Audits of AI Systems

Brenda Leong

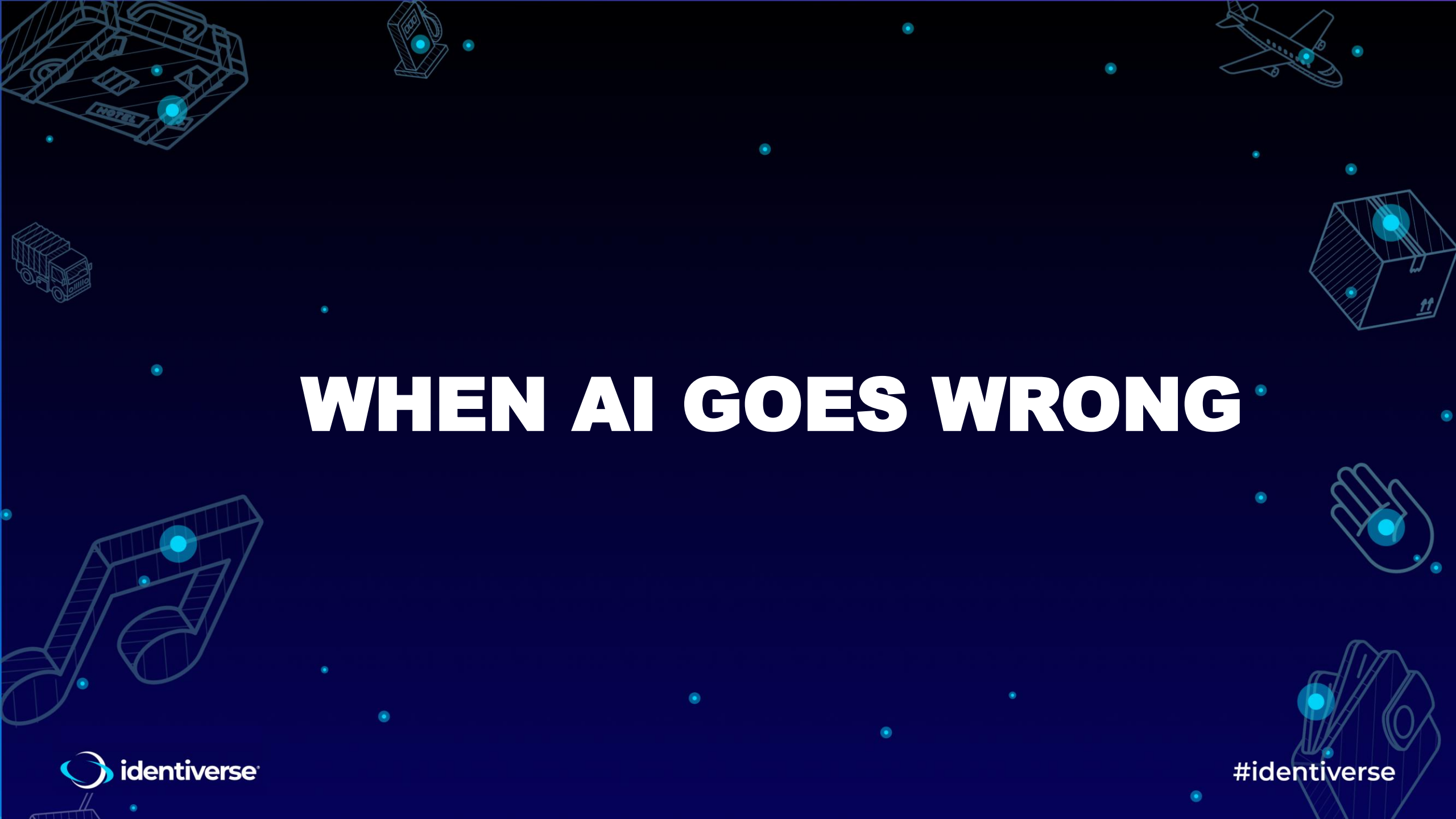
Partner

BNH.AI

Tatiana Rice

Senior Policy Counsel

Future of Privacy Forum



WHEN AI GOES WRONG

Government's Use of Algorithm Serves Up False Fraud Charges

Instagram blames GDPR for failure to tackle rampant self-harm and eating-disorder images

