# Access and Evidence in AI Auditing

**Sarah H. Cen** | Stanford (incoming Asst Prof at CMU)
Rohan Alur | MIT

# Audits

*An audit is the systematic evaluation of a system, often to determine whether it satisfies a predetermined set of criteria*

# Audits

*An audit is the systematic evaluation of a system, often to determine whether it satisfies a predetermined set of criteria*

Used for many reasons, including:

**Compliance testing**. Determines compliance with the law or contracts
**Verifying specifications**. Tests the company's or developer's own claims
**Risk assessment**. Evaluates possible risks, often before deployment
**Finding vulnerabilities**. Pinpoints weak points that can be exploited
**Ongoing monitoring**. Observe behavior "in the wild," after deployment
**Public accountability**. Checks alignment with industry or public standards

# Audits

*An audit is the systematic evaluation of a system, often to determine whether it satisfies a predetermined set of criteria*

Used for many reasons, including:

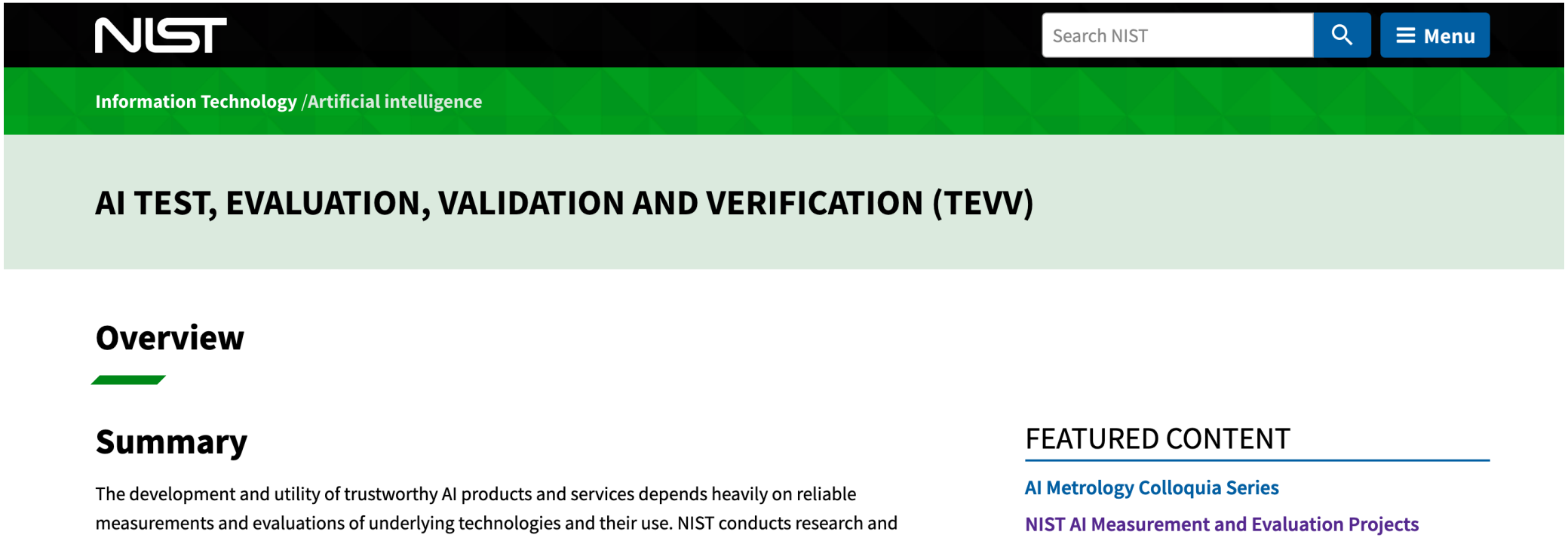
**Compliance testing**. Determines compliance
**Verifying specifications**. Tests the company's
**Risk assessment**. Evaluates possible risks, ofte
**Finding vulnerabilities**. Pinpoints weak points
**Ongoing monitoring**. Observe behavior "in th
**Public accountability**. Checks alignment with

