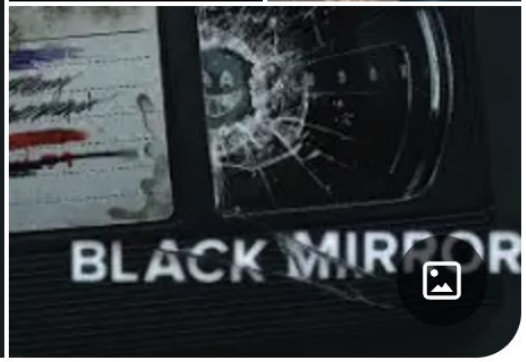
CSCI 5622 Machine Learning

Lecture 13
Sociotechnical issues in AI/ML

Black Mirror

2011 · Sci-fi · 6 seasons

Overview | Watch show | Episode guide | Cast | Reviews



Netflix

Watch Black Mirror | Netflix Official Site

Black Mirror · Joan Is Awful · Loch Henry · Beyond the Sea · Mazey Day · Demon 79 · Striking Vipers · Smithereens · Rachel,…

JUSTICE NEWS

Department of Justice

Office of Public Affairs

FOR IMMEDIATE RELEASE

Thursday, May 12, 2022

Justice Department and EEOC Warn Against Disability Discrimination

Employers' Use of Artificial Intelligence Tools Can Violate the Americans with Disabilities Act

The Department of Justice and the Equal Employment Opportunity Commission (EEOC) today each released a technical assistance document about disability discrimination when employers use artificial intelligence (AI) and other software tools to make employment decisions.

US FDA Proposed Regulatory Framework

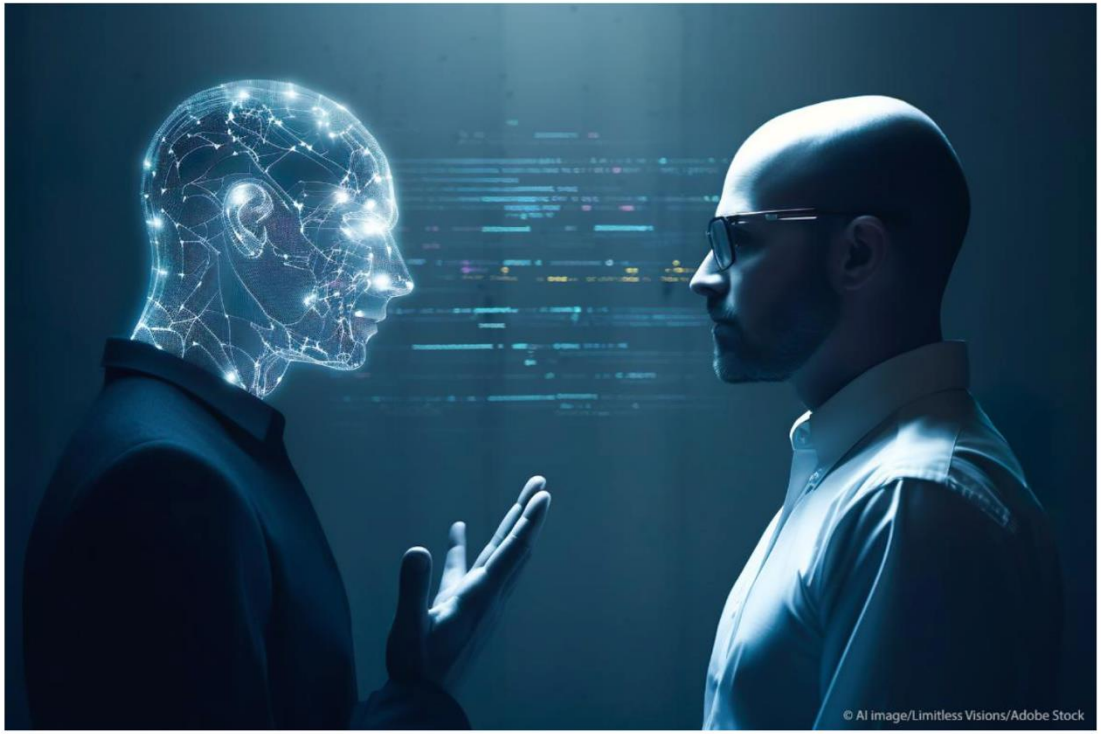
The May 2023 Report by the U.S. Food and Drug Administration (FDA) recognizes the transformative potential of AI in healthcare, but emphasizes the importance to address core issues like human-led governance, data quality, model development standards, accountability, and transparency.

Food and Drug Administration. (2023). Proposed regulatory framework for modifications to artificial intelligence/machine learning (AI/ML)-based software as a medical device (SaMD). Discussion paper and Request for Feedback

EU AI Act: first regulation on artificial intelligence

Society Updated: 14-06-2023 - 14:06
Created: 08-06-2023 - 11:40

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.



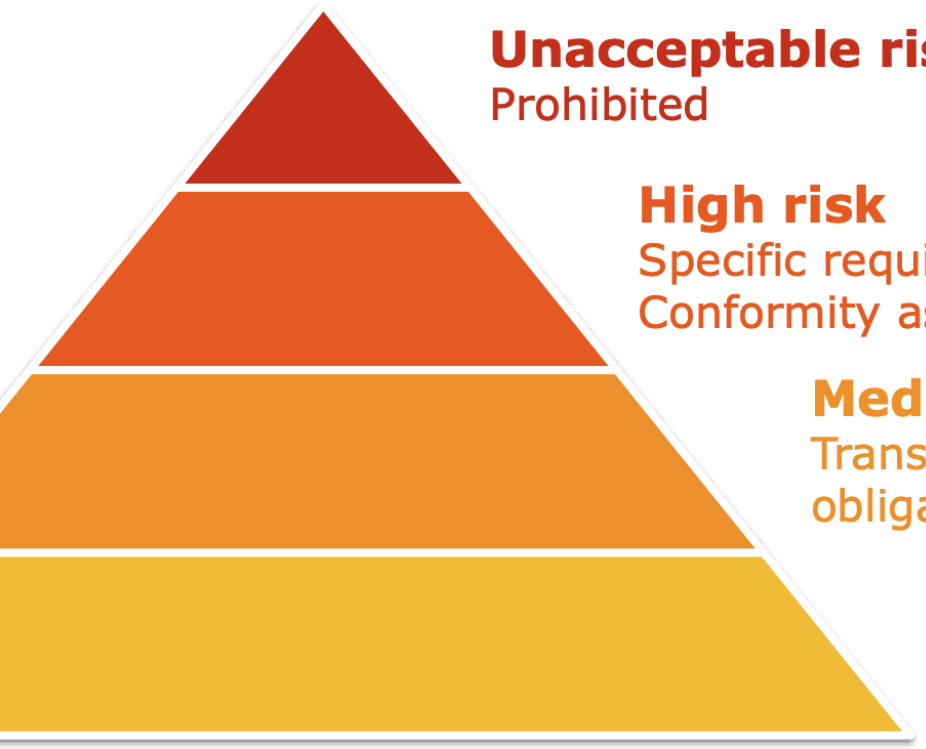
This illustration of artificial intelligence has in fact been generated by AI

EU AI Act

A comprehensive legal framework for AI seeking to ensure that AI systems used in the European Union are safe, transparent, traceable, non-discriminatory and environmentally friendly.

Source:
https://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

EU AI Act



Unacceptable risk
Prohibited

Examples: AI systems deploying deception to distort behavior, social scoring, assessing risk of criminal offense

High risk
Specific requirements & Conformity assessments

Examples: AI systems used as safety components in critical infrastructure, biometric identification, law enforcement, border control management, education

Medium risk
Transparency & Information obligations

Examples: chatbots

Minimal or low risk
Permitted with no restriction

Examples: spam filter

High risk AI providers must:

- establish a **risk management** system
- conduct **data governance** (datasets are **relevant**, sufficiently **representative** and, to the best extent possible, free of errors and complete according to the intended purpose)
- Draw up technical documentation and implement **record-keeping** to enable recording of events
- Design the system to allow **human oversight**
- Design the system to achieve appropriate levels of **accuracy**, **robustness**, and **cybersecurity**

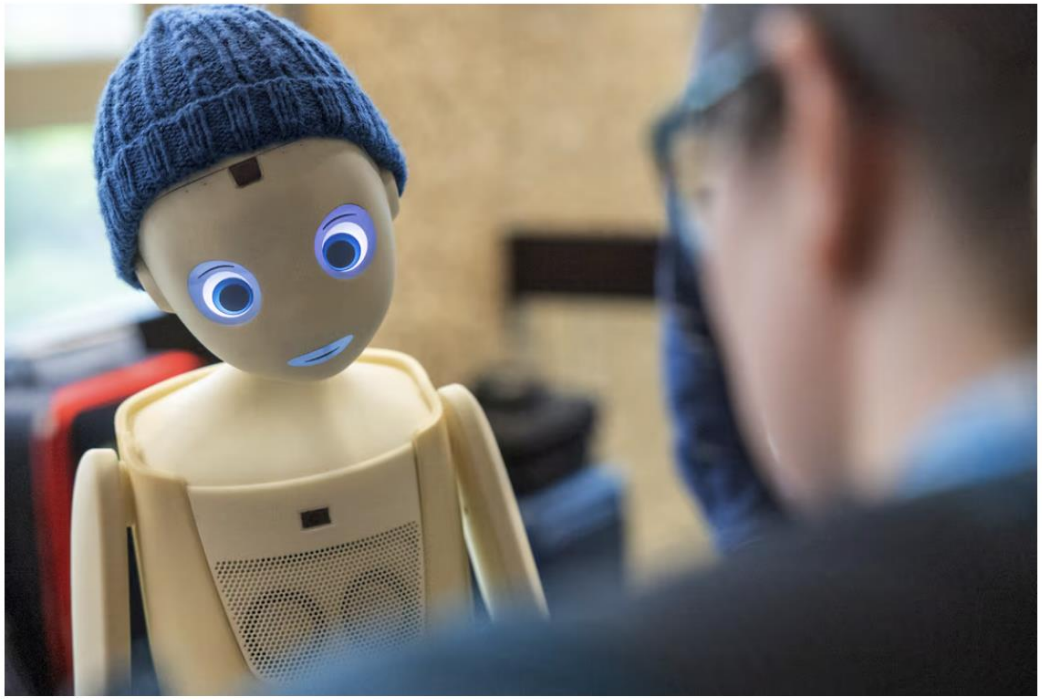
Source: https://artificialintelligenceact.eu/wp-content/uploads/2024/11/Future-of-Life-InstituteAI-Act-overview-30-May-2024.pdf

UN advisory body makes seven recommendations for governing AI

Reuters

By **Supantha Mukherjee**

September 19, 2024 6:44 AM GMT+1 · Updated 6 hours ago



A visitor talks with the Navel robot, by Navel Robotics, during the AI for Good Global summit on artificial intelligence, organised by the International Telecommunication Union (ITU), in Geneva, Switzerland, May 30, 2024. REUTERS/Denis Balibouse/File photo Purchase Licensing Rights

- Global governance
- Global cooperation on establishing common understanding for pursuing common benefits
- Global AI fund to address gaps in capacity and collaboration
- Global AI data framework

United Nations. (2024). Governing AI for Humanity. Report

Why should we, computer scientists, concern ourselves with sociotechnical issues of AI?

How about we allow the society to address these first?