Exclusive: Telegraph investigation found Instagram's algorithms push dangerous content almost two years after it promised to crack down

Laurence Dodds, US TECHNOLOGY REPORTER, SAN FRANCISCO

October 2020 • 9:00pm

Regulators probe racial bias with UnitedHealth algorithm

Regulators says racial bias in algorithm leads to poorer

The New York Times

Opinion

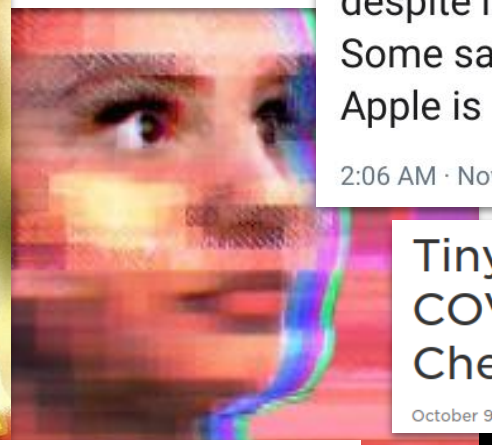
OP-ED CONTRIBUTOR

When a Computer Program Keeps You in Jail

By Rebecca Wexler

Leaving Cert: Why the Government deserves an F for algorithms

Net Results: Invisible code has a significant – and often negative – impact on all our lives



Steve Wozniak
@stevewoz

Replying to @dedwards93 @dhh and @AppleCard

I'm a current Apple employee and founder of the company and the same thing happened to us (10x) despite not having any separate assets or accounts. Some say the blame is on Goldman Sachs but the way Apple is attached, they should share responsibility.

2:06 AM · Nov 10, 2019 · Twitter Web App

A.C.L.U. Accuses Clearview AI of ... 'Nightmare Scenario'

The facial recognition start-up violated the privacy of Illino... residents by collecting their images without their consent, ... liberties group says in a new lawsuit.

Self-Driving Uber Car Kills Ped... in Arizona, Where Robots Roar

Tiny Changes Let False Claims About COVID-19, Voting Evade Facebook Fact Checks

October 9, 2020 · 6:01 AM ET

States Say the Online Bar Exam Was a Success. The Test-Taker Who Peed in His Seat Disagrees

New York, California, and Illinois are among the states reporting that nearly all takers of this week's online bar exam successfully completed the test. But examinees counter that jurisdictions should consider the toll the exam took on them before declaring it a success.

By Karen Sloan | October 07, 2020 at 03:40 PM

Locked Out

Access Denied: Faulty Automated Background Checks Freeze Out Renters

Computer algorithms that scan everything from terror watch lis... eviction records spit out flawed tenant screening reports. And a... ork Tim...

We've Just Seen the First Use of Deepfakes in an Indian Election Campaign

-generated fake videos that are ... iltrating politics.

By Nilesh Christopher

Microsoft's robot editor confuses mixed-race Little Mix singers

Firm's plan to replace editors with AI backfires after wrong image of musician is published

Lawsuit alleges biometric privacy violations from face recognition algorithm training

