alone, however, is not enough to characterize the precise behavior of an AI system. As mentioned in Section 4.1, models with the exact same skeleton can behave very differently, making it difficult to verify whether an AI system complies with a specific rule or meets a given standard from access to the model skeleton alone.

4.4 Option 4: Access to the trained model

We now turn to the last option we consider: access to the final trained model. In contrast to the other three forms of access, access to the trained model allows the auditor to inspect the *specific* model that is or will be released.

There are several different flavors of access to the trained model. *Black-box access* allows the auditor to query the model on the auditor's choice of inputs and observe the outputs but nothing more, yielding a series of input-output pairs. For example, sending ChatGPT queries (or "prompts") and observing how ChatGPT responds is a form of black-box access. Similarly, observing the ratings that a hiring algorithm assigns to job applicants is also a form of black-box access. On the other hand, *model-weight access* allows the auditor to not only query the model, but also to see the entire trained model, including the trained model parameters (i.e., the model weights). By analogy, one can think of black-box access as an auditor being able to crash-test a car, whereas an auditor with model-weight access would also be able to inspect every component of the car. In between these two, there are other forms of access that have been explored, including *fine-tuning access* (the ability to fine-tune the final trained model on a dataset of interest) and *log-probabilities access* (the ability to view "probabilities" the model assigns to different prediction outcomes, before the final prediction is produced in most neural networks). In this way, there are tiers of access to the trained model. For instance, model-weight access is strictly stronger than log-probabilities access (in that one can always obtain log probabilities from model weights), which is stronger than black-box access.

Of the four options considered in this section, only access to the trained model circumvents the problem of "underspecification" or "model multiplicity" [17, 40]. This implies that *access to the trained model is minimally necessary for meaningful audits* because it best indicates how the system behaves when deployed. However, there are limitations to auditing the trained model alone. An auditor cannot gain insight into the developers' process and reasoning from the trained model alone, which is not always satisfactory from a broader accountability perspective. Much of US law, for instance, considers *intent* when determining culpability. Moreover, auditing developers' process and reasoning can encourage safer and more thoughtful industry practices. As such, further access is often needed.

4.5 Key takeaway: Black-box access is minimally necessary

The four types of access described above are not exhaustive. For example, *white-box access* allows the auditor to view the model weights, training procedure, and source code but typically not the training data. Although there are numerous access options, one can view them as follows: putting resource constraints aside, the *stronger* the form of access, the closer the auditor is to being able to *reproduce* the AI system from scratch. Therefore, granting auditors access which is overly broad can introduce its own risks, such as undermining a company's competitive advantage or allowing bad actors to exploit system vulnerabilities.

We argue that black-box access—the weakest form of access to the trained model—should be minimally required because it provides concrete evidence about the specific AI system being audited (as discussed in Section 4.4). In addition, it has several characteristics that may address implementation difficulties that auditors face, as enumerated below:

(1) **Minimal access.** Black-box access only allows the auditor to view the outputs of an AI system. It thus allows the auditor to assess the system that is ultimately deployed without further insight into the underlying implementation, which can allay concerns about revealing trade secrets, compromising data privacy, or exposing security vulnerabilities.
(2) **Prospective.** Although this property is shared with model-weight and log-probabilities access (but not all flavors of trained-model access), black-box access gives the auditor the ability to prospectively inspect the AI system. That is, the auditor can test the AI system on inputs of their choice and observe how the system behaves in situations that may not have arisen. This allows the auditor to perform pre-deployment assessments, and it can also be used to stress-test the system under extreme scenarios that have not yet arisen in natural datasets.
(3) **Model agnostic.** Because black-box audits do not look "under the hood," they are agnostic to the inner workings of the AI system. Perhaps the key benefit is that the audit does not need to be adapted to the underlying AI system, making it possible to develop a *standardized* audit procedure that works across multiple AI systems. For example, even when developers change their model architecture, an auditor does not need to devise a new black-box audit as long as the input-output "types" are consistent. Another benefit of model agnosticism is that it allows the auditor to test how the model behaves end-to-end without necessarily requiring that the auditor be technically proficient (which is often needed if an auditor wishes to conduct model-specific audits).
(4) **Well suited to AI.** Finally, it is worth remarking that AI systems are well-suited to black-box audits. Consider, for example, auditing a firm's (non-AI) hiring practices for discrimination. A black-box audit would require gathering everyone who plays a role in hiring, providing them with a set of applications, asking them to evaluate each as they normally would, and observing their decisions. This process is not scalable, as it would require auditors and the firm to invest significant time and resources.

Perhaps more importantly, those involved with the audit can easily manipulate the outcome by misreporting their true preferences on who to hire. On the other hand, it is straightforward to repeatedly query an AI system, and the results of black-box AI audits are guaranteed to be faithful to how the AI system would behave in practice.

As discussed in this section, no form of access is uniformly better than its alternatives, and the appropriate form of access is context-dependent. We argue however that black-box access is minimally necessary for an informative audit, reduces some of the risks associated with stronger forms of access, and offers other advantages to the auditor. This however begs the question: if black-box access is minimally necessary, *in what contexts is stronger access needed, and how should one develop standardized audit protocols based on the available level of access?* We examine this question next.

5 Hypothesis testing and its connections to audits

In the prior section, we discuss different forms of access to an AI system, and argue that black-box access is often necessary to conduct an informative audit. When, however, does the auditor need more than black-box access, and how should the auditor interpret the evidence that they gather? In this section, we draw parallels between auditing and hypothesis testing, using NYC Local Law 144 (as described in Section 3) as an illustration. We show that this connection helps (i) clarify how an auditor can interpret and communicate the results of an audit, and (ii) produce precise guidance on how much access an auditor needs. We highlight that the null hypothesis can be viewed as a formalization of the relevant legal presumption, and the evidence needed to reject the null hypothesis can be viewed as the corresponding burden of proof. We provide a discussion of related work in hypothesis testing in Appendix A.

5.1 Setup

Consider a model developer or operator, who we refer to as the *AI provider* for the remainder of this section. The provider employs an algorithm $f \in \mathcal{F}$, where $\mathcal{F}$ is a class of mappings from values in $\mathcal{X}$ to distributions over a countable set $\mathcal{Y}$ as denoted by $\Delta(\mathcal{Y})$. For example, in the context of hiring decisions, f could map an applicant's characteristics $x \in \mathcal{X}$ to a prediction $f(x) \in \mathcal{Y} \subset [0,1]$ of the applicant's fit for a given role. Let p_x denote the true (possibly unknown) marginal distribution of x. The auditor is interested in determining whether the provider's algorithm f complies with a requirement of interest. We denote this requirement by $g : \mathcal{F} \to \mathbb{R}$.