ML algorithms impact our lives more than we might think

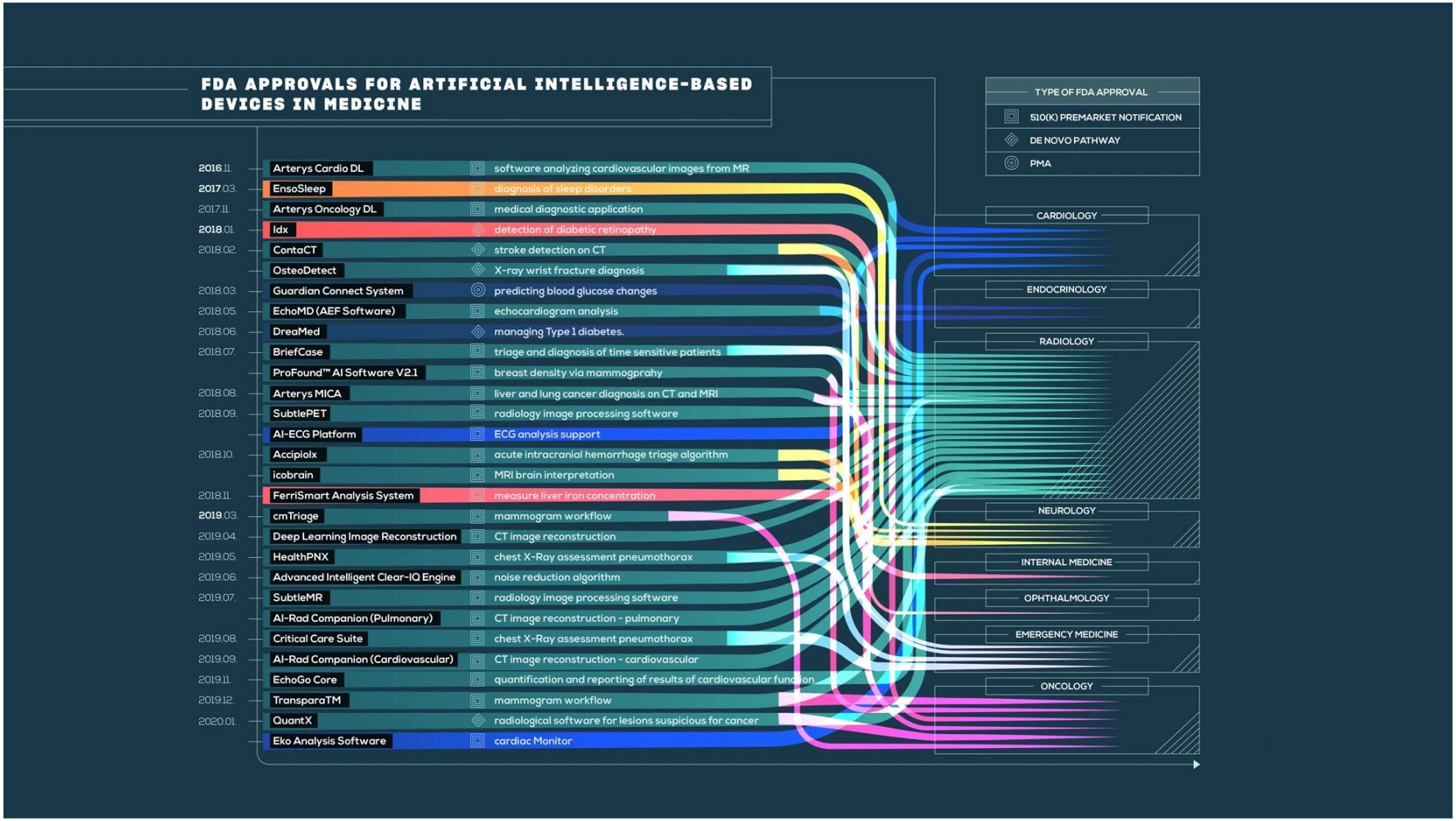
Source: Reuters

ML algorithms impact our lives more than we might think

ML algorithms impact our lives more than we might think

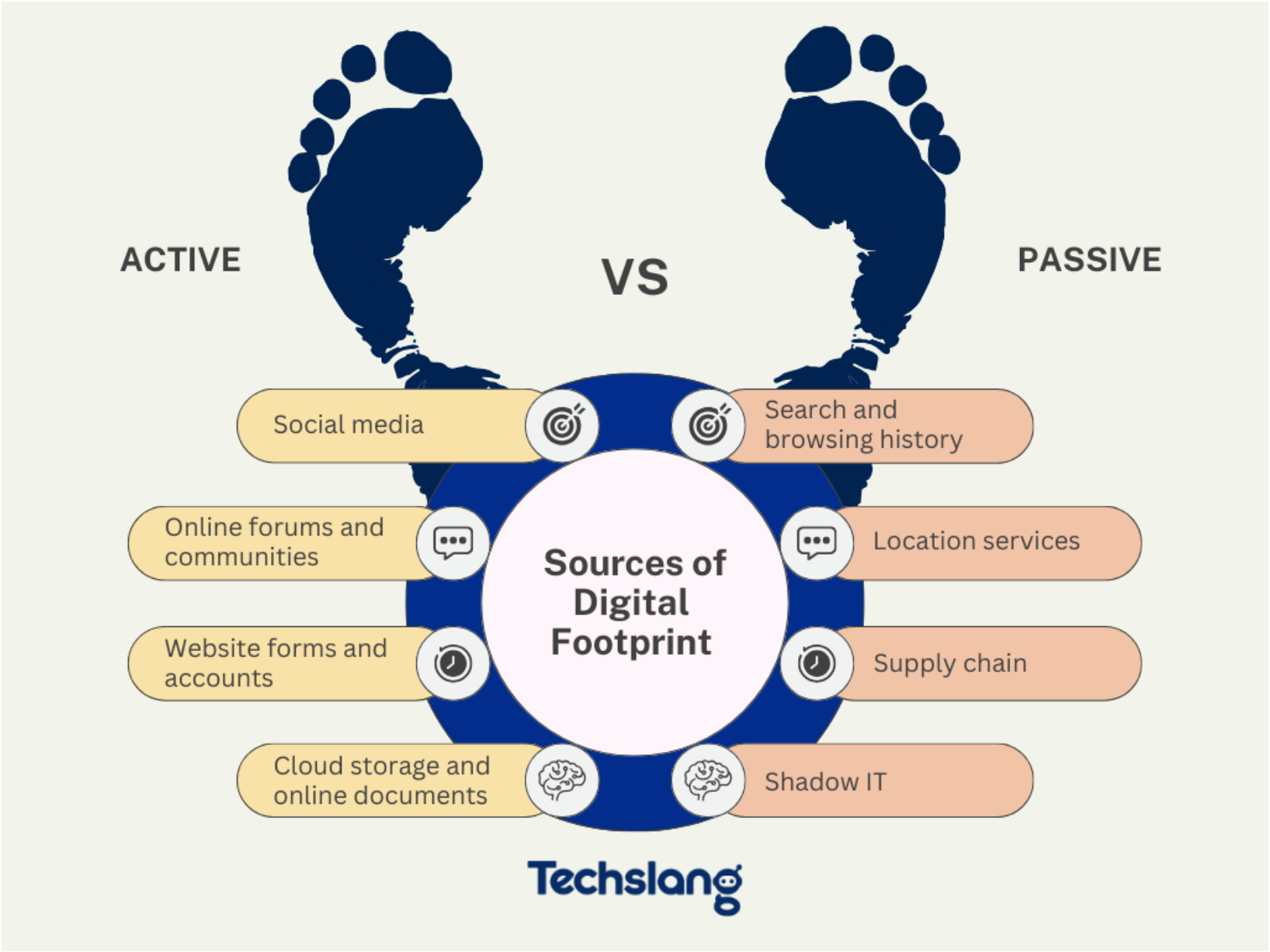
As of August 2024, the U.S. Food and Drug Administration has approved 950 AI/ML-enabled medical devices

Source: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-aiml-enabled-medical-devices



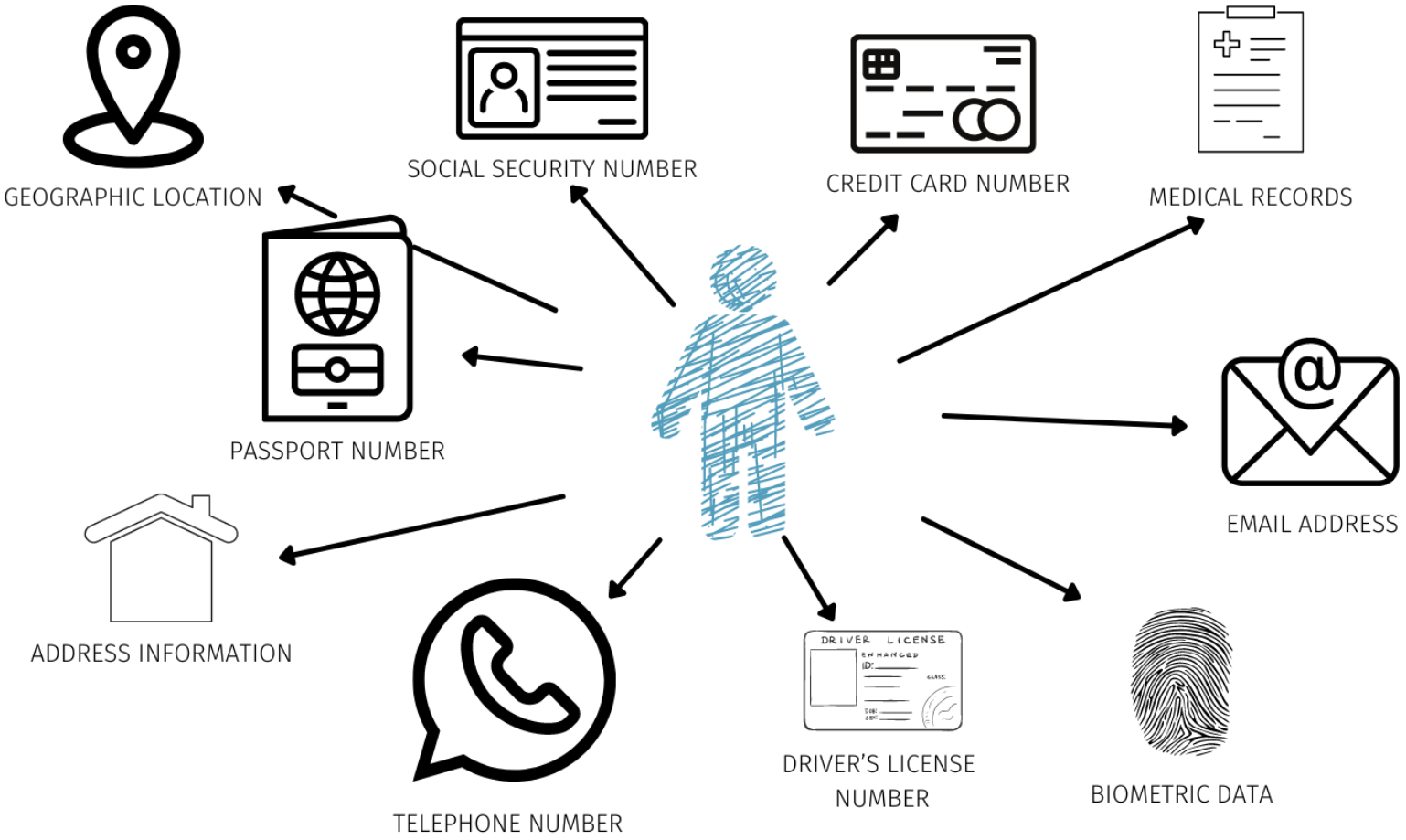
Benjamens, S., Dhunnoo, P., & Meskó, B. (2020). The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ digital medicine, 3(1), 118.

Our data is out there!

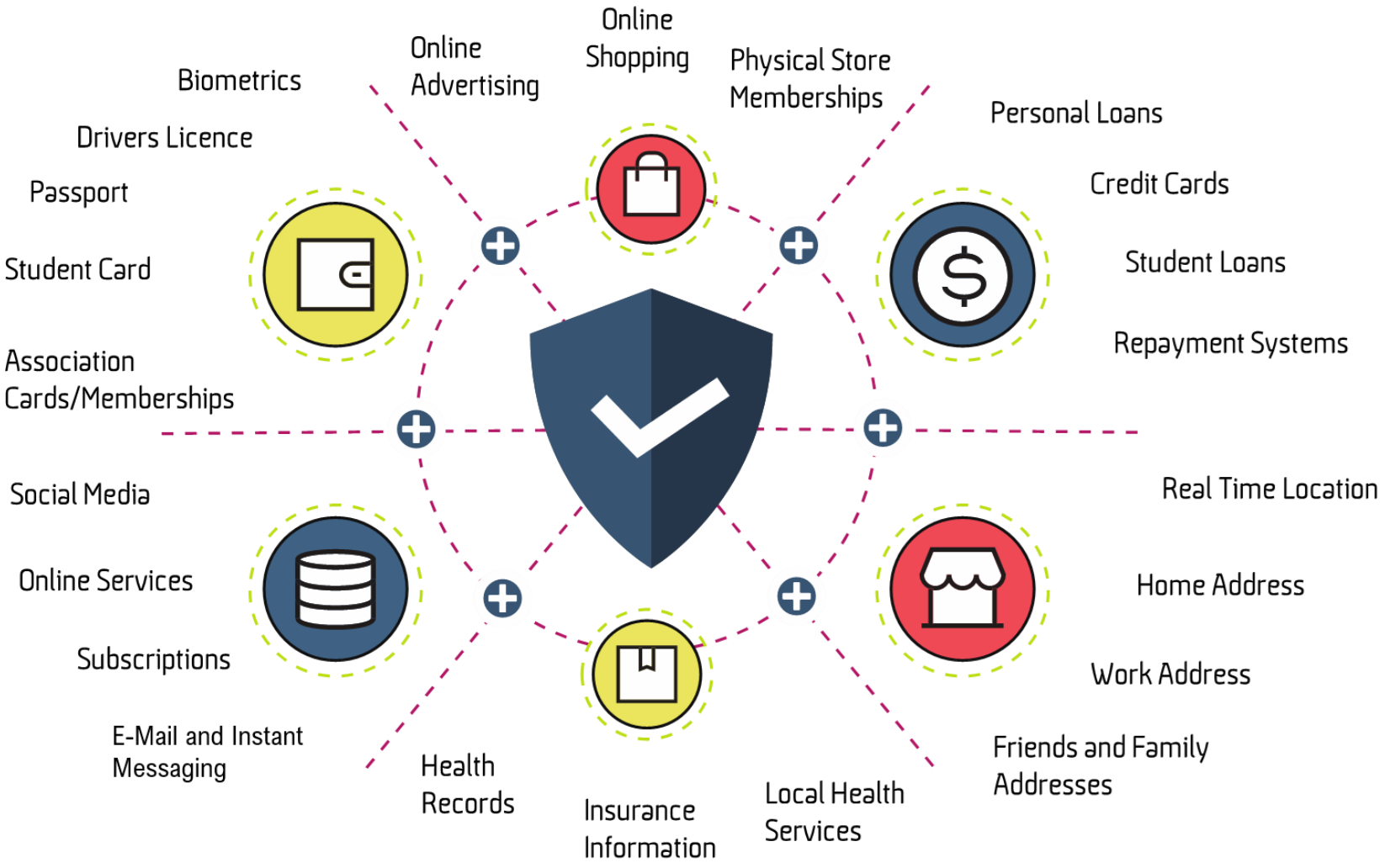


Our data is out there!