Consider the US car industry. Audits help to…

Test compliance with federal safety & emissions regulations

Verify disclosed information (e.g., fuel economy)

# Growing consensus that AI audits matter



# Growing consensus that AI audits matter


| 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. |
| 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 | Determined by member states; Some infringements up to €30M or 6% of annual worldwide turnover, whichever is higher |

# Related Work: AI Auditing

Rich empirical and methodological literature

Bertrand and Mullainathan, 2024; Sweeney, 2013; Ayres et al., 2015; Datta et al., 2015; Luca et al., 2016; Hannák et al., 2017; Metaxa et al., 2021, Hosseinmardi et al., 2023 …

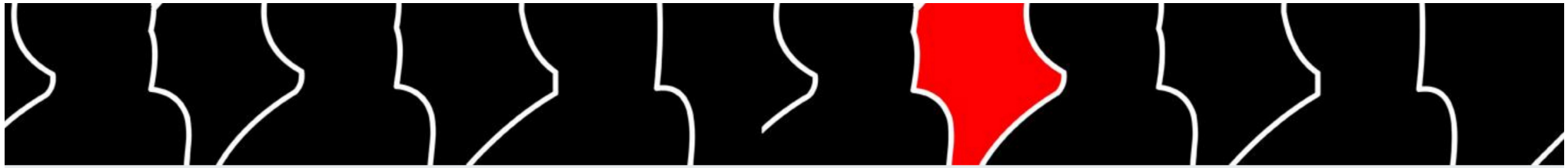
Sandvig et al., 2014; Rastegarpanah et al., 2021, Akpinar et al., 2022, Lee, 2022 …

Frameworks for auditing AI systems

Raji, 2023; Yeung, 2018; Mitchell et al., 2019; Raji et al., 2022; Costanza-Chock et al., 2023; Lam et al., 2023; See Bandy, 2021; Urman et al., 2024 for recent surveys

Auditing as hypothesis testing

Xue et al., 2020 (individual fairness); Cherian and Candès, 2023 (group fairness); Jayaraman and Evans, 2019; Lu et al., 2023; Nasr et al., 2023 (differential privacy)

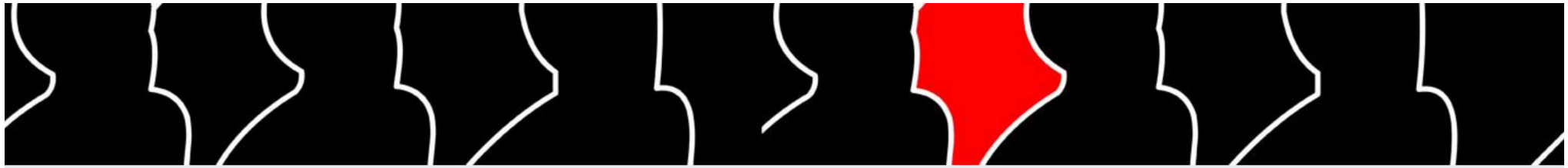


# Auditing Algorithms @ Northeastern

This site is the homepage for the Algorithm Auditing Research Group within the Khoury College of Computer Sciences at Northeastern University. Here, you will find explanations of and links to our work, as well as open-source data and code from our research.

## Why Audit Algorithms?

Today, we are surrounded by algorithmic systems in our everyday life. Examples on the web include Google Search, which personalizes search results to try and surface more relevant content; Amazon and Netflix, which recommend products and media; and Facebook, which personalizes each user's news-feed to highlight engaging content. Algorithms are also increasingly appearing in real world contexts, like surge pricing for vehicles from Uber; predictive policing algorithms that attempt to infer where crimes will occur and who will commit them; and credit scoring systems that determine eligibility for loans and credit cards. The proliferation of algorithms is driven by the explosion of Big Data that is available about people's online and offline behavior.