Definition 1. *We say that an algorithm $f \in \mathcal{F}$ is g-*compliant *if and only if $g(f) \leq 0$.*

When the property g is clear from context, we simply say an algorithm is *compliant*. In a *black-box audit*, the auditor has access to N input-output pairs $(x_i, f(x_i))$. We denote the *evidence* that an auditor has access to (including that gathered through black-box access) by $\mathcal{E}$. Thus, the auditor's task is as follows:

Determine whether f is g-compliant given evidence $\mathcal{E}$.

Example 1 (Maximum loss). *Requiring that f's maximum loss ℓ over some $S \subseteq \mathcal{X}$ is at most η is equivalent to requiring that $g(f) \leq 0$, where $g(f) = \max_{x \in S} \ell(f(x), x) - \eta$. Depending on the definition of loss, one can audit for minimax fairness (by defining loss as negative performance), worst-case harm (by defining loss as the output's harm, e.g., toxicity level), and even copyright infringement (by defining loss as the dissimilarity between x and the copyrighted work).*

Example 2 (Group fairness). *In the area of algorithmic fairness, group fairness typically reflects a notion of parity across groups. For example, one notion of group fairness known as "statistical parity" requires that the rate at which a binary classifier $f : \mathcal{X} \to \{0,1\}$ selects members of group G_1 is at most $\eta > 0$ far from the rate at which f selects members of group G_2 under some distribution p_x over $\mathcal{X}$. This is equivalent to requiring that $g(f) \leq 0$, where $g(f) = |\mathbb{E}[f(x) \mid x_G = G_1] - \mathbb{E}[f(x) \mid x_G = G_2]| - \eta$. The expectation above is taken over $x \sim p_x$, and $x_G \in \{G_1, G_2\}$ is the feature in x denoting group membership.*

Example 3 (Individual fairness). *Another notion of algorithmic fairness requires that "similar individuals be treated similarly," as captured by the criterion: $D(f(x), f(x')) \leq Ld(x, x')$ for all $x, x' \in \mathcal{X}$; distance metrics D and d on $\mathcal{Y}$ and $\mathcal{X}$, respectively; and Lipschitz constant $L > 0$ [45]. This is equivalent to requiring that $g(f) \leq 0$, where $g(f) = \max_{x,x' \in \mathcal{X}} \frac{D(f(x), f(x'))}{d(x,x')} - L$.*

One could similarly cast selection rates or impact ratios (from Section 3) in this format.

5.2 Hypothesis testing and the burden of proof

Given an algorithm f and auditing criterion g, the auditor seeks to determine whether f is g-compliant using $\mathcal{E}$. Below, we discuss two possible hypothesis tests before describing the general hypothesis testing procedure in Section 5.3.

Presumption of compliance. Consider an auditor who seeks to discern which of the following hypotheses holds:

$$H_0 : g(f) \leq 0, \qquad H_1 : g(f) > 0. \tag{1}$$

Let $H \in \{H_0, H_1\}$ denote the ground-truth state. For example, if f is compliant, then $H = H_0$; if f is not, then $H = H_1$.

The auditor does not know H a priori. Therefore, the auditor's goal is to develop a decision test or *rule* $\hat{H}$ such that $\hat{H}(\mathcal{E}) = H_0$ if the auditor believes f is compliant and $\hat{H}(\mathcal{E}) = H_1$, otherwise. The auditor would like $\hat{H}$ to match H for all $f \in \mathcal{F}$ because this means that the auditor has neither false positives nor false negatives, as formalized in Section 5.3.

H_0 is known as the *null hypothesis*. In practice, the implication is that the auditor's *presumption* under (1) is that f is compliant. The auditor therefore assumes (and reports) that f is compliant *unless the evidence allows them to confidently reject this presumption.* This framework therefore highlights what evidence is needed for an auditor to reject H_0 when H_1 is indeed the true hypothesis, i.e., what information the auditor needs to prove that an AI system is non-compliant when it is indeed non-compliant. Though subtle, *this point is crucial.* By viewing auditing as hypothesis testing, it becomes clear what it means for the available information to be (in)sufficient for auditing. Below, we show that the hypothesis test can be reversed to instead presume non-compliance.

Presumption of non-compliance. Consider a different set of hypotheses:

$$J_0 : g(f) > 0, \qquad J_1 : g(f) \leq 0 \tag{2}$$

Relative to (1), the null and alternate hypotheses have been swapped. As before, there is a ground-truth state $J \in \{J_0, J_1\}$, and the auditor's goal is to develop a decision rule $\hat{J}$ such that, given evidence $\mathcal{E}$, the decision $\hat{J}$ approximates J well across all $f \in \mathcal{F}$. In this case, the null hypothesis J_0 (i.e., the legal presumption) is that the algorithm is not compliant.

Burden of proof: Which test should the auditor use? The null hypothesis reflects the auditor's presumption and, accordingly, who bears the burden of proof. For example, NYC Local Law 144 requires bias audits to be "statistically significant" but does not specify the null hypothesis. among other necessary modeling assumptions that we discuss further below. When such audits are required but do not explicitly require the AI provider to disclose any information to auditors, then, under (1), the AI provider is *not incentivized* to disclose information (i.e., to contribute evidence $\mathcal{E}$ to the auditing process). To see this, observe that since the auditor can only reject the null hypothesis $H_0 : g(f) \leq 0$ if they have enough evidence to do so, the burden of proof is on the auditor or governing body. Therefore, NYC Local Law 144, as it stands, incentivizes employers and employment agencies to release minimal amounts of historical data (recall that it does not specify the timeframe of the data that must be provided).

On the other hand, the burden of proof under (2) is on the AI provider. That is, the AI provider is incentivized to give the auditor enough evidence to convince the auditor to reject the null hypothesis J_0, e. g., the employer is incentivized to prove that their hiring process is not biased. In this way, the choice of hypothesis test should reflect the desired legal presumption and corresponding placement of burden of proof. This choice may vary across contexts. For example, if the auditor's evidentiary burden is too great under (1), and the one may wish to shift the evidentiary burden via (2).

5.3 Hypothesis testing procedure

In this section, we describe the procedure for casting an AI audit as a hypothesis test. We refer readers interested in hypothesis testing to Casella and Berger [26] for a textbook treatment. For the remainder of this work, we adopt a presumption of compliance as given in (1), though our results can be equivalently applied to (2). We discuss four main components of hypothesis testing next: the evidence, decision rule, model, and tolerance.

(1) **Evidence.** The auditor has access to evidence $\mathcal{E}$, as defined in Section 5.1. The auditor is generally limited in the amount of evidence they can gather, for example, due to strictly controlled access to algorithm f or its training data, or due to limited resources. In the context of NYC Local Law 144, evidence is the historical or test data.