Paravision's cloud photo storage roots at issue

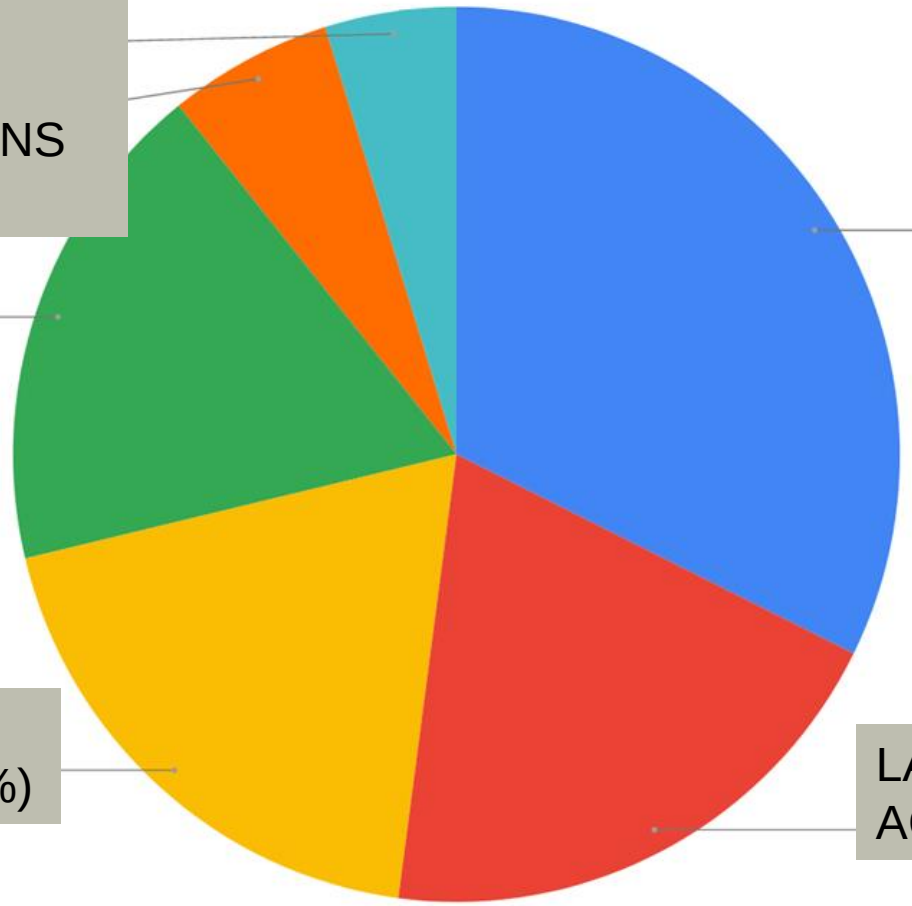
Oct 7, 2020 | Chris Burt

UK passport photo checker shows bias against dark-skinned women

By Maryam Ahmed
BBC News

07 October 2020 | Technology

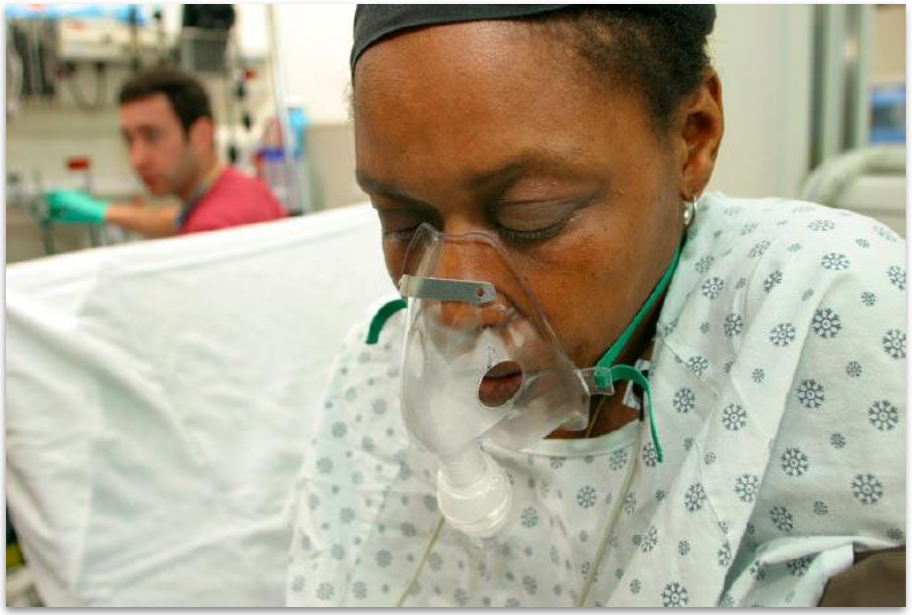
COMMON FAILURE MODES



Harms Are Often Concentrated on Marginalized Groups

- Lack of access
  - Screenout
  - Digital Divide

- Bias and opacity in:
  - Credit scoring and credit products
  - Criminal risk assessment instruments (RIAs)
  - “Ed tech”
  - Employment screening systems
  - Fraud detection systems
  - Healthcare resource allocation
  - “Landlord tech”
  - Predictive policing systems