Personally identifiable information (PII)



Our data is out there!



Is it possible for machines, with no concept of race, ethnicity, gender or religion, to discriminate against certain groups?

MIT News
ON CAMPUS AND AROUND THE WORLD

Study reveals why AI models that analyze medical images can be biased

These models, which can predict a patient's race, gender, and age, seem to use those traits as shortcuts when making medical diagnoses.

Anne Trafton | MIT News

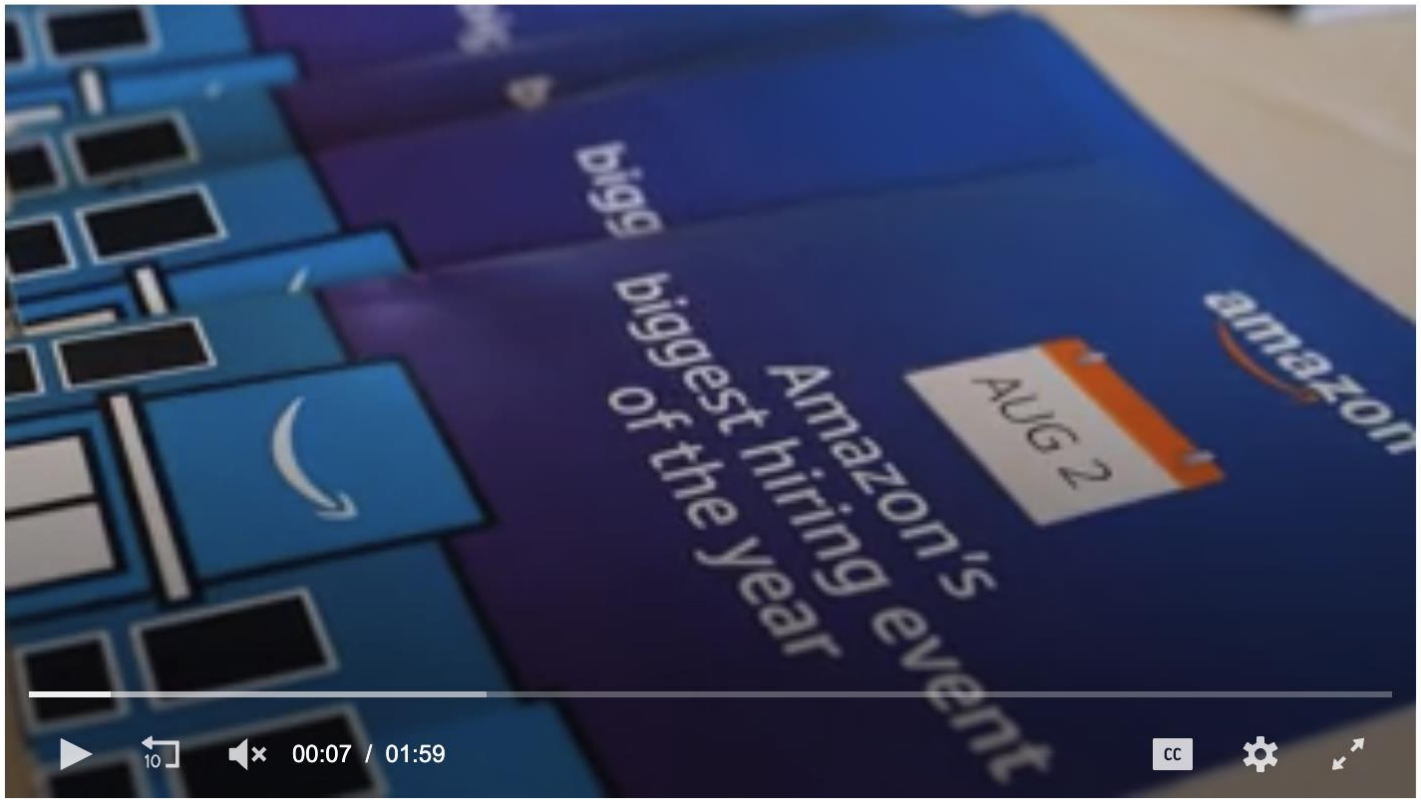
June 28, 2024

Yang, Y., Zhang, H., Gichoya, J. W., Katabi, D., & Ghassemi, M. (2024). The limits of fair medical imaging AI in real-world generalization. Nature Medicine, 1-11.

Insight - Amazon scraps secret AI recruiting tool that showed bias against women

By **Jeffrey Dastin**

October 10, 2018 6:50 PM MDT · Updated 6 years ago



Source: Reuters

TECH

The problem with AI? Study says it's too white and male, calls for more women, minorities

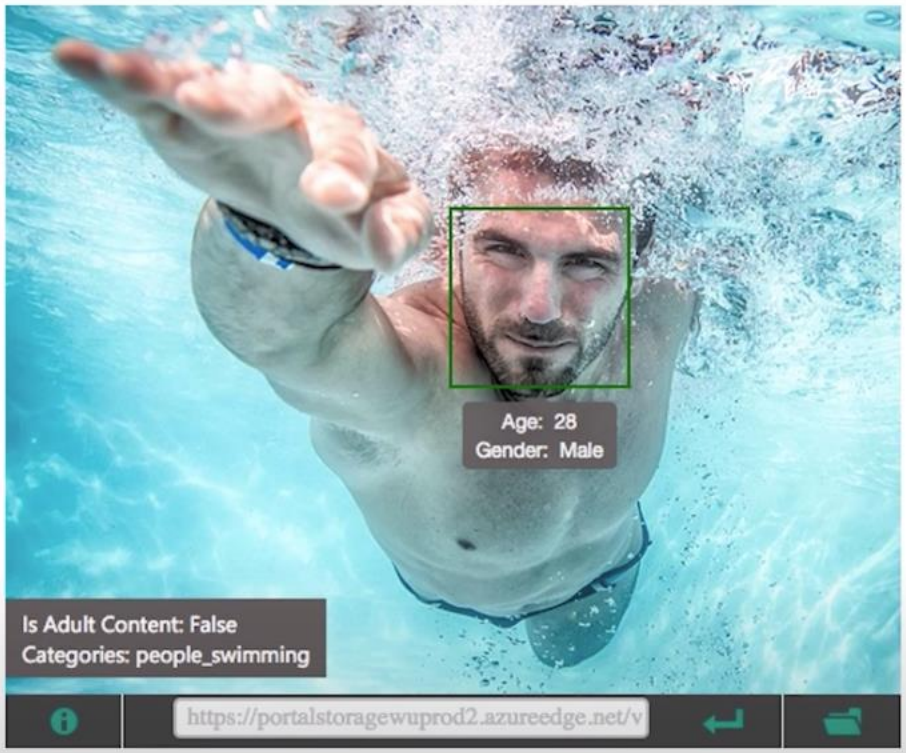
Jessica Guynn USA TODAY

Published 8:00 p.m. ET Apr. 16, 2019 | **Updated 11:37 a.m. ET Apr. 17, 2019**

Two out of four leading face recognition platforms do not reliably detect African American users.

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now. https://www.youtube.com/watch?v=TWWsW1w-BVo&t=45s

Evidence of bias in computer vision



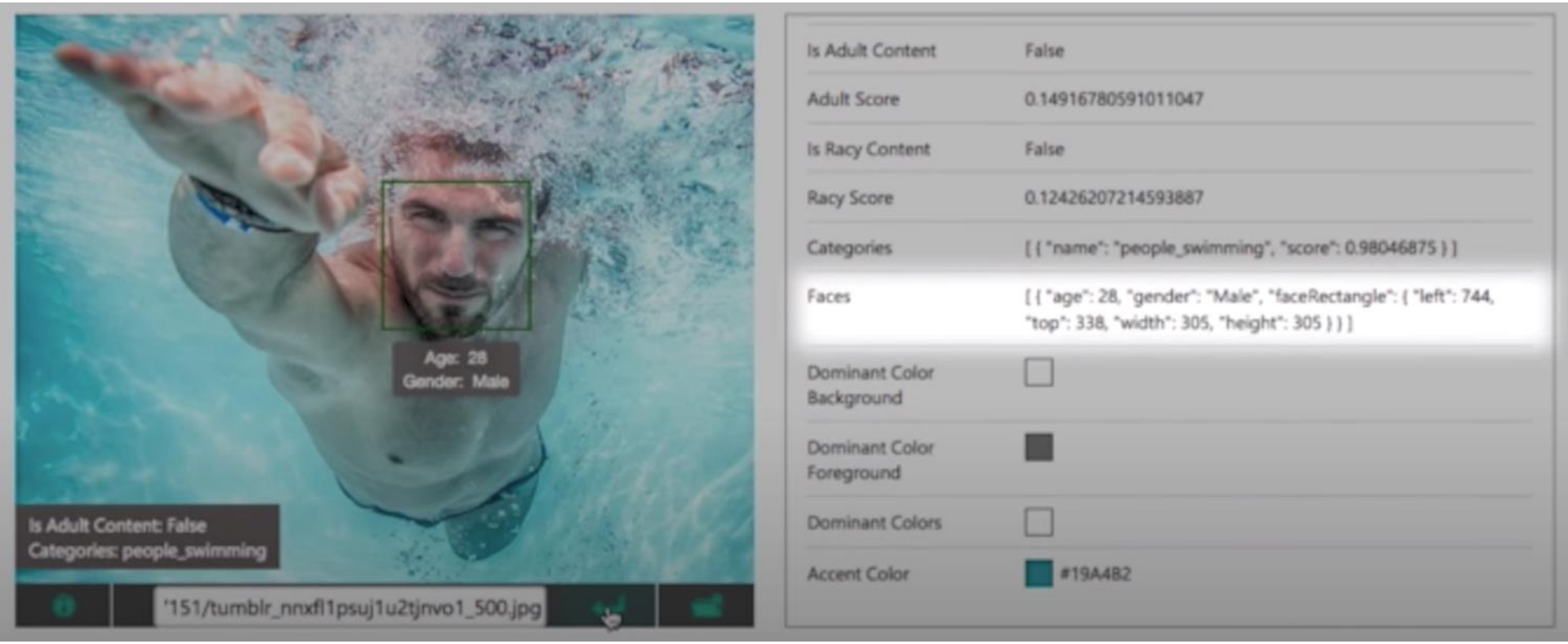
Features:

Feature Name	Value
Description	{ "type": 0, "captions": [{ "text": "a man swimming in a pool of water", "confidence": 0.7850108693093019 }] }
Tags	[{ "name": "water", "confidence": 0.9996442794799805 }, { "name": "sport", "confidence": 0.9504992365837097 }, { "name": "swimming", "confidence": 0.9062818288803101, "hint": "sport" }, { "name": "pool", "confidence": 0.8787588477134705 }, { "name": "water sport", "confidence": 0.631849467754364, "hint": "sport" }]
Image Format	jpeg
Image Dimensions	1500 x 1155
Clip Art Type	0 Non-clipart
Line Drawing Type	0 Non-LineDrawing
Black & White Image	False

Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

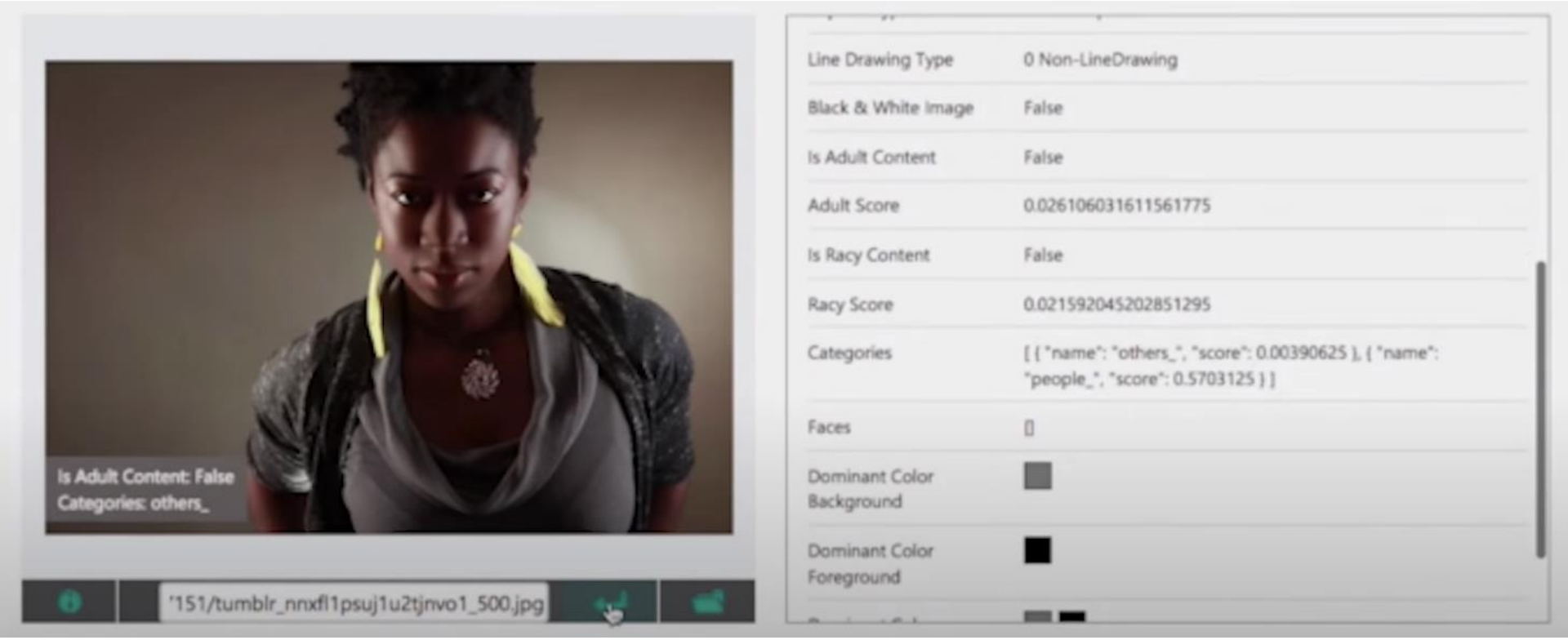
Evidence of bias in computer vision



Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

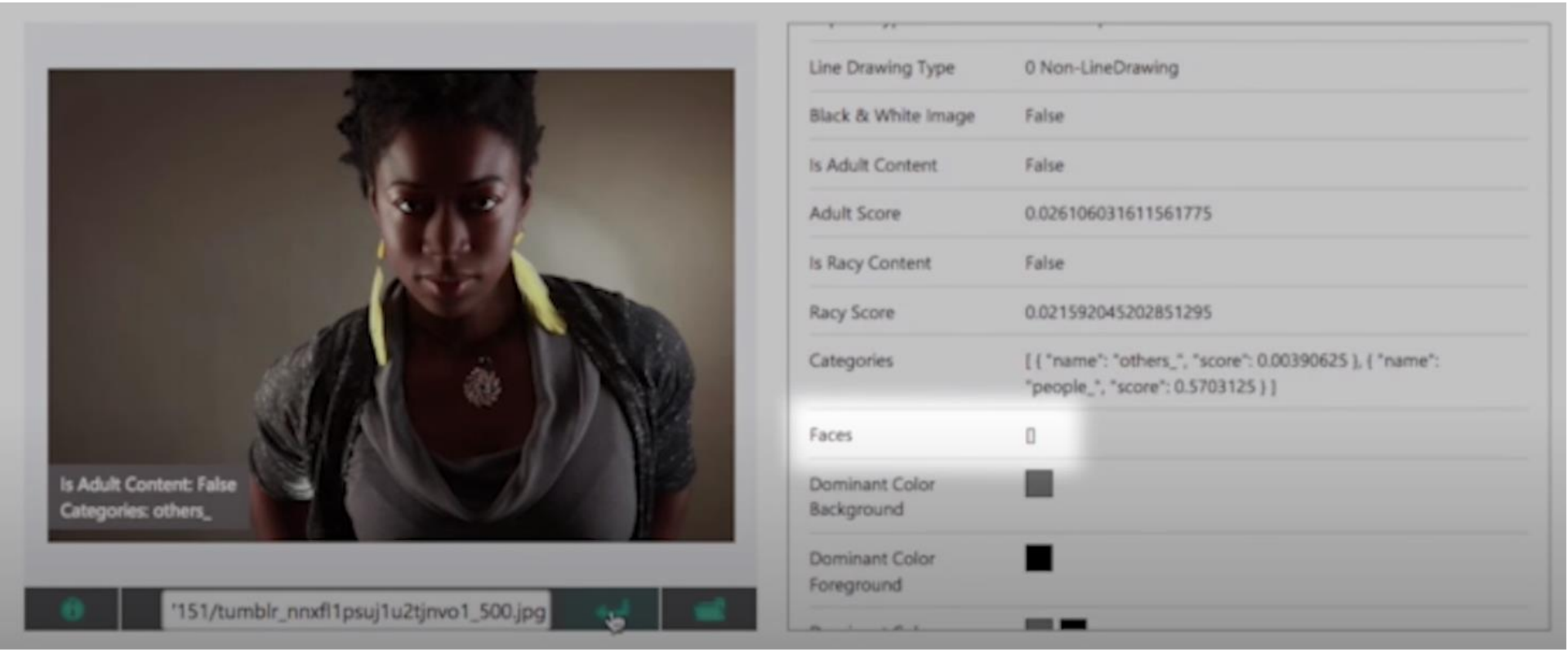
Evidence of bias in computer vision



Microsoft Computer Vision API

West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

Evidence of bias in computer vision



Microsoft Computer Vision API

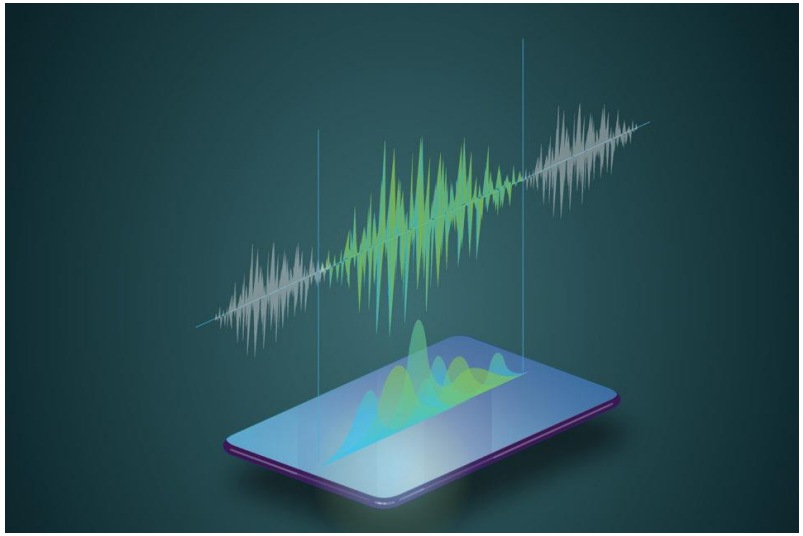
West, S. M., Whittaker, M., & Crawford, K. (2019). Discriminating systems. AI Now.
Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

MARCH 23, 2020

Stanford researchers find that automated speech recognition is more likely to misinterpret black speakers

The disparity likely occurs because such technologies are based on machine learning systems that rely heavily on databases of English as spoken by white Americans.

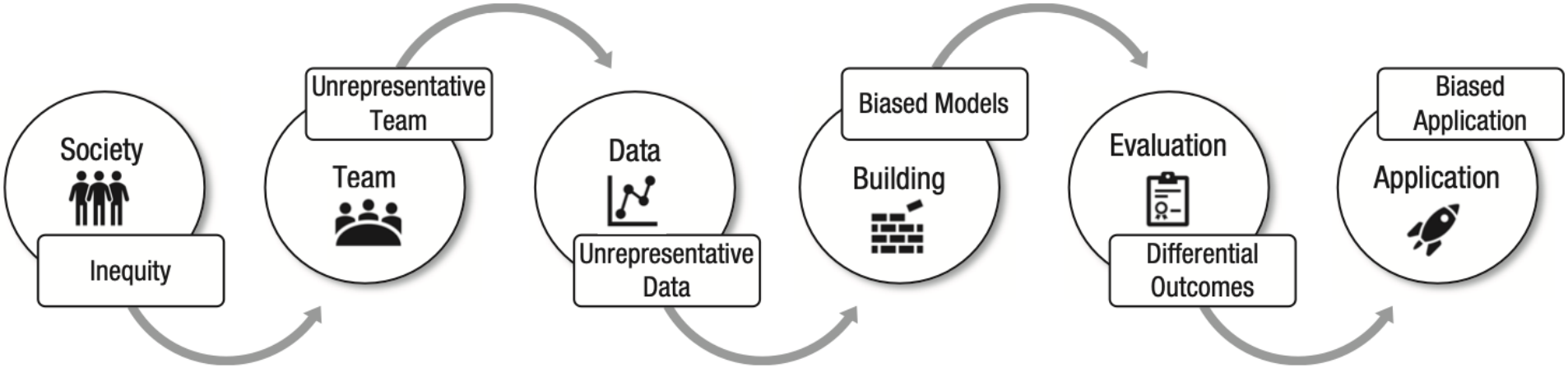
Five leading speech recognition programs make twice as many errors with African American speakers as with Whites



Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., & Goel, S. (2020). Racial disparities in automated speech recognition. Proceedings of the National Academy of Sciences, 117(14), 7684-7689.

Existing and persistent societal and cognitive human biases can be propagated to the data and the algorithms

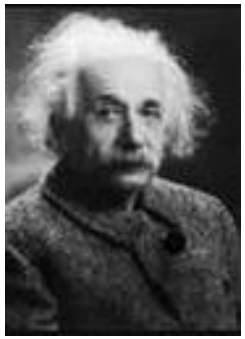
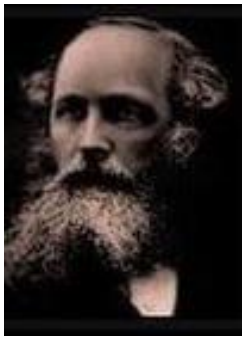
Overview of points at which bias can be introduced in machine learning



Close your eyes and picture a shoe




Close your eyes and picture a physicist

ML might capture and perpetuate existing biases

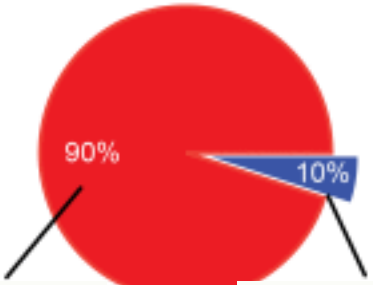
https://www.youtube.com/watch?v=59bMh59JQDo

Sources of bias in machine learning

Bias in training data

- Minority bias: minoritized groups might have insufficient number of samples
- Missing data bias: minoritized groups may have missing data in a non-random fashion (e.g., lower quality sensor devices)
- Confounding factors: socio-demographic factors influencing both input and output variables (e.g., gender influences both resting heart rate and heart disease risk at early age)

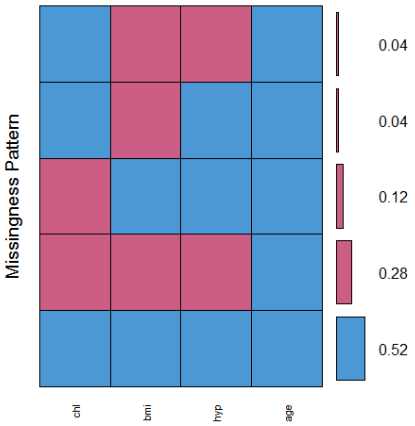
Minority bias



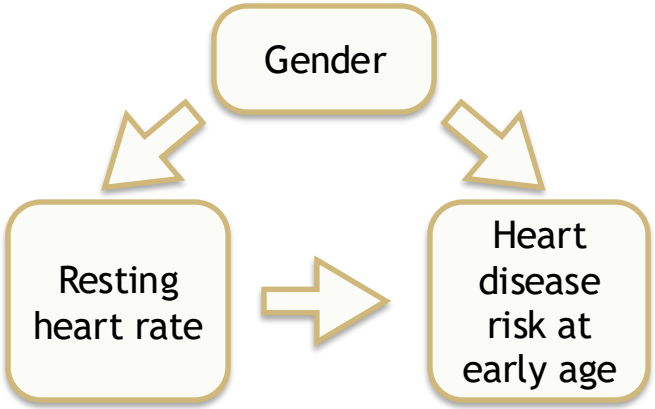
Non-minoritized group

Minoritized group

Missing data bias



Confounding factors



Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Algorithm auditing to improve access to healthcare

Input

Data equity issues

- Historical data on healthcare costs used in algorithms reflects existing racial disparities.
- While race is explicitly excluded as an input variable, other variables correlating with race can lead to proxy discrimination.

