Prepared for Philadelphia Federal Reserve

Fourth Workshop on Payments, Lending, and Innovations in Consumer Finance

Do Deepfakes Discriminate?

Auditing a Deepfake Detection System for Systemic Bias

Patrick Hall, Principal Scientist, BNH.AI
Andrew Burt, Managing Partner, BNH.AI

Based on public reference work from In-Q-Tel Labs:
https://www.iqt.org/ai-assurance-do-deepfakes-discriminate/

BNH.AI

About BNH.AI

We are the **first and only** law firm jointly run by legal and data science personnel to help organizations protect and advance their data, analytics and artificial intelligence investments.

BNH
.AI

Audit: Liability evaluations and model audits enable companies to bet big on AI while understanding its risks.

Address: Guidance and services assist in risk management and deliver documentation that attests to appropriate mitigation.

Automate: Custom software automates routine risk management tasks.

Services and documentation are privileged to the extent feasible.

Background: Bias in AI

Examples of Group bias:

- Overt bias against groups, like **disparate treatment**.
- Unintentional bias against groups, like **disparate impact**.
- **Differential validity**: When a model is less accurate for certain groups of people.

Local or individual bias:

When similar individuals are treated differently because of demographic group membership.

Digital Divide:

Many still cannot even access the internet properly, much less AI-based services.

Screenout:

When employment systems discriminate against those with disabilities.

Reality is even more daunting →

Image source: NIST SP1270 - https://nvlpubs.nist.gov/nistpubs/SpecialPublications/NIST.SP.1270.pdf

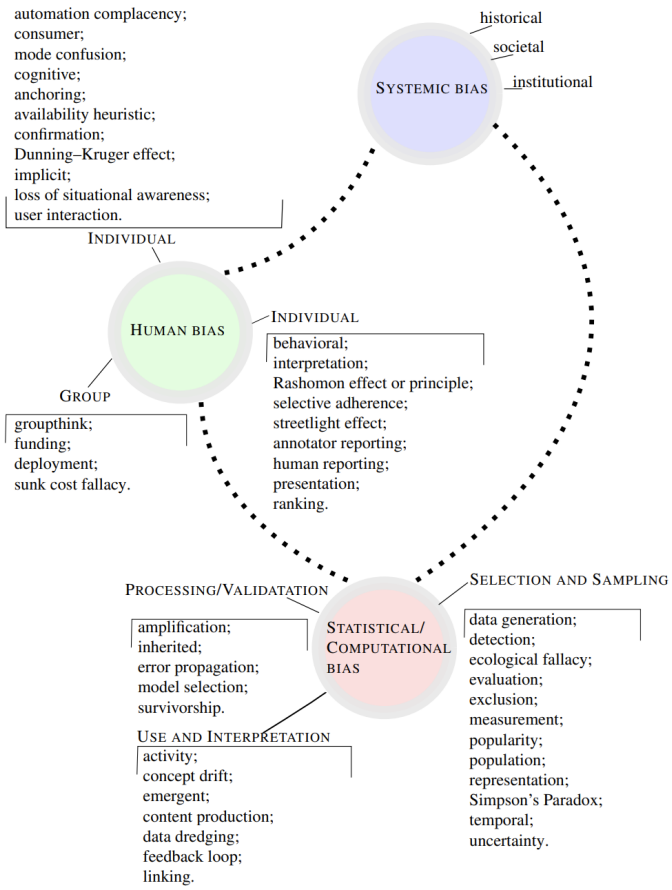


Fig. 2. Categories of AI Bias. The leaf node terms in each subcategory in the picture are hyperlinked to the GLOSSARY. Clicking them will bring up the definition in the Glossary. To return, click on the current page number (8) printed right after the glossary definition.

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

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

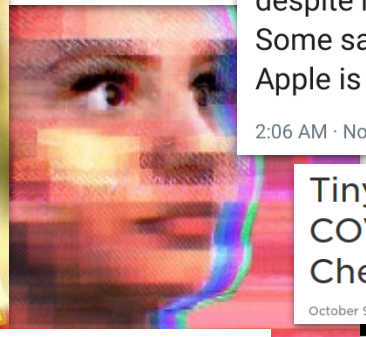
By Rebecca Wexler

Laurence Dodds, US TECHNOLOGY REPORTER, SAN FRANCISCO

October 2020 • 9:00pm

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...