Millions of black people affected by racial bias in health-care algorithms

- Nature, 24 Oct. 2019

Fairness: Discrimination Testing

Testing for Discrimination:

- Group disparities
  (preferably using tests with legal standing)
- Individual disparities
  - Counterfactuals
  - Track decision boundaries
  - Train adversary models or use special training constraints

Discrimination Mitigation:

- Strategy 1: use NO demographic features in the model and check standard discrimination metrics during model selection.
- Strategy 2:
  - Fix your data.
  - Fix your model.
  - Fix your predictions.


MODEL RISK
MANAGEMENT

Managing Risk

A self-driving Uber test car killed a pedestrian in 2018. In their incident report, the NTSB stated Uber's safety culture was immature and that their software included no consideration for jaywalking pedestrians.

“ [M]odel risk cannot be eliminated, so other tools should be used to manage model risk effectively. Among these are establishing limits on model use, monitoring model performance, adjusting or revising models over time, and supplementing model results. ”

— Federal Reserve Bank
Supervisory Guidance on Model Risk Management

Three Lines of Defense

- **1st Line:** Robust model development.
- **2nd Line:** Rigorous model validation.
- **3rd Line:** Audit, governance, and process controls.

For centuries, militaries have relied on redundancy and multiple lines of defense.


AI AUDITS

What is an Audit?

- **Official tracking of adherence to policy, regulation, or law**
  - Legal standards or regulatory frameworks
- **Conducted by independent third parties**
  - Internal, external
  - Transparent and fair
- **Reporting outcomes based on audience and purpose; accountability**
  - Public; regulators; internal overseers
  - Agreed upon standards
  - Certifications

Measuring Bias

- **More than data and models**
  - Unconscious bias and overt prejudice
  - Initial design choices (“screenout”)
  - Homogenous engineers/perspectives
  - Socio-technical applications
- **Measurement Challenges**
  - Data labeling
  - Demographic descriptions (BISG)
  - Language limitations
  - Benchmarks

Tangible Practices

- **Mathematical definitions of bias**
  - Avoid overly complex, sophisticated or unexplainable thresholds
- **Convert outcomes to a binary or single numeric outcome**
  - Apply traditional measures of practical and statistical significance
  - Select appropriate statistical measures (AIR, *t*-tests, Fisher's exact test)
  - Align traditional metrics to existing laws and regulations

Example: Bias Audit Methodology

- Score each video with FakeFinder models.
- Segment scores by demographic group
  (intersectional groups not considered in this case).
- Establish protected groups:
  East Asian, Black, South Asian, and Women.
- Establish control groups: Whites and Men.
- Test for practical and statistical significance in ***outcome*** differences:
  - **Statistical significance:** *t*-test, significance at $p = 0.05$
  - **Practical significance:** adverse impact ratio (AIR)
- Acceptable threshold: 0.8 – 1.25 (4/5th's rule)
- Ideal threshold: 0.9 – 1.11
- Test for practical significance in ***performance*** differences:
  - **Practical significance:** Accuracy, TP, TN, FP, FN rates
- Acceptable threshold: 0.8 – 1.25 (4/5th's rule)
- Ideal threshold: 0.9 – 1.11

Example results

- **Practical Significance: AIR**

For every 1000 deepfakes detected with White faces, we expect 694 deepfakes with S. Asian faces to be detected.

Demographic Groups	AIR
E. Asian-to-White	1.004
Black-to-White	0.821
S. Asian-to-White	0.694
Female-to-Male	1.035

Example results

- **Statistical Significance: t-Tests**

True positive scores for White faces are on average 2.53% lower than for S. Asian faces. This difference is significant, but the actual difference is moderately small. Sample size and a narrow standard deviation for S. Asian scores contribute to the statistical significance, but so does the difference in group means.

