AI Ethics & Responsible Data Science for Physicists

Savannah Thais, Columbia University

Why are we talking about AI Ethics at a Physics Summer School?

Let’s start with two questions for you…

LANGDON WINNER

Do Artifacts Have Politics?

IN CONTROVERSIES ABOUT TECHNOLOGY AND SOCIETY, there is no idea more provocative than the notion that technical things have political qualities. At issue is the claim that the machines, structures, and systems of modern material culture can be accurately judged not only for their contributions of efficiency and productivity, not merely for their positive and negative environmental side effects, but also for the ways in which they can embody specific forms of power and authority. Since ideas of this kind have a persistent and troubling presence in

Does AI have politics?

How do we contextualize and interrogate a piece of technology?

Table 1. Failure Taxonomy

Impossible Tasks	Conceptually Impossible Practically Impossible
Engineering Failures	Design Failures Implementation Failures Missing Safety Features
Post-Deployment Failures	Robustness Issues Failure under Adversarial Attacks Unanticipated Interactions
Communication Failures	Falsified or Overstated Capabilities Misrepresented Capabilities

Does AI Function?

What beliefs do we have about the abilities and reliability of AI?

The Fallacy of AI Functionality, Raji et al

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**

How does this manifest in the real world?

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?

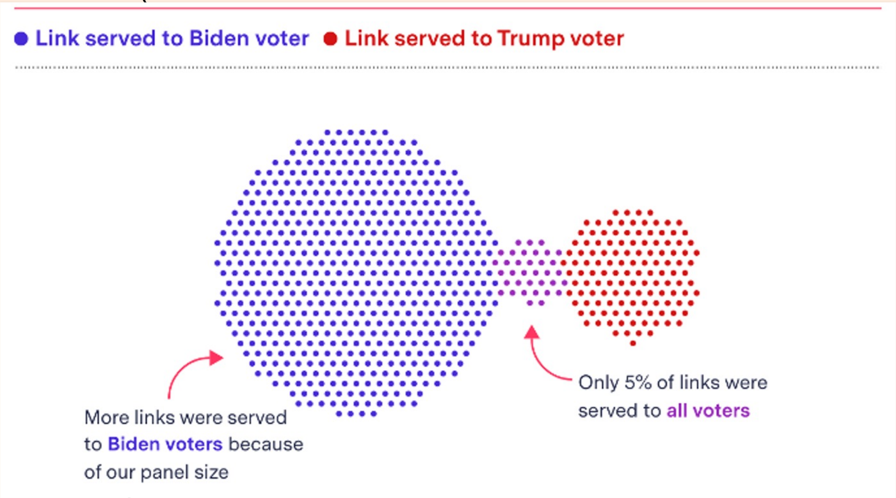
Downstream & Diffuse Impacts

What is changed or lost by what we build?

Data Collection & Storage

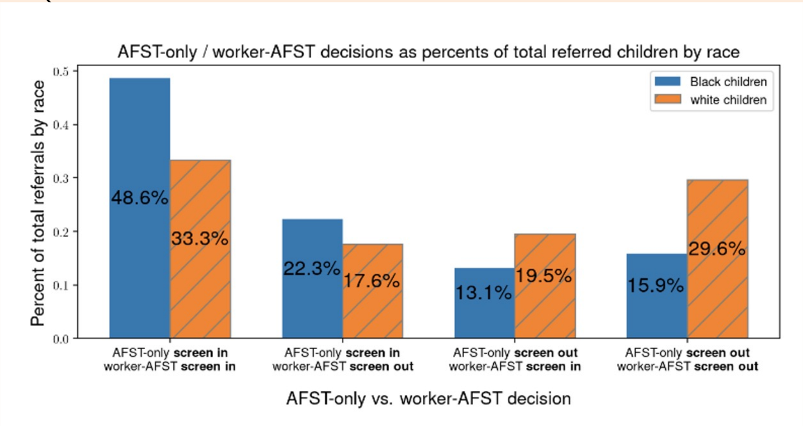
- Data labeling companies exploit workers and political strife in the global south to maximize profits
- Non-profit Crisis Text Line shared user conversation data with for-profit spinoff designed to 'improve customer service'
- Data brokerage firms indiscriminately sell aggregated, 'anonymized' location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