(2) **Decision rule.** Given evidence $\mathcal{E}$, the auditor's goal is to develop an audit—which, in hypothesis testing, is called a decision rule $\hat{H}$—that maps evidence $\mathcal{E}$ to a decision H_0 or H_1 that correspond to deciding whether to report that f is compliant or non-compliant, respectively. Under (1), the auditor adopts the default decision H_0 unless the evidence is convincing enough for the auditor to reject H_0, as we review next.

(3) **Design criteria & tolerance.** A decision rule $\hat{H}$ is evaluated based on two quantities: the false positive rate (FPR) and true positive rate (TPR), $\text{FPR} = \mathbb{P}\left(\hat{H} = H_1 | H = H_0\right)$ and $\text{TPR} = \mathbb{P}\left(\hat{H} = H_1 | H = H_1\right)$, where $\mathbb{P}$ is taken with respect to randomness in the evidence $\mathcal{E}$ and decision rule $\hat{H}$. (Observe that the true negative rate and false negative rate can be computed directly from the FPR and TPR.) Hypothesis testing is largely concerned with finding rules that maximize the TPR while minimizing the FPR. Although we do not review them here, one approach is to restrict the maximum allowable FPR (known as the *significance level*) to ζ and find the decision rule that achieves the maximum TPR among all rules with an FPR no more than ζ and for all possible algorithms in $\mathcal{F}$. This rule is known as the uniformly most powerful (UMP) test and can be treated as an ideal benchmark (though it does not always exist). The allowable FPR can be viewed as the *tolerance* of an audit.

Thus, one interpretation of "statistical significance" for NYC Local Law 144 is to perform the UMP test given a pre-specified tolerance. However, there may be more fundamental issues with the casual use of "statistical significance" in NYC Local Law 144 for the same reasons discussed at the end of Section 5.2. In particular, the term "statistical significance" generally only applies when the null hypothesis is rejected, which implies that "statistical significance" cannot be achieved if one adopts the null hypothesis in (1) and the null hypothesis is indeed true. Therefore, asking that "enough" historical data or additional test data be provided until the audit is statistically significant may be problematic, and more careful consideration of the hypothesis test may be needed.

(4) **Model.** Given evidence $\mathcal{E}$, the auditor's job is to determine whether $g(f) \leq 0$. Doing so necessarily requires some assumptions (i.e., a *model*) about how f generates $\mathcal{E}$. For example, the auditor may assume that x are drawn from some known distribution $\mathcal{D}$. By definition, the auditor does not know the algorithm f that they wish to audit, but the auditor may assume that f belongs to some model family $\tilde{\mathcal{F}} \subset \mathcal{F}$. In this case, the model is determined by $(\mathcal{D}, \tilde{\mathcal{F}})$. Without a model, the auditor lacks the necessary assumptions to inform a decision rule; moreover, both FPR and TPR cannot be defined without a model. Crucially, the model clarifies what additional *information*, if any, is needed to conduct an audit. In the context of NYC Local Law 144, the key model that the auditor needs is the (empirical) test distribution; given this distribution, it is not clear that anything stronger than black-box access is needed since the law seeks to audit for disparate *impact* (i.e., the outcome and not the process) and the metrics of interest are rates (which can be estimated from simple sampling protocols). In other more complex cases (e.g., privacy or "unlearning" audits [25, 115]), further assumptions and further access may be needed.

Therefore, to conduct an audit, the auditor takes the following steps: decide on an appropriate model and tolerance, develop a decision rule, gather evidence, and apply the decision rule.

6 Limitations and future work

Multiple testing. While our work considers testing a single property of a model, auditors are typically interested in auditing for multiple criteria or running repeated audits over time. This can present additional challenges, as (1) the reuse of data across audits will invalidate basic statistical guarantees and (2) even audits run on independent samples will not (on their own) control the family-wise error rate or false discovery rate [11]. These issues are exacerbated when the number of audits is not known ex-ante, and may depend on the results of prior audits.

Explanations, recourse and the limits of auditing. While audits that can be cast as hypothesis test can be powerful tools, they do not necessarily indicate what the provider should do to correct or mitigate these issues. Indeed, as argued in Casper et al. [27] and discussed in Section 4, it is possible that white- and gray-box approaches can be more informative in this regard. For example, hypothesis tests would not necessarily reveal whether the model made a mistake or otherwise behaved unreasonably on a *specific instance*. In such cases, a more localized (or "counterfactual-based") approach to auditing might be appropriate [3, 30, 81]. Furthermore, black-box auditing does not necessarily enable appropriate recourse when an individual is harmed by an algorithm, as the result of a black-box audit does reveal the reasoning behind a developer's design choices or their intentions.

Active learning for auditing. When given black-box access to an algorithm, auditors must choose the inputs that they wish to test, i.e., the set S on which to run f. The simplest methods involve sampling instances independently from some population of interest, or otherwise specifying S *a priori*. However, a natural approach would also consider choosing these instances in an *online* fashion, in which successive samples are chosen conditional on the output of prior queries. An online sampling procedure will naturally improve the power (i.e., the true positive rate) of an audit at any fixed false positive rate. This class of algorithms is sometimes referred to as *active learning* [59].

Manipulation-proofness. The auditor may also have other concerns, such as ensuring that the audit is manipulation-proof. This direction is concerned with removing loopholes that may permit AI developers or companies to pass audits in practice without satisfying the desired criteria in spirit [128] or guaranteeing that the evidence gathered during an audit reflects how the AI system behaves when deployed [122].

Acknowledgments

Thank you to Manish Raghavan, Aleksander Madry, Martha Minow, Cosimo Fabrizio, and James Siderius for their valuable feedback on this work. The authors gratefully acknowledge funding from the MIT-IBM project on Causal Representation, a Stephen A. Schwarzman College of Computing Seed Grant, the National Science Foundation (NSF) grant CNS-1955997, and the Air Force Research Laboratory (AFOSR) grant FA9550-23-1-0301.

References

[1] 2023. DCWP - Automated Employment Decision Tools (AEDT). https://www.nyc.gov/site/dca/about/automated-employment-decision-tools.page. Accessed: 2024-9-20.

[2] Philip Adler, Casey Falk, Sorelle A. Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. 2016. Auditing Black-box Models for Indirect Influence. arXiv:arXiv:1602.07043

[3] Nil-Jana Akpinar, Liu Leqi, Dylan Hadfield-Menell, and Zachary Lipton. 2022. Counterfactual Metrics for Auditing Black-Box Recommender Systems for Ethical Concerns. In *ICML 2022 Workshop on Responsible Decision Making in Dynamic Environments*. Baltimore, Maryland, USA.

[4] Rohan Alur, Loren Laine, Darrick K. Li, Manish Raghavan, Devavrat Shah, and Dennis Shung. 2023. Auditing for Human Expertise.