Demographic Groups	Control Mean	Comparison Mean	Percent Difference	*p*-value
E. Asian-to-White	0.948	0.964	-1.69	3.39E-04
Black-to-White	0.948	0.926	2.32	6.65E-02
S. Asian-to-White	0.948	0.972	-2.53	4.06E-04
Female-to-Male	0.955	0.948	0.73	1.62E-01

Example results

- **Performance and Error Rates**

E. Asian faces experience 644% of the false positive rate that White faces experience.

Demographic Groups	Acc. Ratio	TPR Ratio	FPR Ratio	TNR Ratio	FNR Ratio
E. Asian-to-White	1.005	1.012	6.438	0.973	0.394
Black-to-White	0.969	0.951	0.000	1.005	3.488
S. Asian-to-White	1.017	1.020	0.000	1.005	0.000
Female-to-Male	0.988	0.987	#DIV/0!	0.992	2.276

Example: Audit Conclusions

- **Do deep fake (detectors) discriminate?** Yes, of course they do, like nearly all other socio-technical AI systems.
- **Bias tests indicate disparity in both outcomes and performance.** Performance ratios point to problems in erroneous decisions. (High-confidence erroneous decisions are a common failure of neural networks.)
- **Biased and wrong deep fake detection could have serious consequences.** Bias causes wrong decisions and allows for adversarial exploitation.
- **Remediation via technical or process means is essential.**
- **Analysis via causal or explainable AI (XAI) methods is required to understand drivers of bias.**

LAWS AND REGULATIONS

LEGAL ROADMAP

Regional Jurisdictions:

•**U.S. Federal:** FTC, unless sector-specific.

•**U.S. State and Local:** Host of new developments arising (facial recognition bans, biometric laws, general privacy legislation, and sector-specific regulation).

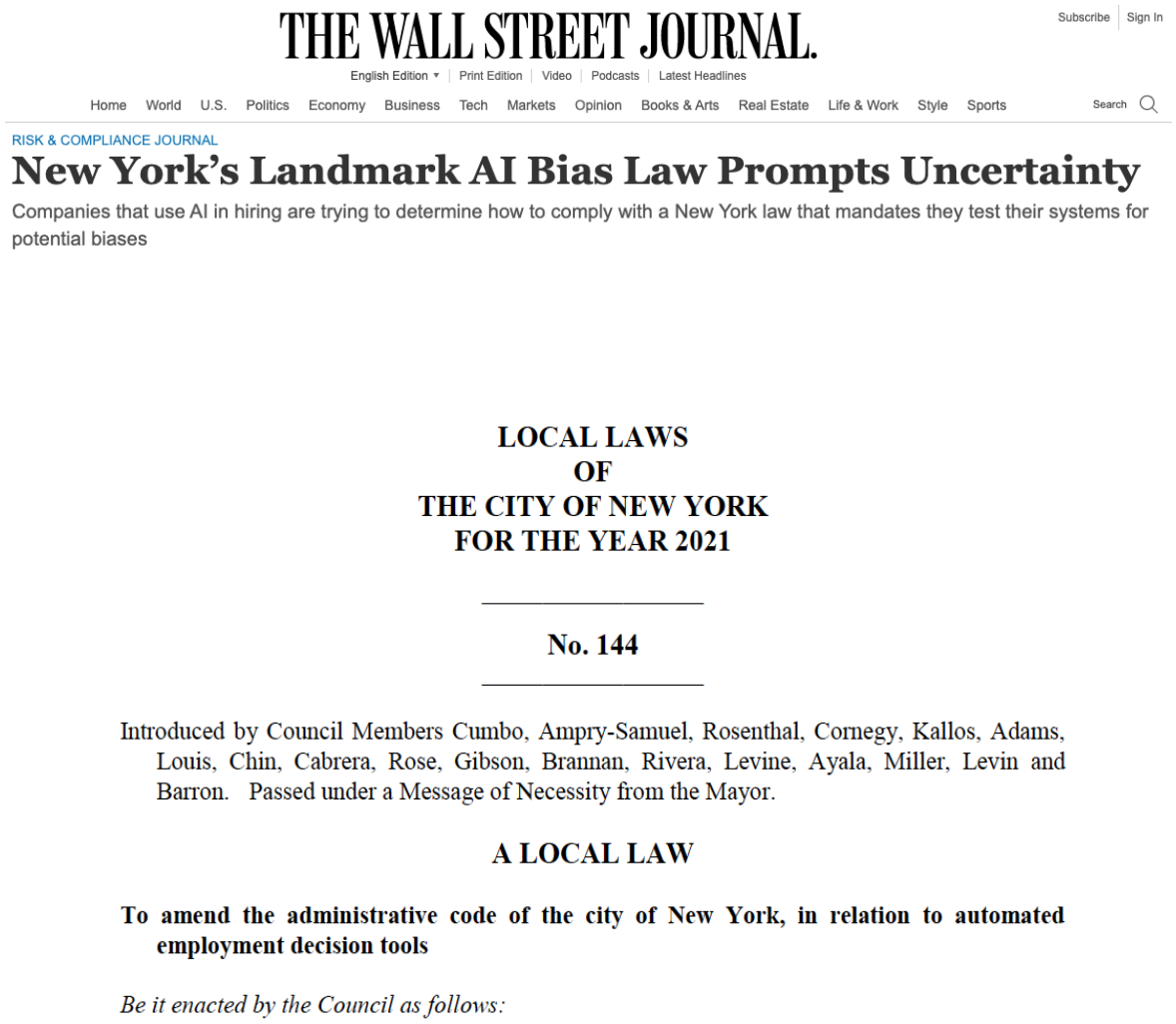
•**International:** E.U. AI Act.; China, Singapore, Korea, Brazil, others