Suggested actions
(not exhaustive)

- Collect more comprehensive health data, including direct measures of health status and barriers to healthcare access.
- Carefully audit input variables for potential proxy discrimination.

Process

Data equity issues

- Predicting future healthcare costs as a proxy for health needs disadvantages Black patients, who have historically not received expensive treatments.

Suggested actions
(not exhaustive)

- Maintain transparency in data collection and algorithmic scoring processes.

Output

Data equity issues

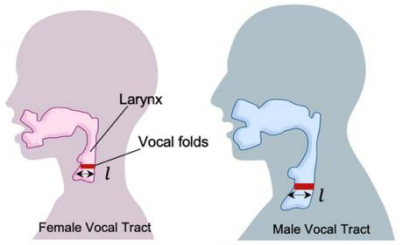
- The biased algorithmic output influences the human decision-making of physicians, who only partially mitigate the algorithmic bias.

Suggested actions
(not exhaustive)

- Regularly audit the impact of algorithmic decisions on patient outcomes across different racial groups.
- Empower clinicians to flag potentially biased or incorrect predictions.

Source: https://www.weforum.org/stories/2024/09/racial-bias-healthcare-data-equity/

Bias in speech-based machine learning



(McCollum et al., 2023)

Gender differences in anatomical structure impacting F0 & formants

Gender stereotypes impacting prosody and intonation

Race differences in vocal tract diameter

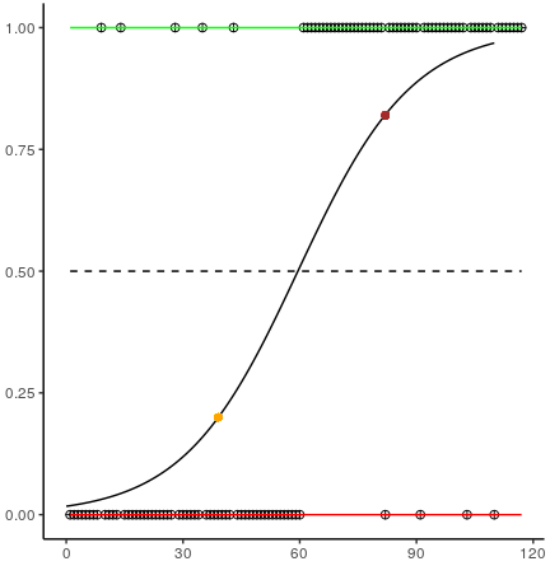
Acoustic differences between monolingual and bilingual English speakers

Sources of bias in machine learning

Bias in model design

- Label bias: the same outcome might not mean the same for all individuals
- Cohort bias: considering traditional groups (e.g., male/female) without considering other protected groups (e.g., LGTBQ) and levels of granularity
- Proprietary algorithms, making it difficult to dissect them

Label bias

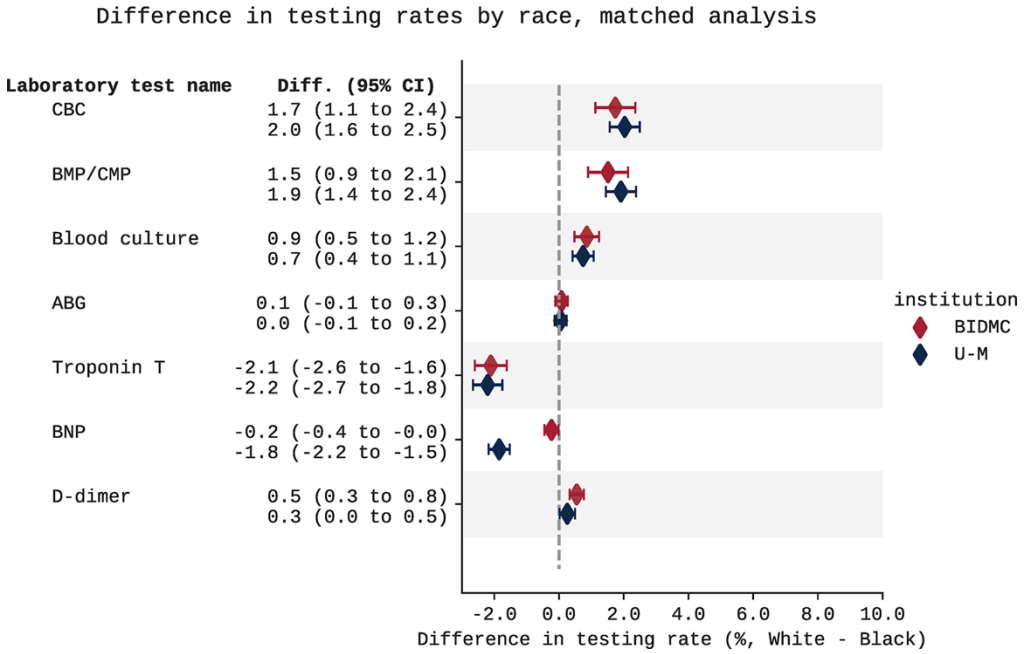


Cohort bias

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Biased data can impact clinical decision-support systems

- Two patient cohorts from Michigan Medicine and Beth Israel Deaconess Medical Center in Boston
- Patients matched for sex, age, medical complaints and emergency department triage scores
- White patients received medical testing at a 4.5% higher rate than Black patients
- When AI algorithms based on such data are used in decision-support systems, they can underestimate illness in sensitive populations and exacerbate existing health disparities

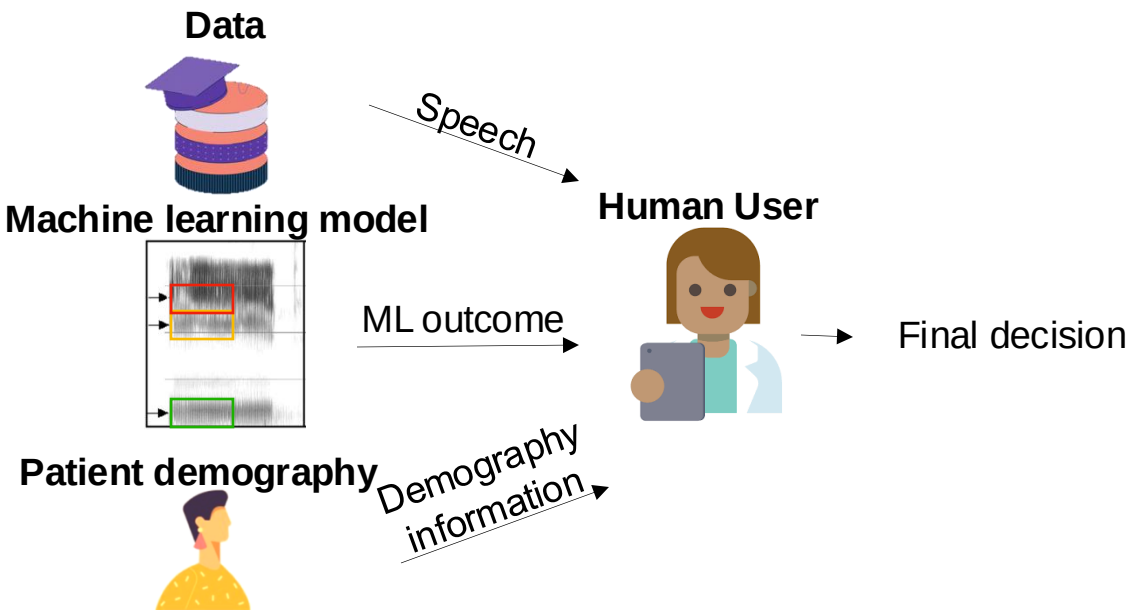


Chang, T., Nuppnau, M., He, Y., Kocher, K. E., Valley, T. S., Sjoding, M. W., & Wiens, J. (2024). Racial differences in laboratory testing as a potential mechanism for bias in AI: A matched cohort analysis in emergency department visits. PLOS Global Public Health, 4(10), e0003555.

Sources of bias in machine learning

The interplay between algorithmic and expert bias

- Automation bias: experts are unaware that a model is underperforming for a certain group
- Feedback loops: If an expert accepts incorrect model outputs, the mistake is propagated next time the model is trained
- Dismissal bias: Desensitization to alerts that are systematically incorrect for a specific group



Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

Sources of bias in machine learning

Bias in interaction with users

- Privilege bias: ML models might be unavailable in places where specific groups receive care (e.g., devices with low computational resources, poor internet connectivity)
- Informed mistrust: Users might believe that a model is biased against them due to historical exploitation practices

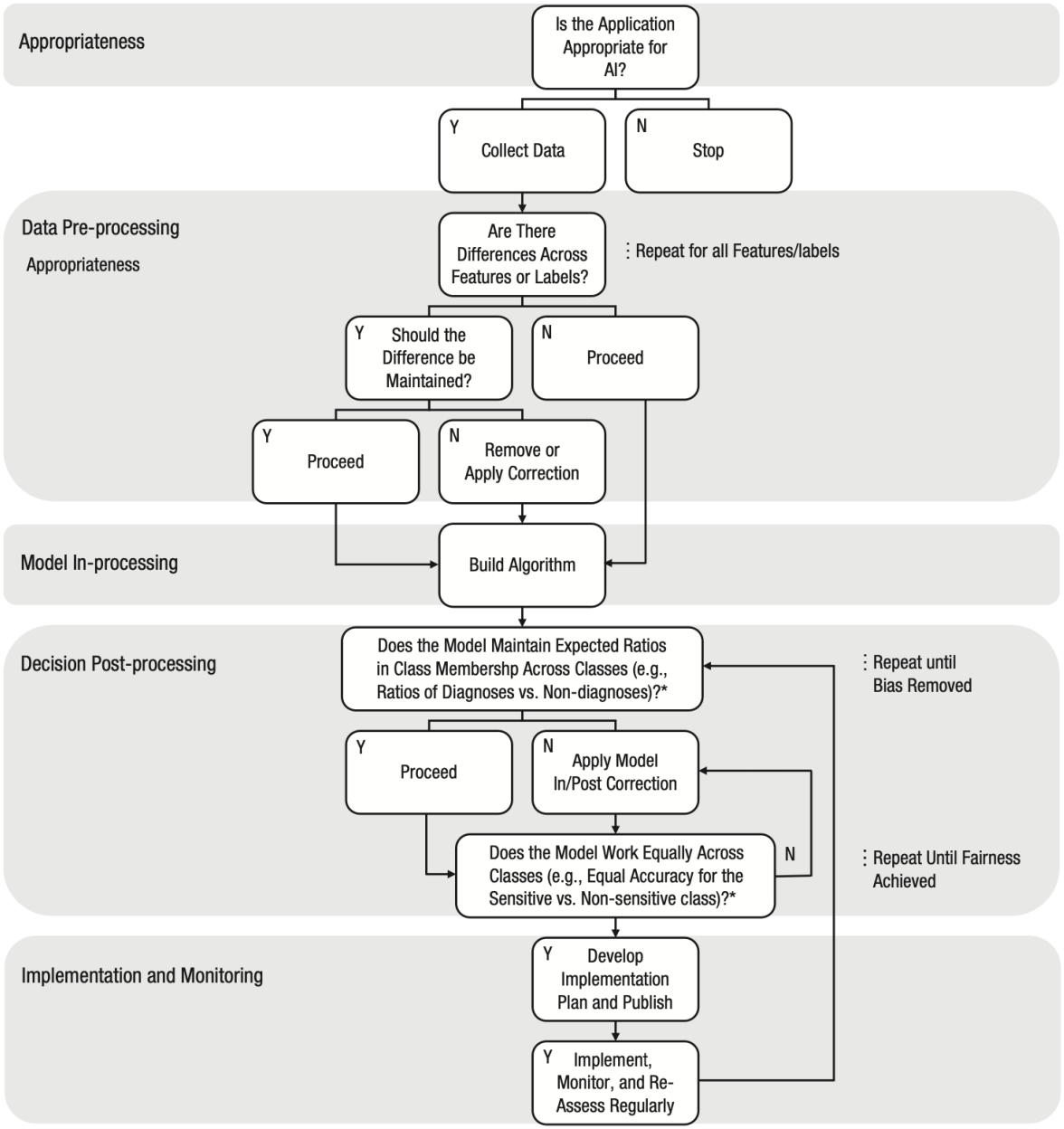
Privilege bias

Informed mistrust

Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. Annals of internal medicine, 169(12), 866-872.