# Auditing Algorithms @ Northeastern

This site is the homepage for the Algorithm Auditing Research Group within the Khoury College of Computer Sciences at Northeastern University. Here, you will find explanations of and links to our work, as well as open-source data and code from our

## Sociotechnical Audits: Broadening the Algorithm Auditing Lens to Investigate Targeted Advertising

**Authors: Michelle S. Lam, Ayush Pandit, Colin H. Kalicki, Rachit Gupta, Poonam Sahoo, Danaë Metaxa**

### Abstract

Algorithm audits are powerful tools for studying black-box systems without direct knowledge of their inner workings. While very effective in examining technical components, the method stops short of a sociotechnical frame, which would also consider users themselves as an integral and dynamic part of the system. Addressing this limitation, we propose the concept of sociotechnical auditing: auditing methods that evaluate algorithmic systems at the sociotechnical level, focusing on the interplay between algorithms and users as each impacts the other. Just as algorithm audits probe an algorithm with varied inputs and observe outputs, a sociotechnical audit (STA) additionally

# Today

## I. What are the legal requirements around AI audits?

Survey of recent legislation

## II. What type of access is needed for AI auditing?

Discuss four types of access

Recommend, at minimum, black-box access

## III. How do we connect auditing techniques to the law?

Hypothesis testing mirrors legal procedure & informs who bears burden of proof

Clearly delineates what assumptions & *further* access (beyond black-box) are needed

# Background

# Various audit practices

**Audit purposes:** test for compliance, determine whether a technology meets standards, validate claims made by system designers, monitor an internal practices, uncover vulnerabilities, and more!

**Three types of auditors:** internal (within organization), external (outside but financially tied), independent (outside and financially independent)

**Timing of audits:** retrospective, prospective, ongoing

(Will not discuss metrics, measurement methods, and standards today)

| Law | Enforced by | Performed by | Audit frequency and requirements | Penalty |
|---|---|---|---|---|
| EU GDPR (2016) | Data Protection Authorities (DPAs) in EU member states. Overseen by European Data Protection Board (EDPB). | Data controllers (typically internal), potentially with the help of third parties and data processors. | Mandates Data Protection Impact Assessments (DPIAs): descriptions of data processing, purposes, risks to rights & freedoms of subjects, measures to address risks, as laid out in Article 35. DPIAs are required before high-risk data processing and when there is a change of the risk represented by processing. | Up to €10M or 2% of annual worldwide turnover (whichever is higher). Up to €20M or 4% of annual worldwide turnover (whichever is higher) for severe violations. |
| EU AI Act (2023) | National competent authorities in EU member states. Overseen by European Commission (EC). | AI system providers (internal) or notified bodies (third-party), depending on the existence of harmonized standards or common specifications. | High-risk AI systems must undergo conformity assessments to ensure they meet requirements for safety, transparency, human oversight, data, and more. Requires assessment before system is on the 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. |

| Law | Enforced by | Performed by | Audit frequency and requirements | Penalty |
|---|---|---|---|---|
| EU DSA (2022) | Digital Service Coordinator (DSC) in each EU member state and the EC. | Audits to be performed by independent auditor (external), with some guidelines (e.g., cannot audit >10 consecutive years). Risk assessments and ongoing monitoring to be conducted internally. | Requires independent audits of providers of very large online platforms and of very large online search engines that test compliance with the obligations set out in Chapter III of the DSA to be conducted annually. Also requires that they perform assessments of systemic risks and continuous monitoring of risk mitigation strategies. | Up to 6% of annual worldwide turnover for failure to comply; periodic penalties must not exceed 5% of average daily worldwide turnover or income per day. |
| NYC 144 (2021) | NYC Dept. of Consumer & Worker Protection (DCWP) | Independent auditor (external) | Requires bias audit (impartial evaluation) that tests whether automated employment decision tool's disparate impact on persons of any "component 1 category" to be conducted annually and prior to first use. A summary must be made publicly available. Conducted prior to first use and annually. | Up to $1.5K per instance; others determined by enforcement body. |

# Not all audits are legally mandated!

# Not all audits are legally mandated!

Laws that indirectly affect the use of AI (e.g., employment discrimination or fair lending laws)

Many audits are conducted by academic researchers, investigative journalists, non-profits, and more.