Vertical-specific Regimes:

- **Employment:** Title VII, EEOC.
- **Consumer Finance:** ECOA, FCRA, SR 11-7 and "effective challenge."
- **OSHA:** Guidelines for robotics safety and "hazard analysis."
- **NHTSA:** Six levels of autonomy in vehicles, "Safety Assessment Letters," and federal exemptions.
- **FDA:** Systemic approval processes for "Software as a Medical Device."

NYC Local Law 144

- **Requirements:** Companies using an automated decision tool in their employment process provide notice, audit the system, and publish results
  - Employers or employment agencies who use such tools must use an external auditor
  - Output tables are provided, in alignment with existing EEOC practices
- **Areas of Contention:** Definitions, required data, relationship between vendor providers and complying employers
- **Milestones:** First formal “AI Audit” law



THE WALL STREET JOURNAL.

Subscribe | Sign In

English Edition | Print Edition | Video | Podcasts | Latest Headlines

Home World U.S. Politics Economy Business Tech Markets Opinion Books & Arts Real Estate Life & Work Style Sports Search

RISK & COMPLIANCE JOURNAL

New York’s Landmark AI Bias Law Prompts Uncertainty

Companies that use AI in hiring are trying to determine how to comply with a New York law that mandates they test their systems for potential biases

LOCAL LAWS
OF
THE CITY OF NEW YORK
FOR THE YEAR 2021

No. 144

Introduced by Council Members Cumbo, Ampry-Samuel, Rosenthal, Cornegy, Kallos, Adams, Louis, Chin, Cabrera, Rose, Gibson, Brannan, Rivera, Levine, Ayala, Miller, Levin and Barron. Passed under a Message of Necessity from the Mayor.

A LOCAL LAW

To amend the administrative code of the city of New York, in relation to automated employment decision tools

Be it enacted by the Council as follows:

IL Biometric Information Privacy Act (BIPA)

- **Requirements:** Companies using ”biometric” technologies must:
  - Obtain informed *written* consent
  - Provide retention/destruction schedule
  - No sharing without consent; Full prohibition on sale
- **Areas of Contention:** “Biometric identifier,” consent, relationship between vendor providers and deploying entities
- **Milestones:** First consumer privacy PRA

White Castle could face multibillion-dollar judgment in Illinois privacy lawsuit

(740 ILCS 14/5)
Sec. 5. Legislative findings; intent. The General Assembly finds all of the following:

(a) The use of biometrics is growing in the business and security screening sectors and appears to promise streamlined financial transactions and security screenings.

(b) Major national corporations have selected the City of Chicago and other locations in this State as pilot testing sites for new applications of biometric-facilitated financial transactions, including finger-scan technologies at grocery stores, gas stations, and school cafeterias.