[5] Rohan Alur, Manish Raghavan, and Devavrat Shah. 2024. Distinguishing the Indistinguishable: Human Expertise in Algorithmic Prediction. arXiv:arXiv:2402.00793

[6] Joshua D. Angrist and Jörn-Steffen Pischke. 2008. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

[7] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (May 23 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[8] Ian Ayres, Mahzarin Banaji, and Christine Jolls. 2015. Race effects on eBay. *The RAND Journal of Economics* 46, 4 (2015), 891–917.

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

[10] Robert P Bartlett, Adair Morse, Nancy Wallace, and Richard Stanton. 2019. Algorithmic Accountability: A Legal and Economic Framework. http://faculty.haas.berkeley.edu/morse/research/papers/AlgorithmicAccountability_BartlettMorseStantonWallace.pdf

[11] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300.

[12] Richard A. Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50 (2017), 3 – 44. https://api.semanticscholar.org/CorpusID:12924416

[13] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* 94, 4 (Aug. 2004), 991–1013.

[14] Felix Bieker, Michael Friedewald, Marit Hansen, Hannah Obersteller, and Martin Rost. 2016. A Process for Data Protection Impact Assessment Under the European General Data Protection Regulation. In *Annual Privacy Forum*. https://api.semanticscholar.org/CorpusID:7904695

[15] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2017. Evasion Attacks against Machine Learning at Test Time. ECML PKDD, Part III, vol. 8190, LNCS, pp. 387–402. Springer, 2013. (2017). https://doi.org/10.1007/978-3-642-40994-3_25 arXiv:arXiv:1708.06131

[16] Battista Biggio and Fabio Roli. 2017. Wild Patterns: Ten Years After the Rise of Adversarial Machine Learning. (2017). https://doi.org/10.1016/j.patcog.2018.07.023 arXiv:arXiv:1712.03141

[17] Emily Black, Manish Raghavan, and Solon Barocas. 2022. Model Multiplicity: Opportunities, Concerns, and Solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM, Seoul, Republic of Korea, 23. https://doi.org/10.1145/3531146.3533149

[18] Emily Black, Samuel Yeom, and Matt Fredrikson. 2019. FlipTest: Fairness Testing via Optimal Transport. (2019). https://doi.org/10.1145/3351095.3372845 arXiv:arXiv:1906.09218

[19] Laura Blattner, Scott Nelson, and Jann Spiess. 2021. Unpacking the Black Box: Regulating Algorithmic Decisions. arXiv:arXiv:2110.03443

[20] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2017. Decision-Based Adversarial Attacks: Reliable Attacks Against Black-Box Machine Learning Models. *ArXiv* abs/1712.04248 (2017). https://api.semanticscholar.org/CorpusID:2410333

[21] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *FAT*. https://api.semanticscholar.org/CorpusID:3298854

[22] Nadia Burkart and Marco F. Huber. 2020. A Survey on the Explainability of Supervised Machine Learning. Journal of Artificial Intelligence Research (JAIR), 70:245-317, 2021. (2020). https://doi.org/10.1613/jair.1.12228 arXiv:arXiv:2011.07876

[23] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. 2009. Building Classifiers with Independency Constraints. *2009 IEEE International Conference on Data Mining Workshops* (2009), 13–18. https://api.semanticscholar.org/CorpusID:3945595

[24] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21 (2010), 277–292. https://api.semanticscholar.org/CorpusID:12856537

[25] Yinzhi Cao and Junfeng Yang. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*. IEEE, 463–480.

