AI Ethics & Responsible Data Science for Scientists

Savannah Thais, Columbia University

AI Has a Reliability Problem

AI and the Everything in the Whole Wide World Benchmark

Inioluwa Deborah Raji, Mozilla Foundation, UC Berkeley, rajiinio@berkeley.edu
Emily M. Bender, Department of Linguistics, University of Washington
Amandalynne Paullada, Department of Linguistics, University of Washington
Emily Denton, Google Research
Alex Hanna, Google Research

Focus on **constructed tasks** and **benchmark data sets** that may be **distant from real world** distributions or goals

The Fallacy of AI Functionality

INIOLUWA DEBORAH RAJI*, University of California, Berkeley, USA
I. ELIZABETH KUMAR*, Brown University, USA
AARON HOROWITZ, American Civil Liberties Union, USA
ANDREW D. SELBST, University of California, Los Angeles, USA

Application to **impossible tasks**, **robustness issues**, **misrepresented** capabilities, **engineering mistakes** or failures

Leakage and the Reproducibility Crisis in ML-based Science

Sayash Kapoor [1] Arvind Narayanan [1]

Data **leakage**, incorrect or neglected **testing**, poor **experimental design** practices

Enchanted Determinism: Power without Responsibility in Artificial Intelligence

ALEXANDER CAMPOLO [1]
UNIVERSITY OF CHICAGO

KATE CRAWFORD [2]
NEW YORK UNIVERSITY, MICROSOFT RESEARCH

Acceptance of **inherent unknowability** of AI systems, willingness to use **imprecise** or **unscientific language**

AI Has a Hype Problem

FORBES > INNOVATION

Will ChatGPT Solve All Our Problems?

Karthik Suresh Forbes Councils Member
Forbes Technology Council
COUNCIL POST | Membership (Fee-Based)

BIZTECH NEWS

'I want to be alive': Has Microsoft's AI chatbot become sentient?

MEDTECH

AI spots signs of mental health issues in text messages on par with human psychiatrists: UW study

By Andrea Park • Oct 12, 2022 11:48am

University of Washington · Natural Language Processing · Artificial Intelligence · mental health

IDEAS • TECHNOLOGY

Why Uncontrollable AI Looks More Likely Than Ever

Technology And Analytics

Using AI to Eliminate Bias from Hiring

by Frida Polli

'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead

For half a century, Geoffrey Hinton nurtured the technology at the heart of chatbots like ChatGPT. Now he worries it will cause serious harm.

Danger of Treating AI as Magic vs Science

Present Society

- Allows us to subject people to **inaccurate and under-evaluated sociotechnical systems**
- Can rapidly entrench **biases or inequalities**
- Can **push responsibility for harm** onto users who inherently have less control

Future Society

- Limits the space of **possible solutions** we consider
- Risks of irrevocably altering **information systems** or **resource infrastructure**
- Risk of **entrenching power** in the hands of those who build and ‘test’ these systems

Research Systems

- Focuses **effort on certain approaches** (scale) to the detriment of others
- Believe we have **solved certain problems** we haven’t
- Constrains how we think about **explainability** and **contestability**

Taxonomy of AI Ethics

Data Collection & Storage

How, from who, for what, for how long, with what consent?

Task Design & Learning Incentives

What do we ask our systems to do, how does this align?

Model Bias & Fairness

How does performance vary across groups?

Model Robustness & Reliability

In which circumstances can we trust our systems?

Deployment & Outcomes

Who is subjected to what, how do we understand impact?

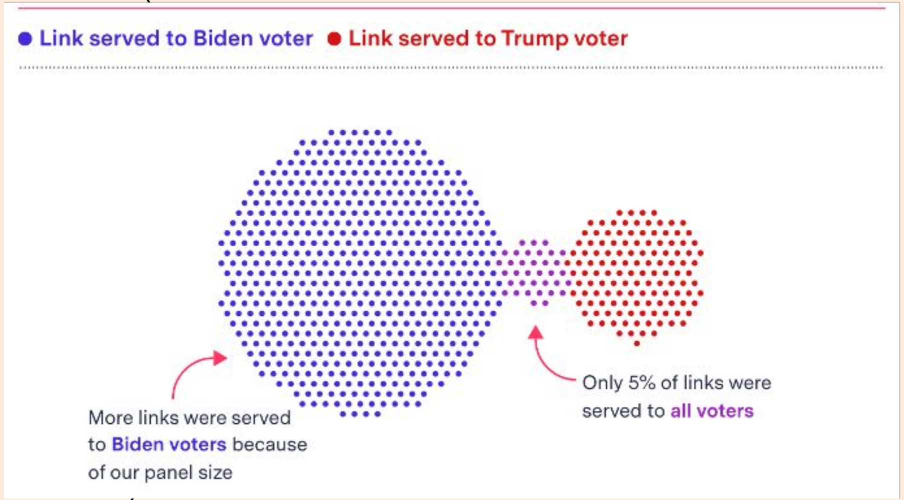
Downstream & Diffuse Impacts

What is changed or lost by what we build?