(c) Biometrics are unlike other unique identifiers that are used to access finances or other sensitive information. For example, social security numbers, when compromised, can be changed. Biometrics, however, are biologically unique to the individual; therefore, once compromised, the individual has no recourse, is at heightened risk for identity theft, and is likely to withdraw from biometric-facilitated transactions.

(d) An overwhelming majority of members of the public are weary of the use of biometrics when such information is tied to finances and other personal information.

(e) Despite limited State law regulating the collection, use, safeguarding, and storage of biometrics, many members of the public are deterred from partaking in biometric identifier-facilitated transactions.

(f) The full ramifications of biometric technology are not fully known.

(g) The public welfare, security, and safety will be served by regulating the collection, use, safeguarding, handling, storage, retention, and destruction of biometric identifiers and information.
(Source: P.A. 95-994, eff. 10-3-08.)

identiverse

rse

Federal Trade Commission (FTC)

“Hold yourself accountable – or be ready for the FTC to do it for you.”

- **Section 5:** Investigate and enforce “unfair and deceptive” trade practices.
- **AI-Related Guidance**
  - Watch for discriminatory outcomes
  - Embrace transparency; Provide understandable disclosures about how data is used
  - Keep your AI claims in check
    - Don’t exaggerate capability; Have evidence to support performance claims
  - Generative AI:
    - Make it clear to consumers whether content is “real” and reflects a commercial relationship with a known person or entity, or the product of AI

Federal Trade Commission (FTC)

- **Biometric Policy Statement:**
  - Deceptive practices→ Marketing false or unsubstantiated claims relating to the validity, reliability, accuracy, performance, or lack of bias of a technology
    - Claims of accuracy are deceptive if only true for certain populations and such limitations are not clearly stated
    - Claims of accuracy are deceptive if tests or audit results do not replicate real-world conditions or use by intended users
  - Unfair practices→
    - Surreptitious surveillance without consumer awareness or ability to avoid
    - Failing to assess foreseeable harms to consumers prior to deployment (incl. whether "human-in-the-loop is sufficient to mitigate risk)
    - Failing to address known risks through organizational controls
    - Failing to evaluate practices and capabilities of third parties (compliance, oversight mechanisms)
    - Failing to provide appropriate training for employees and contractors (security, management of data)
    - Failing to conduct ongoing monitoring

More Federal and State Activity

Schumer Launches Major Effort To Get Ahead Of Artificial Intelligence

New York, N.Y. – *Today, Senate Majority Leader Chuck Schumer (D-NY) launched a first-of-its-kind effort to advance and manage one of the fastest moving, and most consequential industries across the globe: artificial intelligence (AI). Time is of the essence to ensure this powerful new technology and it's potentially wide-ranging impact on society is put to proper*

NIST National Institute of Standards and Technology U.S. Department of Commerce

Just Released! NIST Identity & Access Management Roadmap

National Institute of Standards and Technology (NIST) sent this bulletin at 04/21/2023 12:51 PM EDT

NIST View As Web Page

NIST Cybersecurity and Privacy Program

What is the EU AI Act?

The AI Act is a proposed European law on artificial intelligence (AI) – the first law on AI by a major regulator anywhere. The law assigns applications of AI to three risk categories. First, applications and systems that create an **unacceptable risk**, such as government-run social scoring of the type used in China, are banned. Second, **high-risk applications**, such as a CV-scanning tool that ranks job applicants, are subject to specific legal requirements. Lastly, applications not explicitly banned or listed as high-risk are largely left unregulated.