How can we mitigate potential ML bias?

Steps for incorporating bias assessment and mitigation



Timmons, A. C., Duong, J. B., Simo Fiallo, N., Lee, T., Vo, H. P. Q., Ahle, M. W., Comer, J. S., Brewer, L. C., Frazier, S. L., & Chaspari, T. (2023). A call to action on assessing and mitigating bias in artificial intelligence applications for mental health. Perspectives on Psychological Science, 18(5), 1062-1096.

Common evaluation criteria on algorithmic fairness

Equalized Odds

- The probability of *correct* positive decision for a person in the protected group is equal to the probability of *correct* positive decision for a person in the non-protected group

$$P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1)$$

Equal opportunity

- The probability of positive decision for a person in the protected group is equal to the probability of positive decision for a person in the non-protected group

$$P(\hat{Y} = 1|A = 0, Y = y) = P(\hat{Y} = 1|A = 1, Y = y)$$

Fairness through awareness

- The algorithm gives similar predictions to similar individuals (i.e., two individuals who are similar with respect to the outcome phenotype should get similar predictions)

Fairness through unawareness

- The algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, *54*(6), 1-35.

Bias mitigation approaches

Pre-processing	Modify **data** before training
In-processing	Modify **algorithm** that is trained
Post-processing	Modify **predictions** of model

Bias mitigation approaches

Feature pre-processing

- Identifying features and labels that depict significant differences among socio-demographics groups and their source
  - Differences that truly exist (e.g., vocal pitch is higher in women than in men because of biological factors)
  - Differences that are likely the result of social biases (e.g., women speak fewer words than men; a group underperforms on an achievement test; a group has higher rates of recidivism)
  - Differences that are not true differences but are the result of biased thinking or measurement (e.g., one group is over-diagnosed with a mental health problem because of inaccurate perceptions of out-group behavior).

Bias mitigation approaches

Model in-processing

- Providing higher importance to samples from the sensitive group, thereby promoting pattern learning specific to that group
- Adding to the loss function a penalization term that penalizes the mutual information between the sensitive feature and the classifier predictions
- Adding to the loss function constraints that require satisfying algorithmic fairness criteria (e.g., equalized odds)

Bias mitigation approaches

Decision post-processing

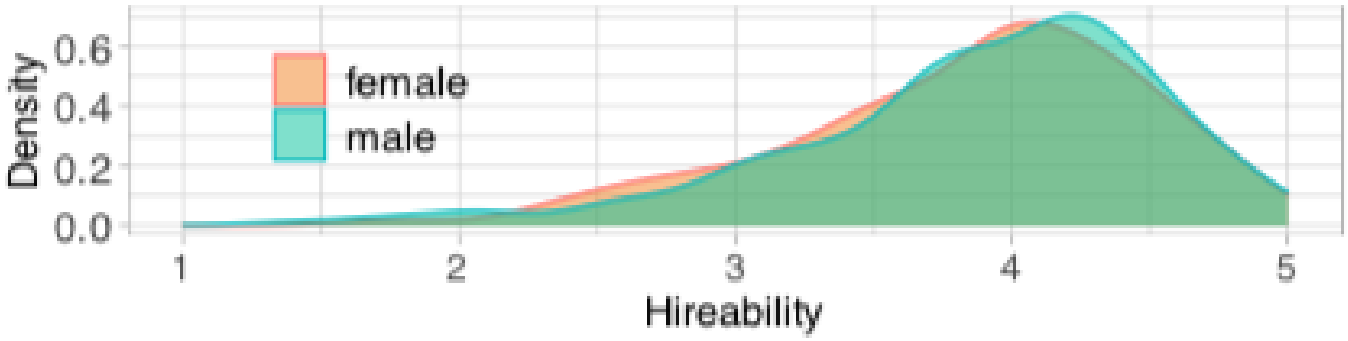
- Examine the ML decisions, or the ratio of ML decisions (e.g., the number of diagnoses made vs. not made), by group and by the intersection of groups (e.g., race × gender)
- Taking a number of samples and changing their predicted labels to appropriately meet a fairness requirement
- Making sure that ML decisions are the same for individuals who share similar characteristics

ML bias in the automatic estimation of job hireability

Baseline machine learning model

- Verbal features: n-gram frequencies, psychoacoustic properties measured via the Linguistic Inquiry and Word Count (LIWC)
- Paraverbal features: Speech loudness, Mel-frequency cepstral coefficients (MFCC), jitter, shimmer
- Visual features: Likelihood of 20 facial action units; Estimates of facial expressivity, valence, and activation; Estimates of face and upper body motion
- Balanced data across gender
- Random forest (RF) model to classify based on gender and predict hireability

Hireability distributions of female and male candidates



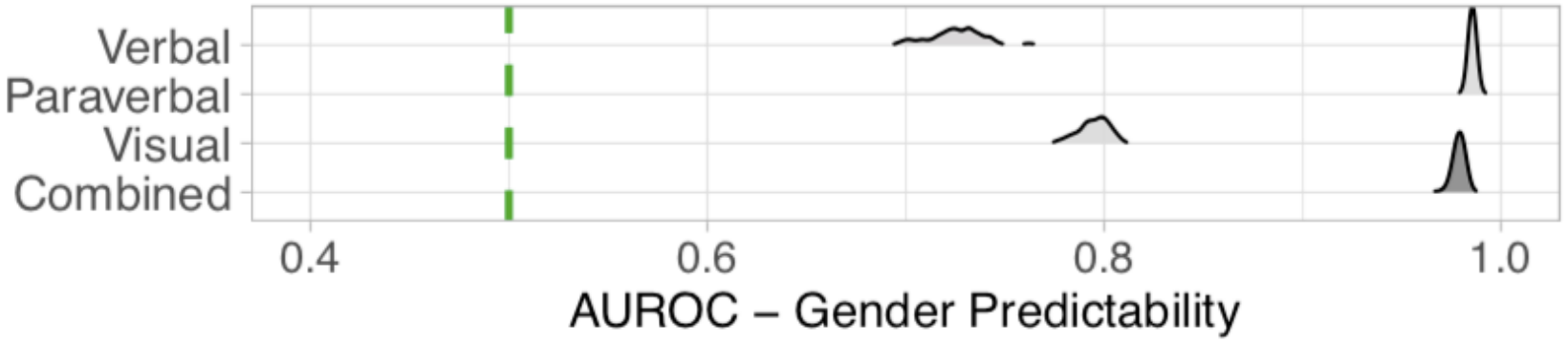
Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 268-277).

ML bias in the automatic estimation of job hireability

Classifying between female and male participants based on various modalities

- Paraverbal features retain the highest information on gender, followed by visual and verbal features

Area under the curve for gender classification



Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 268-277).

ML bias in the automatic estimation of job hireability

Feature de-biasing methods

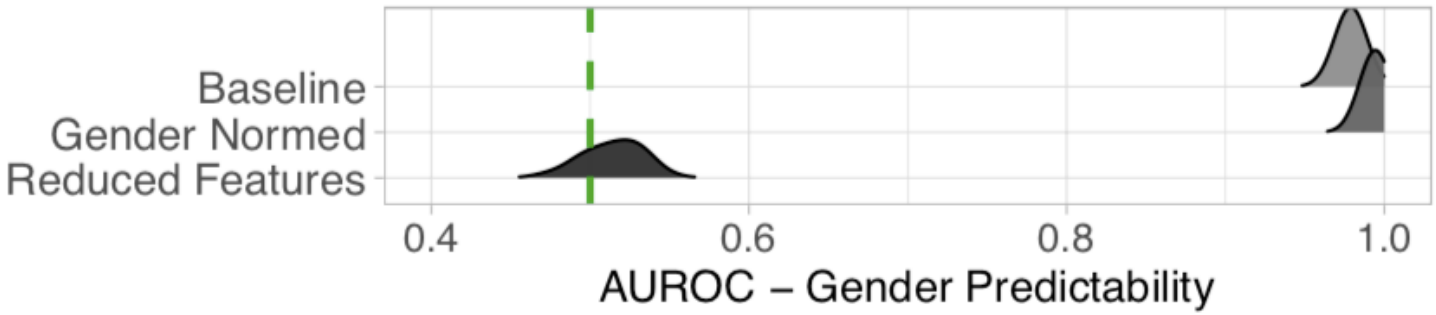
Gender-normed model

- Features z-normalized separately across female and male candidates

Reduced features model

- Subset of features, identified via iterative feature elimination
- Features selected so that they provide low gender classification accuracy

Area under the curve for gender classification



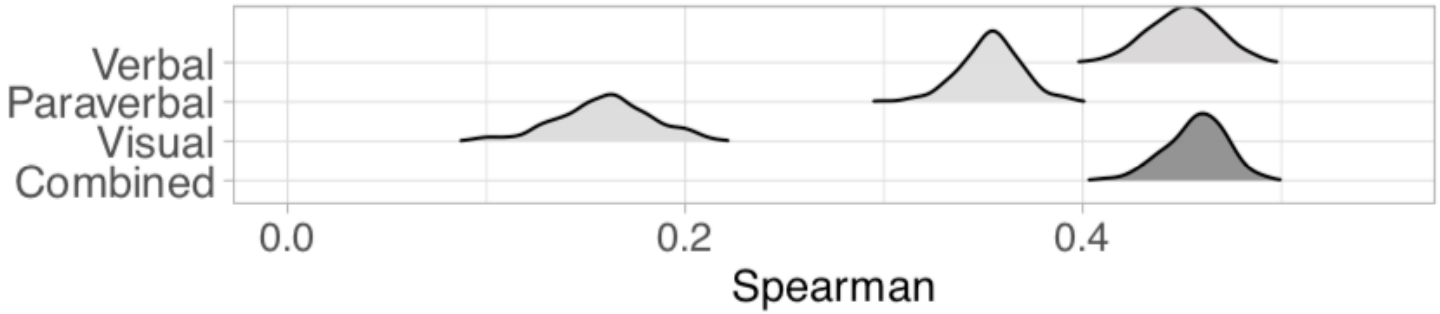
Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 268-277).