# Okay, maybe we're convinced that AI audits are important. So, what's the problem?

# Operational challenges

There are many open operational questions for AI audits, including:

- What should we be evaluating or measuring?
- How often should audits be run?
- Who audits the auditors?

"Digital Services Coordinators and the Commission shall use the data accessed pursuant to paragraph 1 only for **the purpose of monitoring and assessing compliance** with this Regulation and shall take due account of the rights and interests of the providers of very large online platforms or of very large online search engines and the recipients of the service concerned, including the **protection of personal data, the protection of confidential information, in particular trade secrets, and maintaining the security of their service**."

EU Digital Services Act, Article 40

"Digital Services Coordinators and the Commission shall use the data accessed pursuant to paragraph 1 only for **the purpose of monitoring and assessing compli**
interes
searc
**protec**
**parti**

"The Commission, market surveillance authorities and notified bodies and any other natural or legal person involved in the application of this Regulation shall, in accordance with Union or national law, **respect the confidentiality of information and data** obtained in carrying out their tasks and activities in such a manner as to protect … the **intellectual property rights and confidential business information or trade secrets of a natural or legal person, including source code**,"

EU Artificial Intelligence Act, Article 78

# Operational challenges

There are many open operational questions for AI audits, including:

- What should we be evaluating or measuring?
- How often should audits be run?
- Who audits the auditors?

**Today**: What access and evidence should auditors be granted?

**If we can't find a problem, we can't address it**

Auditing & transparency go hand-in-hand!

# Four types of access

# Option 1: Access to training data

AI models learn patterns from training data

“Garbage in, garbage out”

# Option 1: Access to training data

AI models learn patterns from training data

“Garbage in, garbage out”

**Benefits**: Auditing datasets or data procedures can

- Flag problems with data privacy or hygiene (e.g., balance)
- Encourage good data practices (e.g., provenance)

# Option 1: Access to training data

AI models learn patterns from training data

“Garbage in, garbage out”

**Benefits**: Auditing datasets or data procedures can

- Flag problems with data privacy or hygiene (e.g., balance)
- Encourage good data practices (e.g., provenance)

**Limitations**: Good data does not preclude harmful/unwanted outcomes

# Option 2: Access to training procedures

Training procedure = steps developer took to train the model

Such as model class, objective function(s), training algorithm

# Option 2: Access to training procedures

Training procedure = steps developer took to train the model

Such as model class, objective function(s), training algorithm

**Benefits**: Auditing training procedure is interpretable

- Provides sanity checks (recall Facebook's overweighting of emotion reacts)
- Is easy to compare to clear industry standards

# Option 2: Access to training procedures

Training procedure = steps developer took to train the model

Such as model class, objective function(s), training algorithm

**Benefits**: Auditing training procedure is interpretable

- Provides sanity checks (recall Facebook's overweighting of emotion reacts)
- Is easy to compare to clear industry standards

**Limitations**: Does not guarantee good outcomes and can be restrictive

# Option 3: Access to model skeleton

Model skeleton = “untrained” model

Exact model class (e.g., neural network architecture or decision tree)

# Option 3: Access to model skeleton

Model skeleton = “untrained” model

Exact model class (e.g., neural network architecture or decision tree)

**Benefits**: Model skeleton provides best birds-eye view:

- Conveys the input type, output type, how components “fit” together, etc.
- Provides sanity checks (e.g., identify discrepancies btw claims & skeleton)

# Option 3: Access to model skeleton

Model skeleton = “untrained” model

Exact model class (e.g., neural network architecture or decision tree)

**Benefits**: Model skeleton provides best birds-eye view:

- Conveys the input type, output type, how components “fit” together, etc.
- Provides sanity checks (e.g., identify discrepancies btw claims & skeleton)

**Limitations**: There are many possible models that can from same model class, and audits of model skeleton require technical fluency

# Option 4: Access to trained model

Includes: white-box, black-box, log-probabilities, fine-tuning access

# Option 4: Access to trained model

Includes: white-box, black-box, log-probabilities, fine-tuning access

**Benefits**: Unlike the other three, can directly test & probe the end product

- Black-box access does not require knowledge of inner workings
- White-box access can be used to probe the final model