NTIA.gov

Artificial Intelligence

Request for Comment

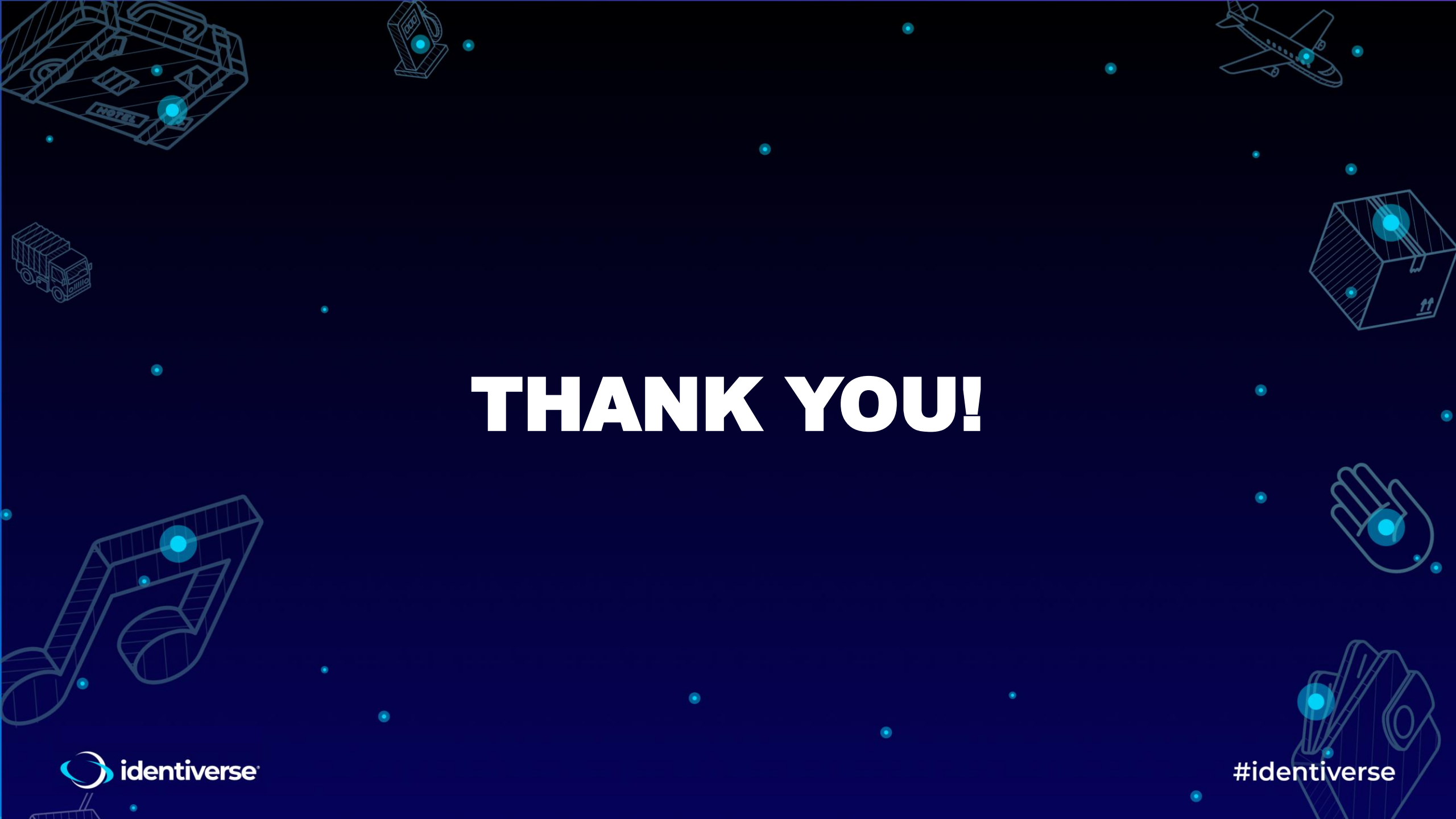
Announcement Event

NTIA's "AI Accountability Policy Request for Comment" seeks feedback on what policies can support the development of AI audits, assessments, certifications and other mechanisms to create earned trust in AI systems. Much as financial audits create trust in the accuracy of a business' financial statements, so for AI, such mechanisms can help provide assurance that an AI system is trustworthy. Just as financial accountability required policy and governance to develop, so too will AI system accountability.

Review and Summary

Model Audits

- **Audits may have different meanings in the context of AI**
  - Internal audits (MRM); Financial audit principles applied to AI; Model/Algorithmic Evaluation
- **To what standard do we audit?**
  - Legal and industry standards
  - Who audits the auditors?
- **Audits are being incorporated** into regulatory and legal oversight systems


THANK YOU!
identiverse
#identiverse