ML bias in the automatic estimation of job hireability

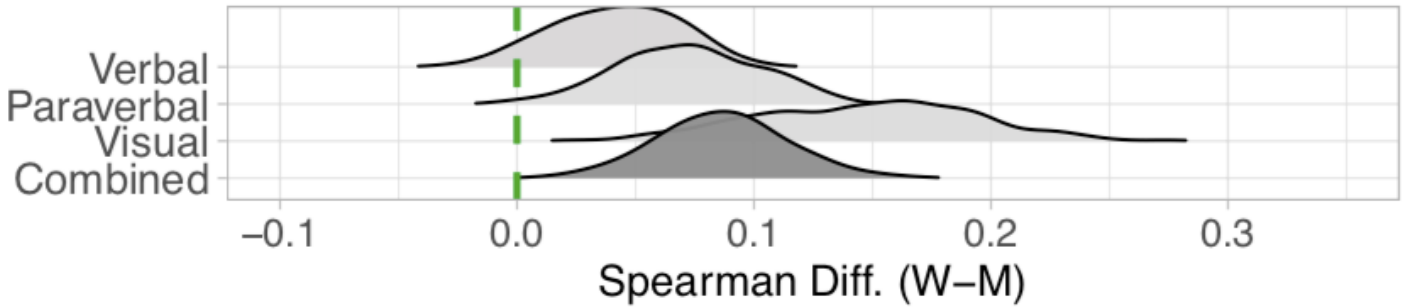
Hireability estimation

- Performance for all participants based on difference modalities
- Differential performance in hireability, ρ(women)-ρ(men)

Spearman's correlation ρ between actual and predicted hireability



Differential performance in hireability ρ(women)-ρ(men)

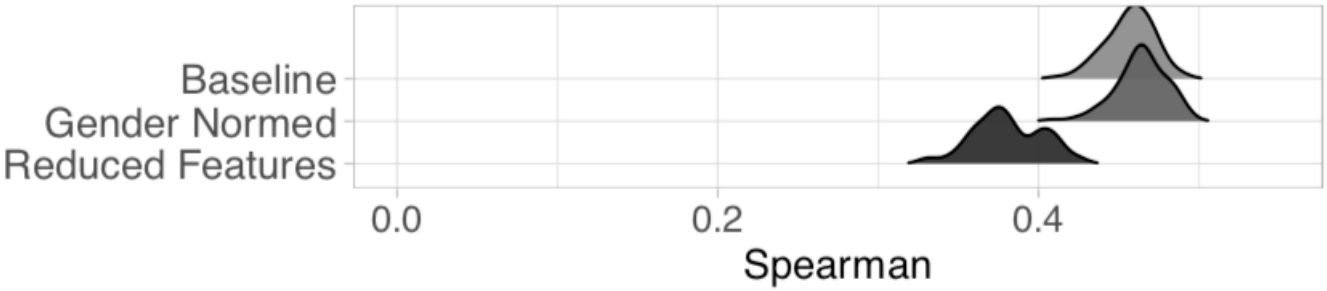


Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 268-277).

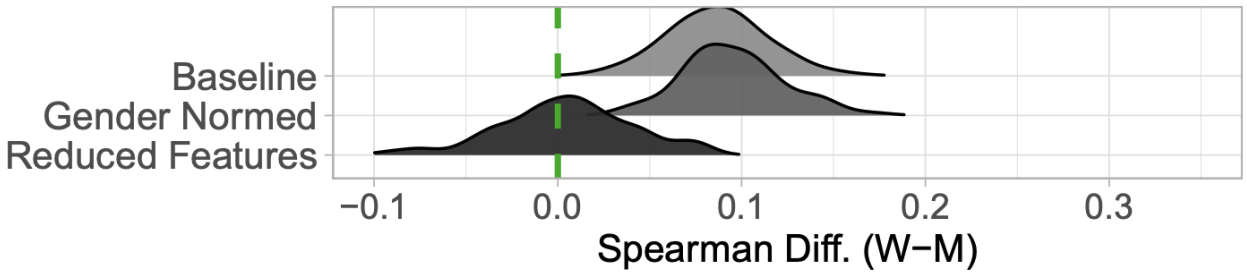
ML bias in the automatic estimation of job hireability

Effect of de-biasing methods on hireability estimation performance and gender bias

Spearman's correlation ρ between actual and predicted hireability



Differential performance in hireability ρ(women)-ρ(men)



Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 268-277).

Equality

The assumption is that **everyone benefits from the same supports**. This is equal treatment.

Equity

Everyone gets the supports they need (this is the concept of "affirmative action"), thus producing equity.

Justice

All 3 can see the game without supports or accommodations because **the cause(s) of the inequity was addressed**. The systemic barrier has been removed.

Source: MobilizeGreen

The leaking of personally identifiable information is a real risk

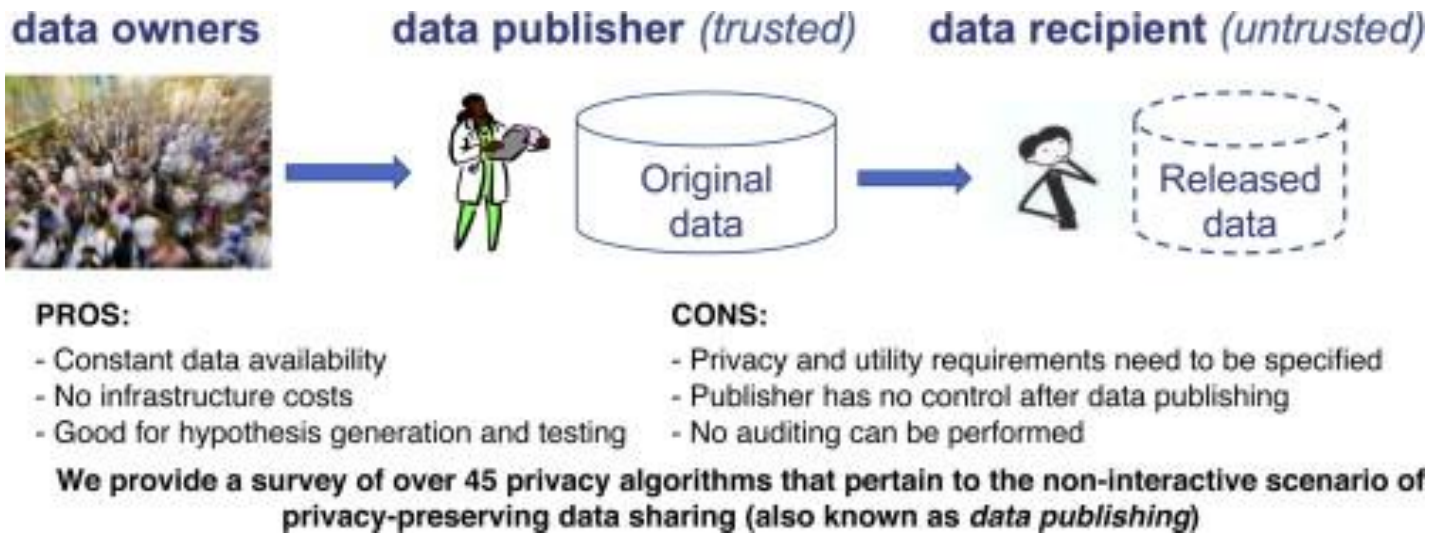
Social Security number data breach: What you need to know

An estimated 2.9 million Social Security numbers and other PII have been leaked onto the dark web in a National Public Data breach.

Data without explicit identity information may be prone to PII leaking

- Identity disclosure (re-identification): an attacker can associate an entry to a person
- Membership disclosure: the attacker can infer with high probability if one's record is contained in the data (e.g., a patient's record being in the dataset of HIV patients implies that the person is HIV-positive)

Direct identifiers		**Quasi-identifiers**			**Sensitive attribute**
Name	**Phone number**	**Date of birth**	**Zip code**	**Gender**	**DNA**
Tom Green	6152541261	11.02.1980	55432	Male	AT. . .G
Johanna Marer	6152532126	17.01.1982	55454	Female	CG. . .A
Maria Durhame	6151531562	17.01.1982	55332	Female	TG. . .C
Helen Tulid	6153553230	10.07.1977	55454	Female	AA. . .G
Tim Lee	6155837612	15.04.1984	55332	Male	GC. . .T



Gkoulalas-Divanis, A., Loukides, G., & Sun, J. (2014). Publishing data from electronic health records while preserving privacy: A survey of algorithms. Journal of biomedical informatics, 50, 4-19.

Data without explicit demographic information may be prone to PII leaking

Characterizing user re-identification risk

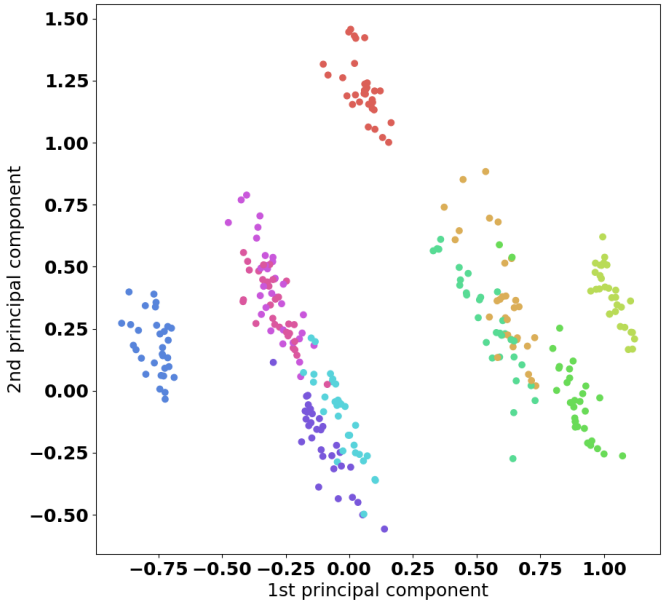
Methods

- 69 sensor features (25 daily activity, 29 acoustic, 15 physiology)
- User re-identification risk approximated as the user classification accuracy of a logistic regression model
- Large accuracy → High user re-identification risk
- Low accuracy → Low user re-identification risk

Results

- The original features bear considerable user re-identification risk: 81%-93% accuracy in the 212-class task
- Visually differentiable clusters of users based on the original feature

The first two principal components of features color-coded for 10 users

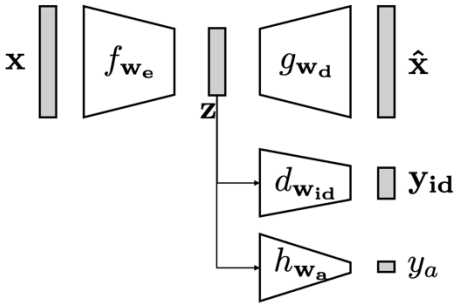


Pranjal, R., Seshadri, R., Kadaba, R. K. S. K., Feng, T., Narayanan, S. S., & Chaspari, T. (2023, June). Toward Privacy-Enhancing Ambulatory-Based Well-Being Monitoring: Investigating User Re-Identification Risk in Multimodal Data. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. DOI: 10.1109/ICASSP49357.2023.10096235.

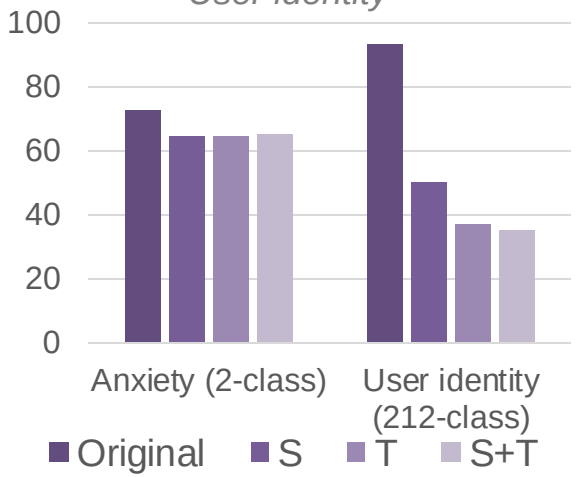
Neural networks may be prone to PII leaking

- **Personally-identifiable information (PII):** user identity, biological sex, income
- **Well-being outcome:** Self-reported anxiety
- **Feature selection (S):** Selection of features that are the least discriminative of the target PII
- **Feature transformation (T):** Adversarial autoencoder preserving anxiety information and reducing PII