# Option 4: Access to trained model

Includes: white-box, black-box, log-probabilities, fine-tuning access

**Benefits**: Unlike the other three, can directly test & probe the end product

- Black-box access does not require knowledge of inner workings
- White-box access can be used to probe the final model

**Limitations**: Does not account for intention or process. Plus, without further information, knowing how to query/probe is hard

# Considering all the options

**Auditing the final model provides the least ambiguity**

# Considering all the options

**Auditing the final model provides the least ambiguity**

**Auditors should, at minimum, receive black-box access**:

Minimal access

Good for security, proprietary tech and data, and technical fluency reasons

Model-agnostic

Does not need to be tailored to specific model → good for scalability, flexibility

Prospective

Can see how model would behave on hypothetical inputs

# Considering all the options

## Auditing the final model provides the least ambiguity

**Auditors should, at minimum, receive black-box access**:

Minimal access

Good for security, proprietary tech and data, and te

Model-agnostic

Does not need to be tailored to specific model → s

Prospective

Can see how model would behave on hypothetical

# Considering all the options

**Auditing the final model provides the least ambiguity**

**Auditors should, at minimum, receive black-box access**:

Minimal access

M

Pr

Can see how model would behave on hypothetical inputs

Black-box access alone can be **inefficient (or ineffective)**.
How much more information is needed for a meaningful audit?

# Determining access using HT

Hypothesis testing connects statistical methods to evidence & the law

# Hypothesis testing basics

Hypotheses

Null hypothesis $H_0$

Alternate hypothesis $H_1$

Decision Rule $\widehat{H}$

$$\max_{\widehat{H}} \mathbb{P}\left(\widehat{H} = H_1 \middle| H = H_1\right)$$

True Positive Rate (TPR)

$$\min_{\widehat{H}} \mathbb{P}\left(\widehat{H} = H_1 \middle| H = H_0\right)$$

False Positive Rate (FPR)

# Hypothesis testing basics

| Hypotheses | Null hypothesis $H_0$ | Alternate hypothesis $H_1$ |
| --- | --- | --- |
| Decision Rule $\widehat{H}$ | $\max_{\widehat{H}} \mathbb{P}(\widehat{H} = H_1 \mid H = H_1)$ | $\min_{\widehat{H}} \mathbb{P}(\widehat{H} = H_1 \mid H = H_0)$ |

$\mathbb{P}$ implies a set of assumptions

Allowable FPR is tolerance!

# Hypotheses ⟺ evidentiary burden

| | | |
|---|---|---|
| Test 1 | $H_0$: Compliant | $H_1$: Non-compliant |
| Test 2 | $H_0$: Non-compliant | $H_1$: Compliant |

# Hypotheses ⟺ evidentiary burden

| | | |
|---|---|---|
| Test 1 | $H_0$: Compliant | $H_1$: Non-compliant |
| Test 2 | $H_0$: Non-compliant | $H_1$: Compliant |

Only reject $H_0$ if you have enough evidence for doing so
Maps to **legal presumption** and **burden of proof**

# Benefits of hypothesis testing

**Clearly stated assumptions**. "Access" to model info informs assumptions

**Interpretable parameters**. Can map "tolerance" to FPR $\in [0,1]$

**HT is well studied**. Long line of work with community backing

**Mirrors legal procedure**. Null hypothesis = legal presumption

Can clearly inform what questions of access & evidentiary burdens!

# Today

## I. What are the legal requirements around AI audits?

Survey of recent legislation

## II. What type of access is needed for AI auditing?

Discuss four types of access

Recommend, at minimum, black-box access

## III. How do we connect auditing techniques to the law?

Hypothesis testing mirrors legal procedure & informs who bears burden of proof

Clearly delineates what assumptions & *further* access (beyond black-box) are needed

# Open directions

Future and ongoing directions:

1. Statistical tests that balance audit objective against constraints, such as trade secret protections

   (ongoing – come talk to me!)

2. Designing manipulation-proof audits under access restrictions
3. Characterizing “frontier” of achievable audit objectives

# Thank you!

shcen@stanford.edu