[26] George Casella and Roger L. Berger. 2008. *Statistical Inference*. Thomson Learning.
[27] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. 2024. Black-Box Access is Insufficient for Rigorous AI Audits. arXiv:arXiv:2401.14446
[28] Sarah H. Cen, Cosimo L. Fabrizio, James Siderius, Aleksander Madry, and Martha Minow. 2023. Auditing AI: How Much Access Is Needed to Audit an AI System? https://aipolicy.substack.com/p/ai-accountability-transparency-2.
[29] Sarah H. Cen, Aleksander Madry, and Devavrat Shah. 2023. A User-Driven Framework for Regulating and Auditing Social Media. arXiv:arXiv:2304.10525
[30] Sarah H. Cen and Manish Raghavan. 2022. The Right to be an Exception to a Data-Driven Rule. arXiv:arXiv:2212.13995
[31] Sarah H. Cen and Devavrat Shah. 2020. Regulating algorithmic filtering on social media. In *Neural Information Processing Systems*. https://api.semanticscholar.org/CorpusID:219721119
[32] Jeffrey Cheng, Marc Marone, Orion Weller, Dawn Lawrie, Daniel Khashabi, and Benjamin Van Durme. 2024. Dated Data: Tracing Knowledge Cutoffs in Large Language Models. arXiv:arXiv:2403.12958
[33] John J. Cherian and Emmanuel J. Candès. 2023. Statistical Inference for Fairness Auditing. arXiv:arXiv:2305.03712
[34] Alexandra Chouldechova. 2016. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. arXiv:arXiv:1610.07524
[35] Alexandra Chouldechova and Aaron Roth. 2018. The Frontiers of Fairness in Machine Learning. arXiv:arXiv:1810.08810
[36] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Z Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2017). https://api.semanticscholar.org/CorpusID:3228123
[37] Sasha Costanza-Chock, Emma Harvey, Inioluwa Deborah Raji, Martha Czernuszenko, and Joy Buolamwini. 2023. Who Audits the Auditors? Recommendations from a field scan of the algorithmic auditing ecosystem. (2023). https://doi.org/10.1145/3531146.3533213 arXiv:arXiv:2310.02521
[38] Cynthia S Crowson, Elizabeth J Atkinson, and Terry M Therneau. 2016. Assessing calibration of prognostic risk scores. *Stat. Methods Med. Res.* 25, 4 (Aug. 2016), 1692–1706.
[39] Nilesh Dalvi, Pedro Domingos, Mausam, Sumit Sanghai, and Deepak Verma. 2004. Adversarial classification. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA) *(KDD '04)*. Association for Computing Machinery, New York, NY, USA, 99–108. https://doi.org/10.1145/1014052.1014066
[40] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yian Ma, Cory McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. 2022. Underspecification presents challenges for credibility in modern machine learning. *J. Mach. Learn. Res.* 23, 1, Article 226 (jan 2022), 61 pages.
[41] Amit Datta, Michael Carl Tschantz, and Anupam Datta. 2015. Automated Experiments on Ad Privacy Settings: A Tale of Opacity, Choice, and Discrimination. In *Privacy Enhancing Technologies Symposium*.
[42] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 248–255.
[43] Department of Consumer and Worker Protection. 2023. Rules of the City of New York. https://rules.cityofnewyork.us/citations/automated-employment-decision-tools/. Title 6, Chapter 5, Subchapter T: Automated Employment Decision Tools.
[44] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS Risk Scales : Demonstrating Accuracy Equity and Predictive Parity Performance of the COMPAS Risk Scales in County. https://api.semanticscholar.org/CorpusID:51920414
[45] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Rich Zemel. 2011. Fairness Through Awareness. arXiv:arXiv:1104.3913
[46] Benjamin Edelman and Zhenyu Lai. 2016. Design of Search Engine Services: Channel Interdependence in Search Engine Results. *Journal of Marketing Research* 53 (03 2016). https://doi.org/10.1509/jmr.14.0528
[47] Benjamin G. Edelman, Michael Luca, and Dan Svirsky. 2015. Racial Discrimination in the Sharing Economy: Evidence from a Field Experiment. http://ssrn.com/abstract=2701902.
[48] Harrison Edwards and Amos J. Storkey. 2015. Censoring Representations with an Adversary. *CoRR* abs/1511.05897 (2015). https://api.semanticscholar.org/CorpusID:4986726
[49] Edwin Farley. 2023. AI Auditing: First Steps Towards the Effective Regulation of Artificial Intelligence Systems. Available at SSRN. https://ssrn.com/abstract=4676184
[50] Federal Trade Commission. 2022. *Policy Statement of the Federal Trade Commission on Education Technology and the Children's Online Privacy Protection Act.* Policy Statement. Federal Trade Commission. https://www.ftc.gov/system/files/ftc_gov/pdf/Policy%20Statement%20of%20the%20Federal%20Trade%20Commission%20on%20Education%20Technology.pdf
[51] Michael Feffer, Anusha Sinha, Wesley Hanwen Deng, Zachary C. Lipton, and Hoda Heidari. 2024. Red-Teaming for Generative AI: Silver Bullet or Security Theater? arXiv:arXiv:2401.15897
[52] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Eduardo Scheidegger, and Suresh Venkatasubramanian. 2014. Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2014). https://api.semanticscholar.org/CorpusID:2077168
[53] Anthony W. Flores, Kristin Bechtel, and Christopher T. Lowenkamp. 2016. False Positives, False Negatives, and False Analyses: A Rejoinder to "Machine Bias: There's Software Used across the Country to Predict Future Criminals. and It's Biased against Blacks". *Federal Probation* 80 (2016), 38. https://api.semanticscholar.org/CorpusID:15391140
[54] Deep Ganguli, Liane Lovitt, John Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Benjamin Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zachary Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom B. Brown, Nicholas Joseph, Sam McCandlish, Christopher Olah, Jared Kaplan, and Jack Clark. 2022. Red Teaming Language Models to Reduce Harms: Methods, Scaling Behaviors, and Lessons Learned. *ArXiv* abs/2209.07858 (2022). https://api.semanticscholar.org/CorpusID:252355458
[55] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92.
[56] Jeffrey Gleason, Desheng Hu, Ronald E Robertson, and Christo Wilson. 2023. Google the Gatekeeper: How Search Components Affect Clicks and Attention. *Proceedings of the International AAAI Conference on Web and Social Media* 17 (2 6 2023), 245–256.
[57] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and Harnessing Adversarial Examples. *CoRR* abs/1412.6572 (2014). https://api.semanticscholar.org/CorpusID:6706414
[58] Anikó Hannák, Gary Soeller, David M. J. Lazer, Alan Mislove, and Christo Wilson. 2014. Measuring Price Discrimination and Steering on E-commerce Web Sites. *Proceedings of the 2014 Conference on Internet Measurement Conference* (2014). https://api.semanticscholar.org/CorpusID:8839357
[59] Steve Hanneke. 2013. A statistical theory of active learning. *Foundations and Trends in Machine Learning* (2013), 1–212.
[60] Anikó Hannák, Piotr Sapieżyński, Arash Molavi Khaki, David Lazer, Alan Mislove, and Christo Wilson. 2017. Measuring Personalization of Web Search. arXiv:arXiv:1706.05011
[61] Anikó Hannák, Claudia Wagner, David Garcia, Alan Mislove, Markus Strohmaier, and Christo Wilson. 2017. Bias in Online Freelance Marketplaces: Evidence from Taskrabbit and Fiverr. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 1914–1933.
[62] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of Opportunity in Supervised Learning. arXiv:arXiv:1610.02413
[63] Muhammad Haroon, Magdalena E. Wojcieszak, Anshuman Chhabra, Xin Liu, Prasant Mohapatra, and Zubair Shafiq. 2023. Auditing YouTube's recommendation system for ideologically congenial, extreme, and problematic recommendations. *Proceedings of the National Academy of Sciences of the United States of America* 120 (2023). https://api.semanticscholar.org/CorpusID:265905645
[64] Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J. Watts. 2023. Causally estimating the effect of YouTube's recommender system using counterfactual bots. arXiv:arXiv:2308.10398
[65] Homa Hosseinmardi, Amir Ghasemian, Miguel Rivera-Lanas, Manoel Horta Ribeiro, Robert West, and Duncan J. Watts. 2023. Causally estimating the effect of YouTube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences of the United States of America* 121 (2023). https://api.semanticscholar.org/CorpusID:261048649
[66] Ferenc Huszár, Sofia Ira Ktena, Conor O'Brien, Luca Belli, Andrew Schlaikjer, and Moritz Hardt. 2022. Algorithmic amplification of politics on Twitter. *Proc. Natl. Acad. Sci. U. S. A.* 119, 1 (Jan. 2022), e2025334119.
[67] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *International Conference on Machine Learning*. https://api.semanticscholar.org/CorpusID:5046541

[68] Guido W. Imbens and Donald B. Rubin. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction.* Cambridge University Press.

[69] Matthew Jagielski, Jonathan Ullman, and Alina Oprea. 2020. Auditing Differentially Private Machine Learning: How Private is Private SGD? arXiv:arXiv:2006.07709

[70] Adrianne Jeffries and Leon Yin. 2020. Google's Top Search Result? Surprise! It's Google. https://themarkup.org/google-thegiant/2020/07/28/google-search-results-prioritize-googleproducts-over-competitors. Accessed: 2023-04-18.

[71] James E. Johndrow and Kristian Lum. 2017. An algorithm for removing sensitive information: Application to race-independent recidivism prediction. *The Annals of Applied Statistics* (2017). https://api.semanticscholar.org/CorpusID:51782788

[72] Faisal Kamiran and Toon Calders. 2009. Classifying without discriminating. *2009 2nd International Conference on Computer, Control and Communication* (2009), 1–6. https://api.semanticscholar.org/CorpusID:1102398

[73] Markelle Kelly, Rachel Longjohn, and Kolby Nottingham. 2024. The UCI Machine Learning Repository. https://archive.ics.uci.edu. Accessed: 2024-09-20.