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

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

By Nilesh Christopher

Lawsuit alleges biometric privacy violations from face recognition algorithm training

Paravision's cloud photo storage roots at issue

Oct 7, 2020 | Chris Burt

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

UK passport photo checker shows bias against dark-skinned women

By Maryam Ahmed
BBC News

07 October 2020 | Technology

Model Audit Background

Audit: Official exercise, tracking adherence to some policy, regulation, or law; conducted by independent parties.

Audit may have several different meanings in the context of AI, including **Internal Audit** (e.g., Model Risk Management), **Financial Audit** principles applied to AI (e.g., ForHumanity), or **Model/Algorithmic Audit** (e.g., Inioluwa Deborah Raji et al., “Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing”).

To what standard do we audit? Who audits the auditors? Without clear standards and ethical guidelines, audits often devolve into tech-washing and marketing exercises.

Despite flaws, audits are being incorporated into laws, e.g., the recent NYC bias audit requirement for AI systems used in hiring (§20-871(a)(1) of Subchapter 25 of Chapter 5 of Title 20 of the New York City Administrative Code).

BNH.AI

FakeFinder Audit

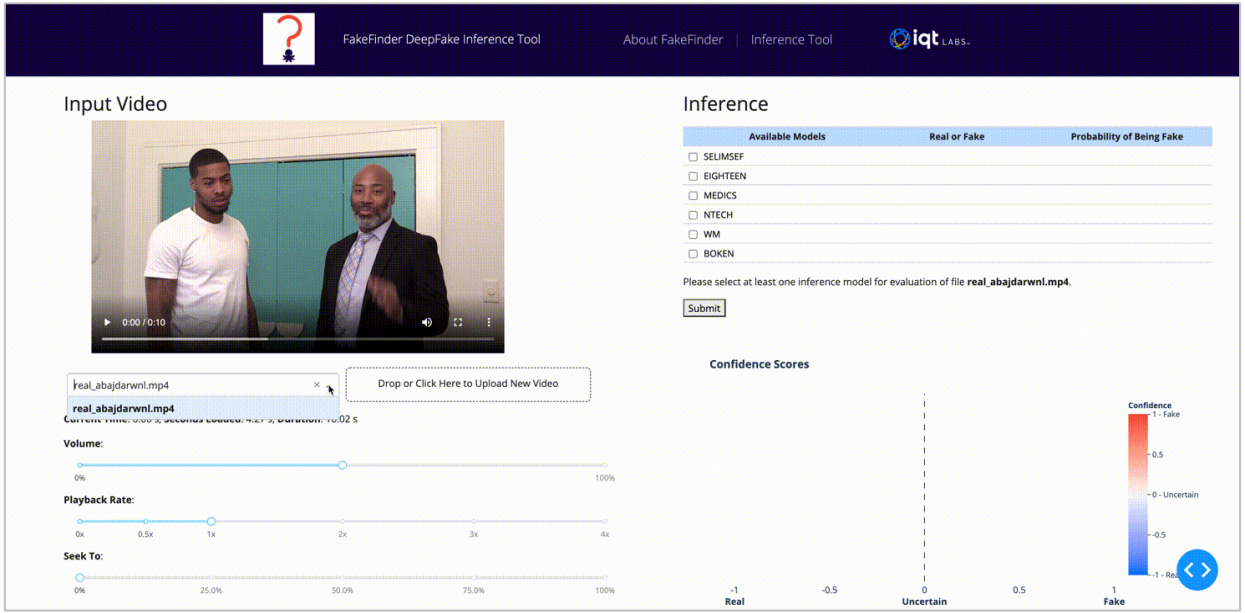
Why: What if an IC deep fake detector works better for Biden than for Obama?