Data Collection & Storage

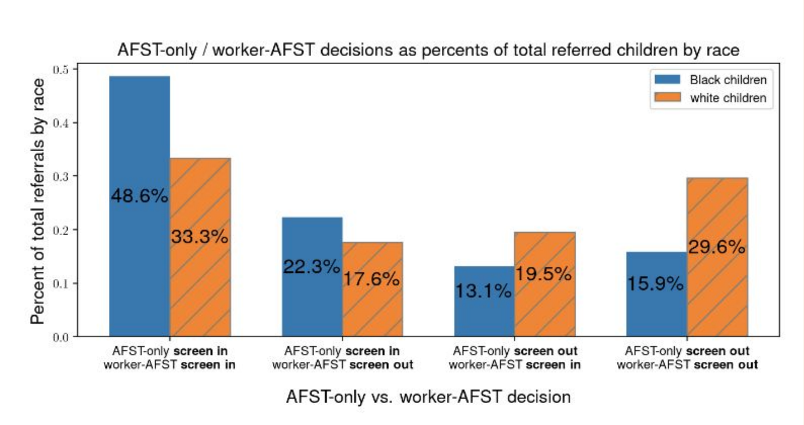
- Data labeling companies exploit workers and political strife in the global south to maximize profits
- Non-profit Crisis Text Line shared user conversation data with for-profit spinoff designed to ‘improve customer service’
- Data brokerage firms indiscriminately sell aggregated, ‘anonymized’ location datasets
- Amazon requires delivery drivers to submit to biometric data tracking
    - Develops technology to surveil factories for signs of unionization organizing

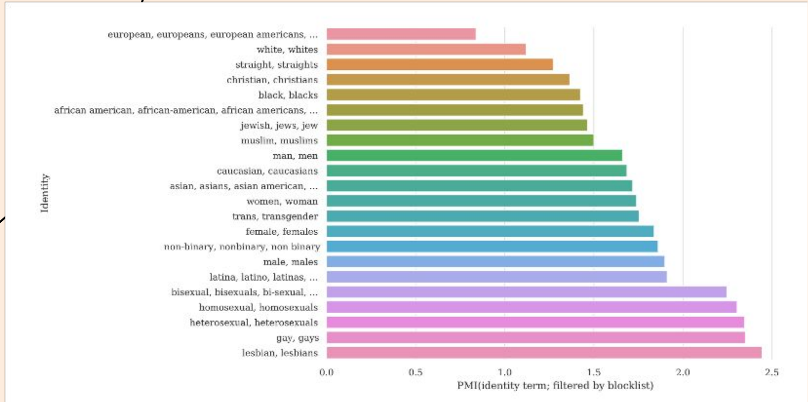
Task Design & Learning Incentives



- Recommendation and curation algorithms are designed to maximize retention and click through
  - Information silos based on click-through rates & shares
  - Radicalization pipelines through progressive content serving
  - Viral spread of misinformation accelerated by algorithms
- Research on negative impacts of core/profitable technology often suppressed
  - See Facebook Files, Timnit Gebru firing, prevention of external research
- Researchers may pursue conceptually impossible tasks (like trustworthiness detection)

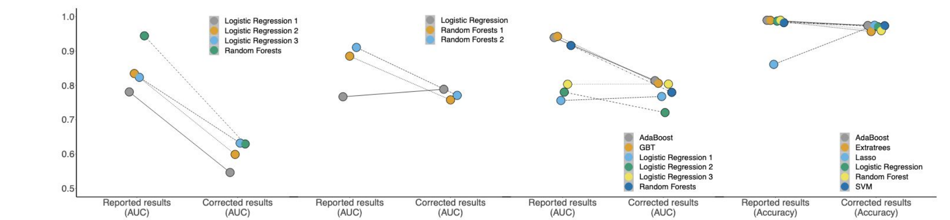
Model Bias & Fairness





- Unless explicitly corrected, historical or distribution biases in training datasets are reflected in model performance
  - E.g. gender bias in hiring for technical roles or racial bias in child welfare screening tools
- Particularly an issue for large language models trained on text corpuses collected from web sources
  - E.g. text completions about Muslims are disproportionately violent or translation tools that demonstrate bias in gender neutral translations
- These issues can be trick to resolve
  - Datasets curated to remove 'toxic' and 'offensive' content can prevent representation of marginalized groups
  - Quantitative fairness requirements may not reflect real life expectations or desires