[74] Daniel Kifer, Solomon Messing, Aaron Roth, Abhradeep Thakurta, and Danfeng Zhang. 2020. Guidelines for Implementing and Auditing Differentially Private Systems. arXiv:arXiv:2002.04049

[75] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. arXiv:arXiv:1609.05807

[76] C. Kliman-Silver, 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 Internet Measurement Conference.*

[77] Dariusz Kloza, Niels van Dijk, Simone Casiraghi, Sergi Vazquez Maymir, Sara Roda, Alessia Tanas, and Ioulia Konstantinou. 2020. Towards a method for data protection impact assessment: Making sense of GDPR requirements. https://api.semanticscholar.org/CorpusID:208222562

[78] Tamar Kricheli-Katz and Tali Regev. 2016. How many cents on the dollar? Women and men in product markets. *Science Advances* 2, 2 (2016).

[79] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. 2016. Adversarial Machine Learning at Scale. *ArXiv* abs/1611.01236 (2016). https://api.semanticscholar.org/CorpusID:9059612

[80] Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, and Danaë Metaxa. 2023. Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 360 (oct 2023), 37 pages. https://doi.org/10.1145/3610209

[81] Seung C. Lee. 2022. A black box approach to auditing algorithms. *Issues In Information Systems* (2022).

[82] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher R'e, Diana Acosta-Navas, Drew A. Hudson, E. Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel J. Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan S. Kim, Neel Guha, Niladri S. Chatterji, O. Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas F. Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. Holistic Evaluation of Language Models. *Annals of the New York Academy of Sciences* 1525 (2023), 140 – 146. https://api.semanticscholar.org/CorpusID:253553585

[83] Zachary C. Lipton. 2016. The Mythos of Model Interpretability. arXiv:arXiv:1606.03490

[84] Daniel Lowd and Christopher Meek. 2005. Adversarial learning. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining* (Chicago, Illinois, USA) *(KDD '05)*. Association for Computing Machinery, New York, NY, USA, 641–647. https://doi.org/10.1145/1081870.1081950

[85] Daniel Lowd and Christopher Meek. 2005. Good Word Attacks on Statistical Spam Filters. In *International Conference on Email and Anti-Spam.* https://api.semanticscholar.org/CorpusID:1933015

[86] Michael Luca, Tim Wu, Sebastian Couvidat, Daniel Frank, and William Seltzer. 2016. Does Google Content Degrade Google Search? Experimental Evidence. https://api.semanticscholar.org/CorpusID:53570166

[87] Scott Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. arXiv:arXiv:1705.07874

[88] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2017. Towards Deep Learning Models Resistant to Adversarial Attacks. arXiv:arXiv:1706.06083

[89] Subha Maity, Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. 2021. Statistical inference for individual fairness. arXiv:arXiv:2103.16714

[90] Natalie Maus, Patrick Chao, Eric Wong, and Jacob Gardner. 2023. Black Box Adversarial Prompting for Foundation Models. arXiv:arXiv:2302.04237

[91] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2019. A Survey on Bias and Fairness in Machine Learning. arXiv:arXiv:1908.09635

[92] Jeremy B. Merrill and Will Oremus. 2021. Five points for anger, one for a 'like': How Facebook's formula fostered rage and misinformation. *The Washington Post* (26 Oct 2021). https://www.washingtonpost.com/technology/2021/10/26/facebook-angry-emoji-algorithm/

[93] Danaë Metaxa, Michelle A. Gan, and James A. Landay. 2021. An Image of Society: Gender and Racial Representation and Impact in Image Search Results for Occupations. https://api.semanticscholar.org/CorpusID:233322946

[94] Danaë Metaxa, Joon Park, James Landay, and Jeff Hancock. 2019. Search Media and Elections: A Longitudinal Investigation of Political Search Results. *Proceedings of the ACM on Human-Computer Interaction* 3 (11 2019), 1–17. https://doi.org/10.1145/3359231

[95] Danaë Metaxa, Joon Sung Park, Ronald E Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing algorithms: Understanding algorithmic systems from the outside in. *Found. Trends® Hum.–Comput. Interact.* 14, 4 (2021), 272–344.

[96] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 220–229.

[97] Christoph Molnar. 2022. *Interpretable Machine Learning.* https://christophm.github.io/interpretable-ml-book/

[98] Milad Nasr, Jamie Hayes, Thomas Steinke, Borja Balle, Florian Tramèr, Matthew Jagielski, Nicholas Carlini, and Andreas Terzis. 2023. Tight Auditing of Differentially Private Machine Learning. arXiv:arXiv:2302.07956

[99] New York City Council. 2023. New York City Administrative Code. https://codelibrary.amlegal.com/codes/newyorkcity/latest/NYCadmin/0-0-0-123415. Title 20, Chapter 5, Subchapter 25: Automated Employment Decision Tools.

[100] Anh Nguyen, Jason Yosinski, and Jeff Clune. 2014. Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. arXiv:arXiv:1412.1897

[101] Brendan Nyhan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, Sandra González-Bailón, Andrew M Guess, Edward Kennedy, Young Mie Kim, David Lazer, Neil Malhotra, Devra Moehler, Jennifer Pan, Daniel Robert Thomas, Rebekah Tromble, Carlos Velasco Rivera, Arjun Wilkins, Beixian Xiong, Chad Kiewiet de Jonge, Annie Franco, Winter Mason, Natalie Jomini Stroud, and Joshua A Tucker. 2023. Like-minded sources on Facebook are prevalent but not polarizing. *Nature* 620, 7972 (Aug. 2023), 137–144.

[102] Judea Pearl. 2009. *Causality: Models, Reasoning, and Inference.* Cambridge University Press.

[103] Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nathan McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. In *Conference on Empirical Methods in Natural Language Processing.* https://api.semanticscholar.org/CorpusID:246634238

[104] Fábio Perez and Ian Ribeiro. 2022. Ignore Previous Prompt: Attack Techniques For Language Models. arXiv:arXiv:2211.09527

[105] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. 2017. On Fairness and Calibration. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (Eds.). 5680–5689. https://proceedings.neurips.cc/paper/2017/hash/b8b9c74ac526fffbeb2d39ab038d1cd7-Abstract.html

[106] Mahima Pushkarna, Andrew Zaldivar, and Oddur Kjartansson. 2022. Data Cards: Purposeful and Transparent Dataset Documentation for Responsible AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency.* 1776–1826.