Objective: Evaluation of systemic bias in alignment with existing legal guidance

Model: FakeFinder deep fake detector by In-Q-Tel Labs

Standard: AI Ethics Framework for the Intelligence Community

Code of Ethics: Washington D.C. Bar

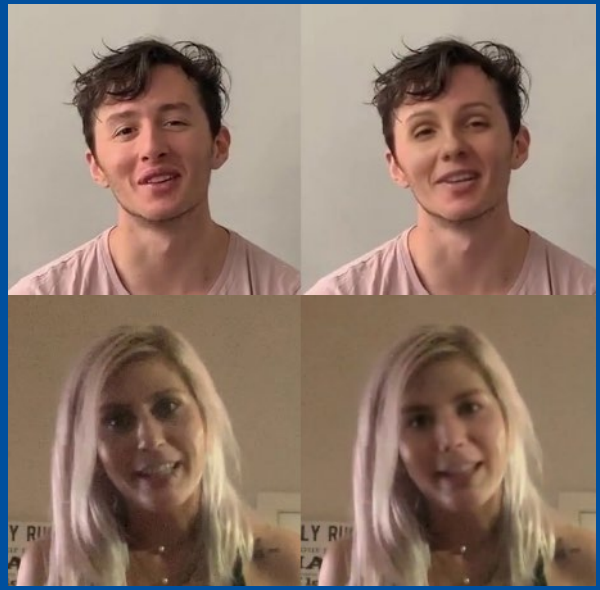


Short animation of FakeFinder user interface in which a real video and then a deep fake are analyzed - https://github.com/IQTLabs/FakeFinder.

BNH.AI

DFDC Data Background

- The Deep Fake Detection Challenge (DFDC) preview dataset was selected for the audit.
- Released by Facebook Research, in preview, and later, full formats.
- Preview set contains roughly 5,000 videos, with 28% “cheap fake”/face swap videos.
- Videos contain people of various races and genders.
- Demographic markers were not included in the preview data, and were assigned post-hoc and manually.*
- Do you assign demographic markers to the original face or new body?! Do you bias-test the faces or the bodies?!



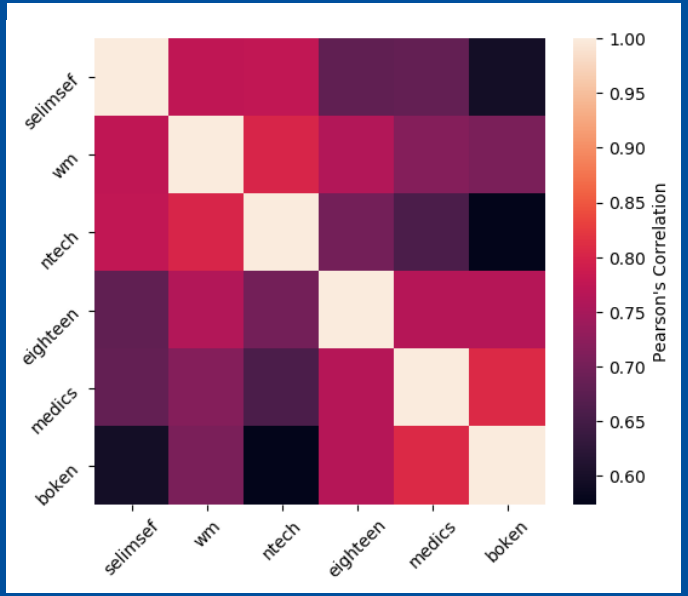
Example snapshots from the DFDC preview data, with real images on the left and deepfakes on the right - https://arxiv.org/pdf/1910.08854.pdf.

BNH.AI

*Not thinking though bias management from the beginning of an ML project is a common driver of incidents.

FakeFinder Background

- Comprised of 6 CNN-based deep fake classifiers, drawn from DFDC and DeepForensics challenge entrants.
- Individual classifiers score between 0.6 and 1.0 accuracy (@0.5) across various competition datasets.
- Each classifier is available within a container with trained weights from GitHub.
- Deep fake detection scores generated via simple Python script or basic GUI (slide 8).



Correlations between the predictions of the six constituent deep fake detector classifiers that make up FakeFinder - https://github.com/IQTLabs/FakeFinder.

BNH.AI

Bias Audit Methodology

1. Score each video with FakeFinder models.
2. Segment scores by demographic group
   (intersectional groups not considered in this case).
3. Establish so-called protected groups:
   East Asian, Black, South Asian, and Women.
4. Establish control groups: Whites and Men.
5. 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
6. Test for practical significance in *performance* differences:

   Practical significance: Accuracy, TP, TN, FP, FN rate protected-to-control ratios.