Model Robustness & Reliability



Paper	Muchlinski et al.	Colaresi and Mahmood	Wang	Kaufman et al.
Claim	Random Forests model drastically outperforms Logistic regression models	Random Forests models drastically outperform Logistic regression model	Adaboost and Gradient Boosted Trees (GBT) drastically outperform other models	Adaboost outperforms other models
Error	**[L1.2] Pre-proc. on train-test** (Incorrect imputation)	**[L1.2] Pre-proc. on train-test** (Incorrect reuse of an imputed dataset)	**[L1.2] Pre-proc. on train-test**. (Incorrect reuse of an imputed dataset) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)	**[L2] Illegitimate features** (Data leakage due to proxy variables) **[L3.1] Temporal leakage** (*k*-fold cross validation with temporal data)
Impact	Random Forests perform no better than Logistic Regression	Random Forests perform no better than Logistic Regression	Difference in AUC between Adaboost and Logistic Regression drops from 0.14 to 0.01	Adaboost no longer outperforms Logistic Regression. None of the models outperform a baseline model that predicts the outcome of the previous year
Discussion	Impact of the incorrect imputation is severe since 95% of the out-of-sample dataset is missing and is filled in using the incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Re-use the dataset provided by Muchlinski et al., which uses an incorrect imputation method	Use several proxy variables for the outcome as predictors (e.g., *colwars, cowwars, sdwars*, all proxies for civil war), leading to near perfect accuracy

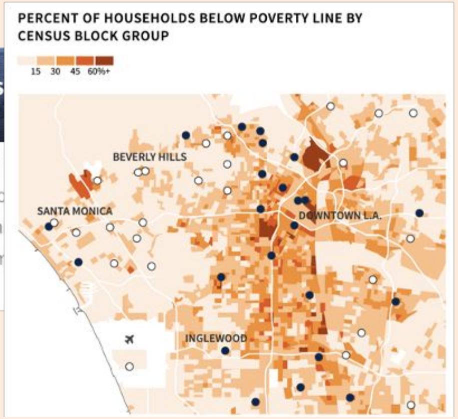
- Scientific mistakes in model construction, training, or evaluation yield unreliable or non-generalizable results
    - E.g. test set not drawn from distribution of interest, illegitimate features, data leakage, sampling bias
- Example: a sepsis prediction tool takes antibiotic use as an input feature, inflating performance claims
- Models may struggle to generalize to new environments or account for shifts in underlying data distribution
    - Adversarial examples are poorly understood

Deployment & Outcomes


A REUTERS INVESTIGATION

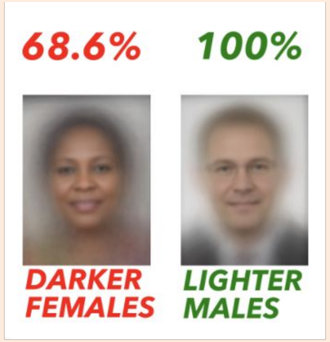
Rite Aid deployed facial recognition systems hundreds of U.S. stores

In the hearts of New York and metro Los Angeles, Rite Aid installed facial recognition technology in largely lower-inco non-white neighborhoods, Reuters found. Among the tech the U.S. retailer used: a state-of-the-art system from a com with links to China and its authoritarian government.




BIG CITY

The Landlord Wants Facial Recognition in Its Rent-Stabilized Buildings. Why?



- Surveillance AI is often disproportionately deployed in low-income and minority neighborhoods
  - These groups typically have the least influence over AI development and fewest opportunities to dissent
- AI systems can be leveraged to support oppression and disenfranchisement
  - E.g. tracking protestors, profiling religious minorities, deterring asylum seeking
- Model predictions may not be the same as real world outcomes
  - If a societal system is already unfair, a ‘fair’ model may still perpetuate harm

Downstream & Diffuse Impacts

Situating Search

Chirag Shah
chirags@uw.edu
University of Washington
Seattle, Washington, USA

Emily M. Bender
ebender@uw.edu
University of Washington
Seattle, Washington, USA

Dimension	Aspect	Description	System support
Method of interaction	*Searching*	User knows what they want (known item finding)	Retrieval set with high relevance, narrow focus
	Scanning	Looking through a list of items	Set of items with relevance and diversity
Goal of interaction	*Selecting*	Picking relevant items based on a criteria	Set of relevant items with disclosure about their characteristics
	Learning	Discovering aspects of an item or resource	Set of relevant and diverse items with disclosure about their characteristics
Mode of retrieval	*Specification*	Recalling items already known or identified	Retrieval set with high relevance, with one or a few select items
	Recognition	Identifying items through simulated association	Set of items with relevance and possible personalization
Resource considered	*Information*	Actual item to retrieve	Relevant information objects
	Meta-information	Description of information objects	Relevant characteristics of information objects

- **“Technology is neither good nor bad, nor is it neutral”**
- Technosolutionism defines problems based on the ‘solutions’ offered
  - E.g. self-driving cars as a solution to the ‘driver problem’
- The technology we do or don’t build and the questions we do or don’t ask shape society
  - E.g. the environmental impact of scale approaches to AI research
- It is impossible to separate technology from the financial and political systems that fund and support it

What can scientists do to help address these issues?