[107] Inioluwa Deborah Raji. 2023. The Anatomy of AI Audits: Form, Process, and Consequences. In *The Oxford Handbook of AI Governance.* Oxford University Press. https://doi.org/10.1093/oxfordhb/9780197579329.013.28 arXiv:https://academic.oup.com/book/0/chapter/421725648/chapter-ag-pdf/54770606/book_41989_section_421725648.ag.pdf

[108] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (2020). https://api.semanticscholar.org/CorpusID:209862020

[109] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel E. Ho. 2022. Outsider Oversight: Designing a Third Party Audit Ecosystem for AI Governance. arXiv:arXiv:2206.04737

[110] Ashesh Rambachan, Jon M. Kleinberg, Sendhil Mullainathan, and Jens Ludwig. 2020. An Economic Approach to Regulating Algorithms. *NBER Working Paper Series* (2020). https://api.semanticscholar.org/CorpusID:214775707

[111] Javier Rando, Daniel Paleka, David Lindner, Lennard Heim, and Florian Tramèr. 2022. Red-Teaming the Stable Diffusion Safety Filter. *ArXiv* abs/2210.04610

(2022). https://api.semanticscholar.org/CorpusID:252780252
[112] Bashir Rastegarpanah, Krishna P. Gummadi, and Mark Crovella. 2021. Auditing Black-Box Prediction Models for Data Minimization Compliance. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2021*.
[113] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. arXiv:arXiv:1602.04938
[114] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. https://api.semanticscholar.org/CorpusID:15686114
[115] Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A Smith, and Chiyuan Zhang. 2024. Muse: Machine unlearning six-way evaluation for language models. *arXiv preprint arXiv:2407.06460* (2024).
[116] Latanya Sweeney. 2013. Discrimination in Online Ad Delivery. http://ssrn.com/abstract=2208240.
[117] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. arXiv:arXiv:1312.6199
[118] Eva Thelisson and Himanshu Verma. 2024. Conformity assessment under the EU AI act general approach. *AI Ethics* (Jan. 2024).
[119] Florian Tramèr, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, Jean-Pierre Hubaux, Mathias Humbert, Ari Juels, and Huang Lin. 2015. FairTest: Discovering Unwarranted Associations in Data-Driven Applications. arXiv:arXiv:1510.02377
[120] Aleksandra Urman, Mykola Makhortykh, and Aniko Hannak. 2024. Mapping the Field of Algorithm Auditing: A Systematic Literature Review Identifying Research Trends, Linguistic and Geographical Disparities. arXiv:arXiv:2401.11194
[121] Claudia Wagner, David Garcia, Mohsen Jadidi, and Markus Strohmaier. 2015. It's a Man's Wikipedia? Assessing Gender Inequality in an Online Encyclopedia. In *International AAAI Conference on Web and Social Media*.
[122] Suppakit Waiwitlikhit, Ion Stoica, Yi Sun, Tatsunori Hashimoto, and Daniel Kang. 2024. Trustless Audits without Revealing Data or Models. *arXiv preprint arXiv:2404.04500* (2024).
[123] Kyle Wiggers. 2023. Twitter reveals some of its source code, including its recommendation algorithm. *TechCrunch* (31 Mar 2023). https://techcrunch.com/2023/03/31/twitter-reveals-some-of-its-source-code-including-its-recommendation-algorithm/
[124] Folkert Wilman. 2022. The Digital Services Act (DSA) - An Overview. (16 12 2022). https://ssrn.com/abstract=4304586
[125] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. 2017. Learning Non-Discriminatory Predictors. *ArXiv* abs/1702.06081 (2017). https://api.semanticscholar.org/CorpusID:2047106
[126] Lei Xu, Yangyi Chen, Ganqu Cui, Hongcheng Gao, and Zhiyuan Liu. 2022. Exploring the Universal Vulnerability of Prompt-based Learning Paradigm. arXiv:arXiv:2204.05239
[127] Songkai Xue, Mikhail Yurochkin, and Yuekai Sun. 2020. Auditing ML Models for Individual Bias and Unfairness. In *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy] (Proceedings of Machine Learning Research, Vol. 108)*, Silvia Chiappa and Roberto Calandra (Eds.). PMLR, 4552–4562. http://proceedings.mlr.press/v108/xue20a.html
[128] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (Eds.). PMLR, 24929–24962. https://proceedings.mlr.press/v162/yan22c.html
[129] Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. 2020. A Survey on Causal Inference. arXiv:arXiv:2002.02770
[130] Karen Yeung. 2018. A Study of the Implications of Advanced Digital Technologies (Including AI Systems) for the Concept of Responsibility Within a Human Rights Framework. *Social Science Research Network* (2018). https://api.semanticscholar.org/CorpusID:158736157
[131] Jiahao Yu, Xingwei Lin, Zheng Yu, and Xinyu Xing. 2023. GPTFUZZER: Red Teaming Large Language Models with Auto-Generated Jailbreak Prompts. *ArXiv* abs/2309.10253 (2023). https://api.semanticscholar.org/CorpusID:262055242
[132] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *Proceedings of the 26th International Conference on World Wide Web* (2016). https://api.semanticscholar.org/CorpusID:1911971
[133] Matthew D Zeiler and Rob Fergus. 2013. Visualizing and Understanding Convolutional Networks. arXiv:arXiv:1311.2901
[134] Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. 2021. Domain Generalization: A Survey. (2021). https://doi.org/10.1109/TPAMI.2022.3195549 arXiv:arXiv:2103.02503

A AI auditing techniques

Background: Audit studies. Audit studies have a long history in the social sciences. Bertrand and Mullainathan [13] find in a study of labor market discrimination that resumes with White-sounding names receive 50 percent more callbacks, on average, than identical resumes with African-American-sounding names. This evidence was gathered by submitting a set of fictitious resumes in response to real help-wanted ads, allowing researchers to experimentally manipulate the perceived race of job applicants in a way akin to the black-box audits described in this work.

Taking inspiration from this tradition, there is now a growing body of literature that audits algorithmic systems for evidence of consumer harm. Investigators have employed audit studies to examine self-preferencing in search results [46, 56, 70, 86], discrimination in online platforms [8, 41, 47, 61, 78, 93, 114, 116, 121], and the effects of algorithmic personalization (particularly on political polarization) [60, 64, 66, 76, 94, 101]. A key challenge is that the inputs to these systems are often highly complex, and may not be directly manipulable by researchers. This motivates other causal identification strategies, e.g., by identifying natural experiments in observational data (see Yao et al. [129] for a recent survey or Angrist and Pischke [6], Imbens and Rubin [68], Pearl [102] for textbook treatments). For additional background on audit studies, including the legal and ethical questions that arise, as well as recommendations for best practices, we refer to Metaxa et al. [95]. For systematic reviews of the algorithm auditing literature, we refer to Bandy [9] and Urman et al. [120].