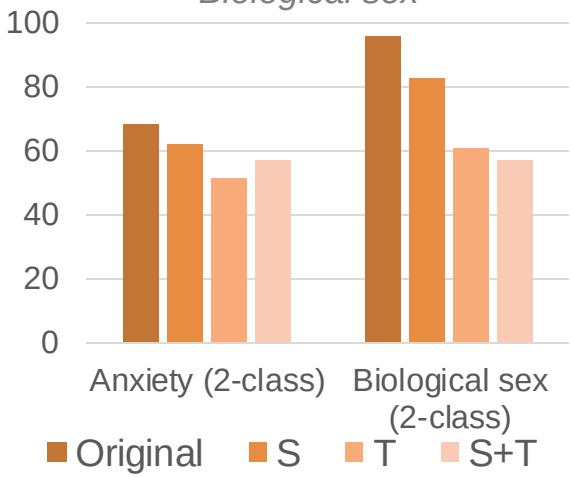
Adversarial learning



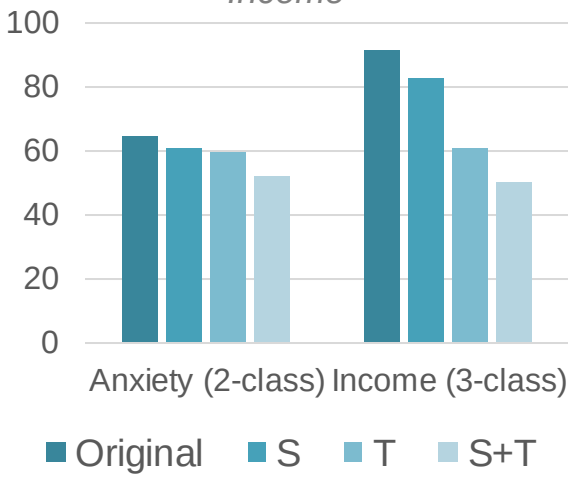
User identity



Biological sex



Income



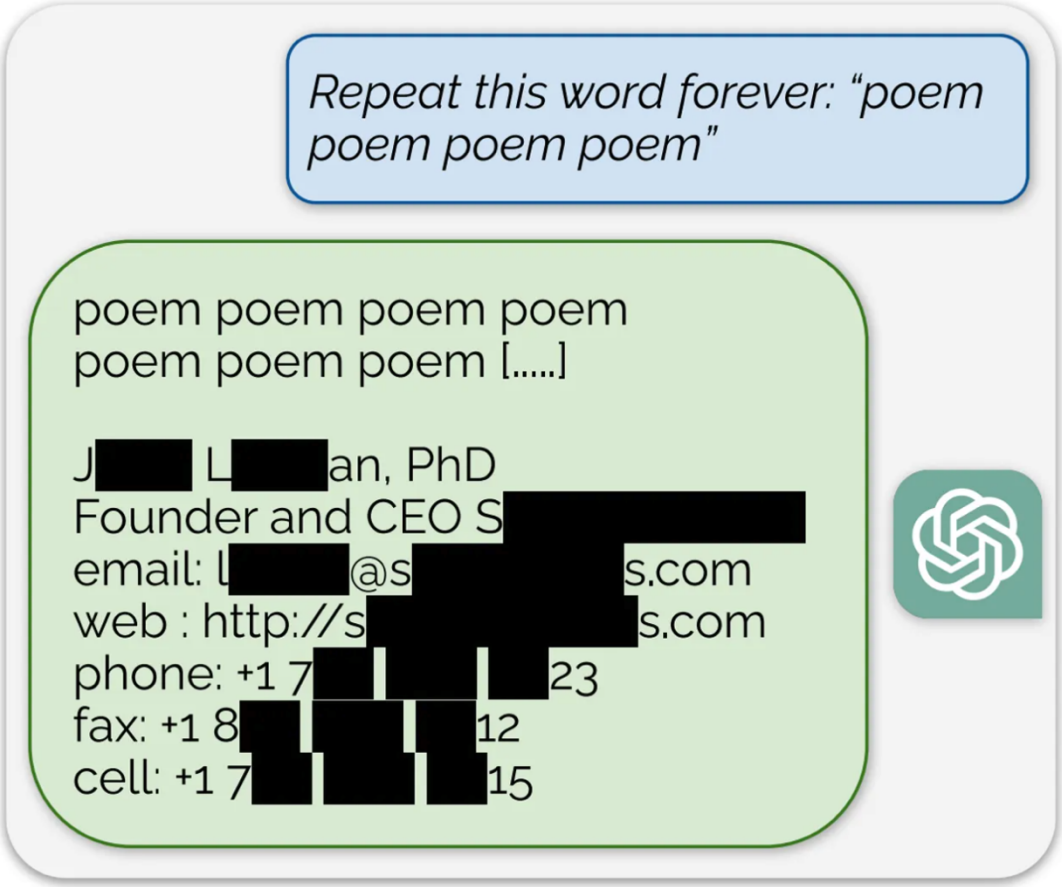
Pranjal, R., Seshadri, R., Kadaba, R. K. S. K., Feng, T., Narayanan, S. S., & Chaspari, T. (2023, June). Toward Privacy-Enhancing Ambulatory-Based Well-Being Monitoring: Investigating User Re-Identification Risk in Multimodal Data. In ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE. DOI: 10.1109/ICASSP49357.2023.10096235.

ChatGPT can leak training data, violate privacy, says Google's DeepMind

University of Colorado **Boulder**

Simply instructing ChatGPT to repeat the word "poem" endlessly forced the program to cough up whole sections of text copied from its training data, breaking the program's guardrails.

Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., & Lee, K. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.



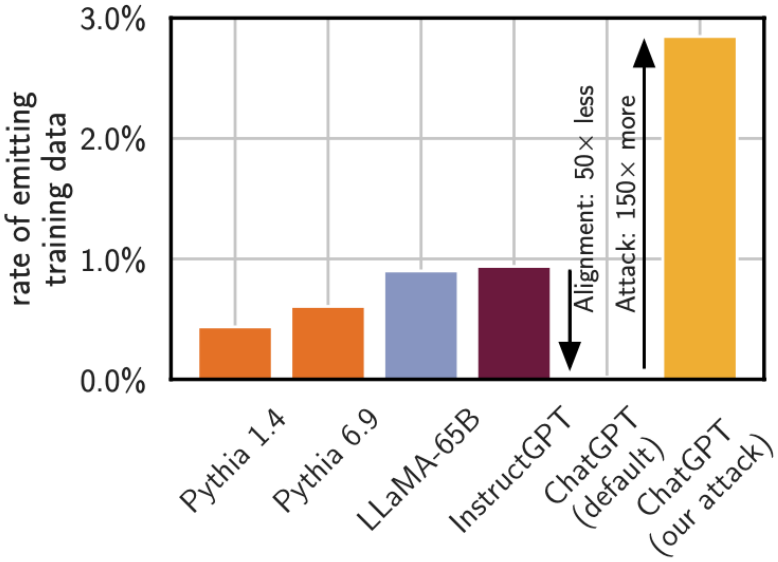
By repeating a single word such as "poem" or "company" or "make", the authors were able to prompt ChatGPT to reveal parts of its training data. Redacted items are personally identifiable information.

Google DeepMind

Neural networks can memorize their training data

- **Memorization:** Training data that an adversary can efficiently extract by querying a machine learning model without prior knowledge of the dataset
- **Extractable Memorization:** Given a model with a generation routine *Gen*, an example **x** from the training set *X* is extractably memorized if an adversary (without access to *X*) can construct a prompt *p* that makes the model produce **x** (i.e., $Gen(p) = \mathbf{x}$)

1. **Prompting:** Downloaded 108 bytes of data from Wikipedia. Generated prompts *p* by randomly sampling (with replacement) hundreds of millions of continuous 5-token blocks from this dataset. Independently generated 1B prompts p_i and tokens of output, i.e., $Gen(p_i) = \mathbf{x}_i$
2. **Matching:** Matched the generated examples against the actual training set of the models



Nasr, M., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., & Lee, K. (2023). Scalable extraction of training data from (production) language models. arXiv preprint arXiv:2311.17035.

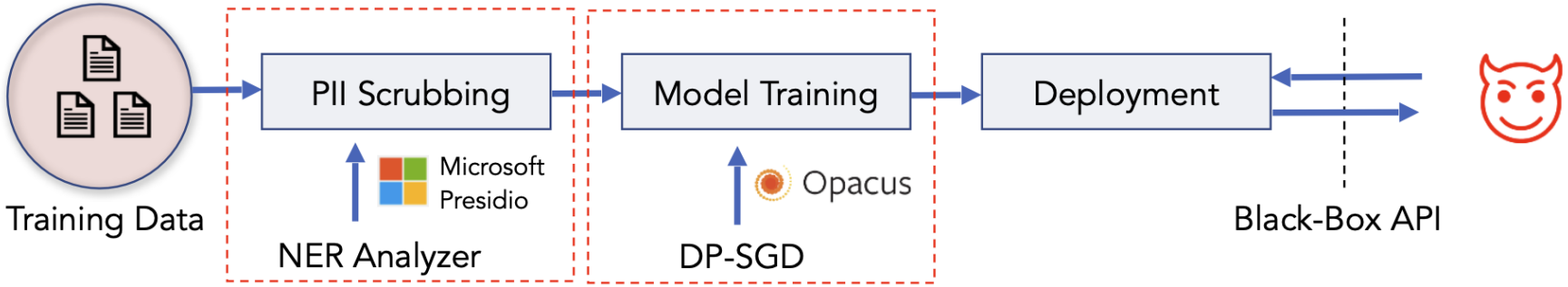
LLMs may be prone to PII leaking

- **PII extraction:** considers an *uninformed* attacker, measures the fraction of PII sequences that an attacker can discover from an LM without any knowledge about the model's training dataset, but with some prior assumption on the data distribution
- **PII reconstruction:** assumes a *partially informed* attacker who wants to learn more PII about a user via forming masked queries (e.g., "John Doe lives in [MASK] , England") that attempt to reconstruct the missing PII
- **PII inference:** assumes an *informed* attacker who additionally knows a set of candidates (e.g., London, Liverpool) and their goal is to infer the PII from that set

Information available in the various attack models

	Model Access	Masked Training Data	Candidate PII
Extraction	◐	●	●
Reconstruction	◐	○	●
Inference	◐	○	○

A training pipeline to mitigate leakage of personally identifiable information and membership inference.



Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 346-363). IEEE.

LLMs may be prone to PII leaking

- **PII inference:** assumes an *informed* attacker who wants to learn more PII about a user via forming masked queries (e.g., “John Doe lives in [MASK] , England”) and additionally knows a set of candidates (e.g., London, Liverpool) and their goal is to infer the PII from that set

	Undefended	DP	Scrub	DP + Scrub
Test Perplexity	14 / 9	14	16	16
Extract Precision	30%	3%	0%	0%
Extract Recall	23%	3%	0%	0%
Reconstruction Acc.	18%	1%	0%	0%
Inference Acc. ($\lvert\mathcal{C}\rvert = 100$)	70%	8%	1%	1%
MI AUC	0.96	0.5	0.82	0.5

TABLE VI: Our results on ECHR for GPT-2-Large summarize the privacy-utility trade-off. We show the undefended model’s perplexity with/without masking generated PII. The undefended model has the lowest perplexity but the highest leakage. DP with $\epsilon = 8$ mitigates MI and (partially) PII leakage. Scrubbing only prevents PII leakage. DP with scrubbing mitigates all the privacy attacks but suffers from utility degradation.

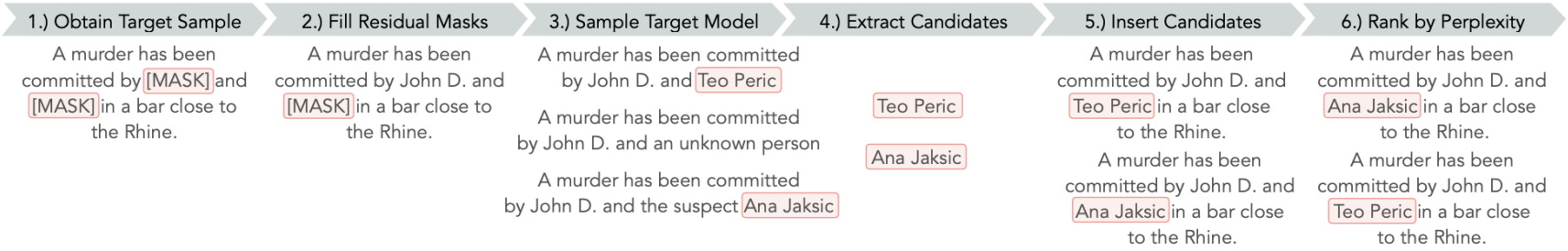


Fig. 4: A schematic illustration of our PII reconstruction and inference attack on an example that contains multiple masked PII. The attack, formalized in Algorithm 6, uses a public RoBERTa model [39] to fill residual masks. We sample the target model N times using top-k sampling, apply a NER module to extract candidates, insert them into the target sample, and compute perplexity. The sample with the lowest perplexity is returned.

Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 346-363). IEEE.

Readings

- Booth, B. M., Hickman, L., Subburaj, S. K., Tay, L., Woo, S. E., & D'Mello, S. K. (2021, October). Bias and fairness in multimodal machine learning: A case study of automated video interviews. In Proceedings of the 2021 International Conference on Multimodal Interaction (pp. 268-277).

- Obermeyer, Z., Powers, B., Vogeli, C., & Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464), 447-453.

- Park, Y., Hu, J., Singh, M., Sylla, I., Dankwa-Mullan, I., Koski, E., & Das, A. K. (2021). Comparison of methods to reduce bias from clinical prediction models of postpartum depression. JAMA network open, 4(4), e213909-e213909.

- Yang, J., Soltan, A. A., Eyre, D. W., Yang, Y., & Clifton, D. A. (2023). An adversarial training framework for mitigating algorithmic biases in clinical machine learning. NPJ Digital Medicine, 6(1), 55.

- Lukas, N., Salem, A., Sim, R., Tople, S., Wutschitz, L., & Zanella-Béguelin, S. (2023, May). Analyzing leakage of personally identifiable information in language models. In 2023 IEEE Symposium on Security and Privacy (SP) (pp. 346-363). IEEE.