   - Acceptable threshold: 0.8 – 1.25 (4/5th's rule)
   - Ideal threshold: 0.9 – 1.11

$$\text{AIR} \equiv \frac{P(\hat{y} = 1 \mid X_p = 1)}{P(\hat{y} = 1 \mid X_c = 1)}$$

AIR is defined as the ratio of the rate of positive outcomes for a protected group divided by the rate of positive outcomes for an associated control group.

$$\text{Acc. Ratio} \equiv \frac{\text{Acc}_p}{\text{Acc}_c}$$

Performance ratios divide some measure of error or performance quality (e.g., Acc., TPR, TNR, FPR, FNR) for a protected group by the same quantity for a control group.

BNH.AI

This process works well for many binary outcomes, across input data formats or functional forms of models.

Example Results: Practical Significance - AIR

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

Remember the political deep fake in slide 4?

BNH.AI

Example Results: Statistical Significance - t-tests

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 interpretation:
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.

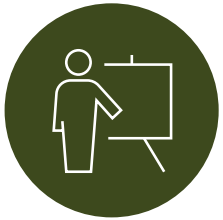
BNH.AI

Example Results: Performance and Error Ratios

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

BNH.AI



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 pratfall with neural networks.)
- Biased and wrong deep fake detection in the IC context could have serious consequences. Bias causes wrong decisions and allows for adversarial exploitation.
- Remediation via technical or process means is prudent.
- Analysis via causal or explainable AI (XAI) methods is required to understand drivers of bias.

BNH.AI

Remediating Bias: Practical Advice

Pre-, in-, post-processing, and model selection, but …

There is much more to bias than datasets and algorithms

- Bias is managed most successfully in a specific operational context.
- Apply the scientific method and experimental designs to AI systems.
- Apply human-centered design to AI systems.
- Setup governance structures for the ***people*** that build and maintain AI systems.

NIST Special Publication 1270

Towards a Standard for Identifying and Managing Bias in Artificial Intelligence

Reva Schwartz
National Institute of Standards and Technology
Information Technology Laboratory

Apostol Vassilev
National Institute of Standards and Technology
Information Technology Laboratory
Computer Security Division

Kristen Greene
National Institute of Standards and Technology
Information Technology Laboratory
Information Access Division

Lori Perine
National Institute of Standards and Technology
Information Technology Laboratory
& The University of Maryland

Andrew Burt
Patrick Hall
BNH.AI

This publication is available free of charge from:
https://doi.org/10.6028/NIST.SP.1270

March 2022

U.S. Department of Commerce
Gina M. Raimondo, Secretary

National Institute of Standards and Technology
James K. Olthoff, Performing the Non-Exclusive Functions and Duties of the Under Secretary of Commerce for Standards and Technology & Director, National Institute of Standards and Technology

BNH.AI

Comprehensive AI Risk Management

Discussing fairness or other trustworthy AI system characteristics apart from one another is somewhat impractical (e.g., performance quality, reliability, robustness, security, privacy, safety, transparency, accountability, etc.).

> When managing risks in AI systems it is important to understand that the attributes of the AI RMF risk taxonomy are interrelated. Highly secure but unfair systems, accurate but opaque and uninterpretable systems, and inaccurate, but fair, secure, privacy-protected, and transparent systems are all undesirable. It is possible for trustworthy AI systems to achieve a high degree of risk control while retaining a high level of performance quality. Achieving this difficult goal requires a comprehensive approach to risk management, with tradeoffs among the technical and socio-technical characteristics.

BNH.AI

Excerpted from "NIST AI Risk Management Framework (Draft)."
Source: https://www.nist.gov/system/files/documents/2022/03/17/AI-RMF-1stdraft.pdf.

Practical Takeaways

Collect demographic data beforehand, can be inferred from name and ZIP code

Get started with simple established tests, with known thresholds

Remember that bias testing is only one part of managing bias

Open Source Bias Testing and Remediation Tools

Aequitas: https://github.com/dssg/aequitas

AI fairness 360: https://github.com/Trusted-AI/AIF360

Fairmodels: https://github.com/ModelOriented/fairmodels (R)

BNH.AI

BNH
.AI

BURT & HALL LLP

Patrick Hall, Principal Scientist, BNH.AI
ph@bnh.ai