Frameworks for algorithmic auditing. Perhaps most closely related to this work are general frameworks for ensuring that algorithms satisfy normative and regulatory constraints. As discussed in Section 2, Raji et al. [108] propose a framework which guides the development life cycle of an algorithmic decision pipeline. In contrast, we provide an in-depth discussion of black-box audits, propose a way to translate between the law and audit procedure, and describe an open problem related to query complexity. Lam et al. [80] take a different perspective, and instead propose the notion of a *socio-technical* audit to directly study the interplay between algorithms and their users. In particular, a socio-technical audit involves experimentally manipulating the *outputs* of an algorithm—for example, via a browser extension which manipulates search results or social media feeds—to study human components of a system (e.g., how user react or modify their behavior) in addition to algorithmic components. Farley [49] advocates for government-mandated audits of AI systems, and outlines a regulatory framework intended to standardize the practice of AI auditing. Farley [49] focuses on the policy aspects of AI auditing; in contrast, our work develops a framework for implementing AI audits. Raji et al. [109] and [37] survey the current practice of auditing and make recommendations for enabling the effective oversight and regulation of algorithms. In the case of Raji et al. [109], these lessons are drawn primarily from other fields where third-party (or "outsider") audits have proven effective.

Finally, [19] propose a framework for regulating algorithms based on model explanations that are tailored to capture specific model characteristics—for example, racial disparities in the model's

predictions—rather than to best explain the model's average performance. We further discuss the relationship of auditing and these interpretability techniques in Appendix B.

Black-box auditing. Our work focuses on black-box auditing, where the auditor may only *query* the model, rather than e.g., inspecting source code, model architecture or training procedures. This approach is intended to enable third party oversight of algorithms [109], even in the face of limited cooperation by algorithm providers [37]. This aligns with the perspective taken in Cen et al. [29], which propose auditing procedures for algorithms which curate content on social media platforms. It is also the approach taken in Rastegarpanah et al. [112], who develop algorithms for testing compliance with the GDPR's data minimization principle (that an algorithm uses only "the minimal information that is necessary for performing the task at hand" [112]). This is also similar to the perspective taken in Akpinar et al. [3], Lee [81], which propose the use of black-box audits to assess *counterfactuals*. For example, such an audit might ask whether, for a given individual, the algorithmic recommendation changes if the individual's race were different. As we discuss in Section 6, these localized audits can be useful for individuals seeking recourse for algorithmic harms.

Finally, contemporaneous work by Casper et al. [27] argues that a black-box approach is insufficient for rigorous auditing, and highlight the limitations of black-box queries. These include (1) the difficulty of developing a global understanding of how a system behaves, (2) the inability to study system components separately, (3) the possibility that overly simplistic black-box audits can produce misleading results, (4) the limitations of black-box interpretability methods and (5) the inability to suggest *remedies* when models are noncompliant. We share the view that broader access (e.g., to model weights, gradients or source code) can enable more in depth auditing of algorithmic systems, and we discuss the benefits and limitations of black-box auditing at length in Section 4. Given the strictly controlled access provided to auditors (see our remark above) and concerns such as privacy, our discussion of black-box audits is driven by a desire to explore what can be achieved with black-box access, which can be supplemented with further access to cover its blind spots.

B Additional related work

Algorithmic fairness. Our work is inspired by a large literature on algorithmic fairness. This methodological work is itself inspired by well-publicized instances of real-world algorithmic discrimination (e.g, Chouldechova [34]). Of particular relevance to our work are the many definitions of fairness which have been proposed, including the notion of individual fairness [45], equalized odds [62], statistical parity or disparate impact [23, 24, 48, 52, 71, 72, 125, 132] and calibration [38, 44], [12, 53, 105]. Choosing a particular fairness measure is highly nontrivial, as imposing fairness constraints generally comes at some cost to model accuracy ([36]). Furthermore, many seemingly natural definitions of fairness turn out to be incompatible with each other ([75, 105]). This motivates alternative approaches to fairness which do not directly alter model training procedures [110].

Our work is most closely related to a smaller but growing literature which develops tests for specific kinds of algorithmic harms or failures. For example, Black et al. [18], Cherian and Candès [33], Yan and Zhang [128] develop tests for disparities in performance on important (and perhaps legally protected) subgroups, Xue et al. [127] and Maity et al. [89] propose algorithms to detect violations of individual fairness, Tramèr et al. [119] and Adler et al. [2] develop methods to understand how protected attributes influence model behavior (including indirectly). Alur et al. [4, 5] propose tests to detect whether algorithms fail to incorporate contextual information which may be available to a human decision maker, and Bartlett et al. [10] propose a framework for detecting 'input' or proxy discrimination. For additional background we refer to Chouldechova and Roth [35] and Mehrabi et al. [91] for surveys of the literature.

Explainable machine learning. Our work is also related to a large and growing literature on *explainable* (or interpretable) machine learning. Although we cannot provide a complete overview here, notable works include *LIME* [113], a technique for providing explanations for individual model predictions via black-box access, and *SHAP* [87], a technique for attributing individual model predictions to specific inputs ('features'). Zeiler and Fergus [133] propose a method for visualizing intermediate layers of a convolutional neural network. These works are broadly motivated by a desire to understand *why* and *how* machine learning models (particularly nonlinear models) make predictions. For additional background, including the challenges of defining model interpretability, we refer to Lipton [83]. For a survey and book-length treatment of specific techniques for model interpretability, we refer to Burkart and Huber [22], Molnar [97], respectively.

Adversarial attacks. Finally, our work on black-box auditing is complementary to a rich literature on adversarial machine learning, which seeks to discover (or mitigate against) adversarial inputs—often small perturbations of non-adversarial inputs—which 'fool' an algorithm into producing incorrect or incoherent outputs. Indeed, the robustness of algorithmic predictors to adversarial attacks is itself a natural property of interest for both internal and external auditors. Furthermore, the task of *generating* adversarial inputs using a sequence of black-box queries is very similar to the problem of auditing for extreme values, and both are naturally addressed via the machinery of online convex optimization.

Work on adversarial attacks against machine learning models dates to early email spam filters [39, 84, 85]. Much of the more recent literature on the vulnerability of deep neural networks to adversarial attacks can be traced to Szegedy et al. [117], who document the sensitivity of neural networks to imperceptible perturbations of their inputs. Notable work on adversarial attacks of deep neural networks includes Biggio et al. [15], Brendel et al. [20], Goodfellow et al. [57], Ilyas et al. [67], Kurakin et al. [79], Nguyen et al. [100]. To address these vulnerabilities, Madry et al. [88] propose an approach for training adversarially robust neural networks. More recently, Maus et al. [90], Perez and Ribeiro [104], Xu et al. [126] propose techniques for generating adversarial *prompts* for modern foundation models. For additional background on adversarial machine learning, we refer to Biggio and Roli [16].