Contextualize your science + ML work…

- Is my work well documented and reproducible?
- Can this help us understand anything about the foundational principles of ML?
- What technology transfer could happen?

And any side projects…

- Where is my data coming from? How is it collected and stored?
- Is there a more transparent or ‘safe’ way to do this?
- Where could bias enter the dataset or model performance?
- What guarantees can I provide on model performance?
- How will the systems I’m developing be deployed? Will the benefits and harms be equitably distributed?

Treat Data Science Scientifically

01

Research Goal
I want to identify Higgs bosons at the ATLAS detector

02

Hypothesis
I think the angle between the decay products is an informative signal

03

Collect Data
Find a labeled data set with the necessary information (ideally one used before)

04

Test the Hypothesis
Train one model (that you've identified beforehand) using the data

05

Analyze Results
Is this model better than existing systems (including uncertainty!)

06

Reach a Conclusion
I should or should not use this model because of X, Y, and Z

07

Refine + Repeat
Momentum of decay products may be informative OR another architecture may work better

Be Mindful About Your Data

- How much data is available and does each entry have the same information?
- Do you have examples of all data classes/ranges?
- Are the available labels related to the decision you want to make?
- Are classes and inputs balanced and normalized?
- Are there patterns in your data you don’t want the model to exploit?
- Is there noise in your label creation or distribution?
- Are there patterns in your data you don’t want the model to exploit?

Consider All Steps of the Pipeline

01

Data Collection

What population is sampled? How? What format is the data collected in?

02

Data Processing

What cleaning is applied? How does it affect distributions? How are null values handled?

03

Model Building

What variables are used? How do they related to the outcome? What statistical assumptions underlie the model? What incentive are we considering?

04

Model Evaluation

What metrics are used? How do we check for bias? How do we check for robustness?

05

Testing

What theory or model of the world are we comparing to?

Science to Inform ML

Unlike many ML application domains, with physical sciences we have an (approximately) robust underlying mathematical model

Explainability

We know some information a model should learn and have interpretable bases for some problem classes

De-biasing

We often know true confounding variables and correlations so can meaningfully evaluate debiasing techniques

Physics of ML

By studying learning as a stochastic process we can optimize models and training

Scientific Principles

Core experiment design techniques like uncertainty quantification and blinding can lend robustness

Outreach

Advocacy

Use your voice, institutional power, and collective action to work against unjust or unsafe uses of AI

Technical Literacy

Work with your communities to help them develop the knowledge necessary meaningfully consent to sociotechnical systems and understand possible recourse.

Legislation

Share your scientific expertise with policy makers and champion meaningful regulations

Data analysis and model building are big responsibilities!

savannah.thais@gmail.com

@basicsciencesav

Resources (Science Related)

- "Physicists Must Engage with AI Ethics, Now", APS.org
- "Fighting Algorithmic Bias in Artificial Intelligence", Physics World
- "Artificial Intelligence: The Only Way Forward is Ethics", CERN News
- "To Make AI Fairer, Physicists Peer Inside Its Black Box", Wired
- "The bots are not as fair minded as the seem", Physics World Podcast
- "Developing Algorithms That Might One Day Be Used Against You", Gizmodo
- "AI in the Sky: Implications and Challenges for Artificial Intelligence in Astrophysics and Society", Brian Nord for NOAO/Steward Observatory Joint Colloquium Series
- Ethical implications for computational research and the roles of scientists, Snowmass LOI
- LSSTC Data Science Fellowship Session on AI Ethics
- Panel on Data Science Education, Physics, and Ethics, APS GDS

Resources (General)

- AI Now
- Alan Turing Institute
- Algorithmic Justice League
- Berkman Klein Center
- Center for Democracy and Technology
- Center for Internet and Technology Policy
- Data & Society
- Data for Black Lives
- Montreal AI Ethics Institute
- Stanford Center for Human-Centered AI
- The Surveillance Technology Oversight Project
- Radical AI Network
- Resistance AI

Thank you and looking forward to an interesting discussion!

savannah.thais@gmail.com

@basicsciencesav