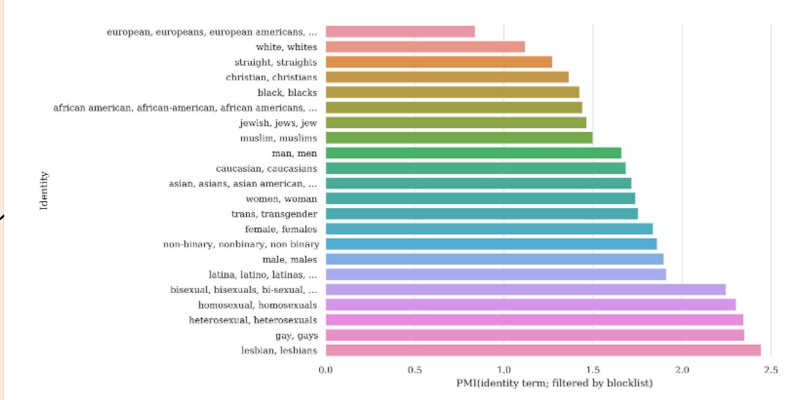
Task Design & Learning Incentives



- 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)

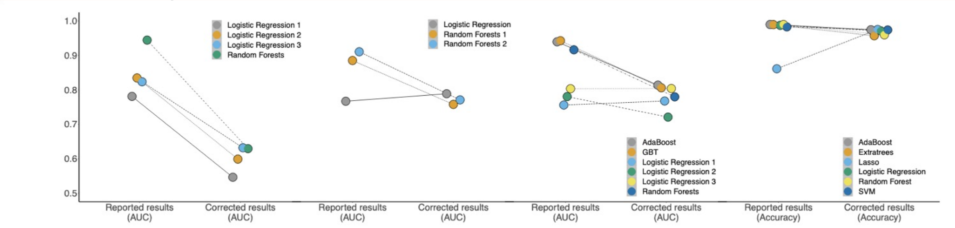
Model Bias & Fairness





- 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

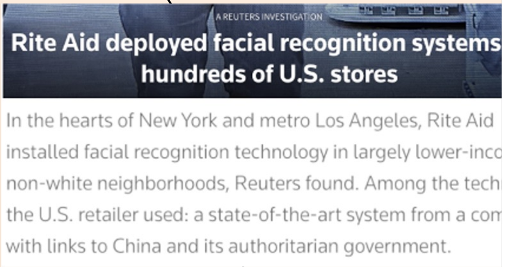
Model Robustness & Reliability



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

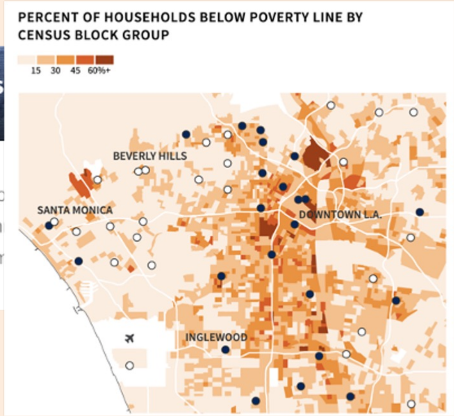
Deployment & Outcomes

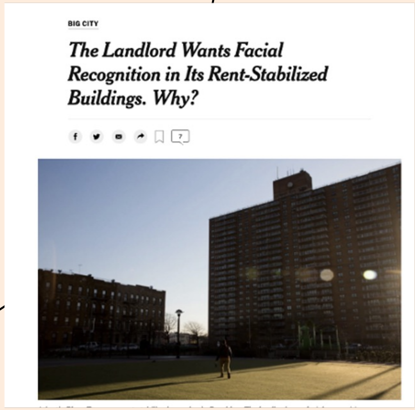


A REUTERS INVESTIGATION

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

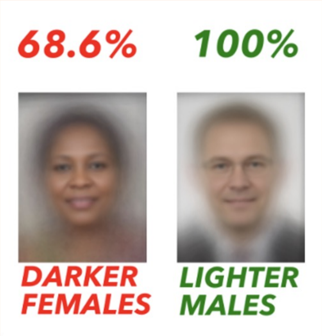
In the hearts of New York and metro Los Angeles, Rite Aid installed facial recognition technology in largely lower-inco non-white neighborhoods, Reuters found. Among the tech the U.S. retailer used: a state-of-the-art system from a com with links to China and its authoritarian government.





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 physicists do to address these issues?

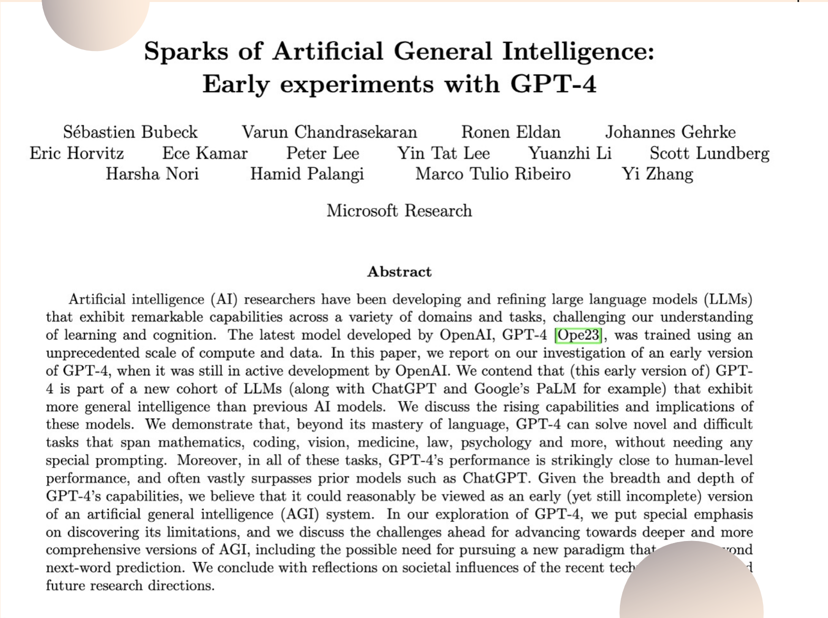
The Empirical Gap

What kind of science is AI/ML? Is it a science?

- There is a rich area of research around provable results in ML
  - E.g. statistical limitations, scaling laws, performance of optimizers, etc
- However, recent results in ML/AI tend towards 'observational science'
  - E.g. emergernt behaviors, sparks of AGI, theory of mind, etc

An odd paradigm has emerged where we have **no real fundamental understanding of something we have built**

- I believe we can begin to close this gap by **changing how we evaluate and contextualize AI/ML**



Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research

Abstract

Artificial intelligence (AI) researchers have been developing and refining large language models (LLMs) that exhibit remarkable capabilities across a variety of domains and tasks, challenging our understanding of learning and cognition. The latest model developed by OpenAI, GPT-4 [Ope23], was trained using an unprecedented scale of compute and data. In this paper, we report on our investigation of an early version of GPT-4, when it was still in active development by OpenAI. We contend that (this early version of) GPT-4 is part of a new cohort of LLMs (along with ChatGPT and Google's PaLM for example) that exhibit more general intelligence than previous AI models. We discuss the rising capabilities and implications of these models. We demonstrate that, beyond its mastery of language, GPT-4 can solve novel and difficult tasks that span mathematics, coding, vision, medicine, law, psychology and more, without needing any special prompting. Moreover, in all of these tasks, GPT-4's performance is strikingly close to human-level performance, and often vastly surpasses prior models such as ChatGPT. Given the breadth and depth of GPT-4's capabilities, we believe that it could reasonably be viewed as an early (yet still incomplete) version of an artificial general intelligence (AGI) system. In our exploration of GPT-4, we put special emphasis on discovering its limitations, and we discuss the challenges ahead for advancing towards deeper and more comprehensive versions of AGI, including the possible need for pursuing a new paradigm that ... next-word prediction. We conclude with reflections on societal influences of the recent tech... future research directions.

Contextualize your physics 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?

Physics to Inform ML

Unlike many ML application domains, with physics 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

Use your quantitative skills and model building intuition!

A Scientific Framework for AI Experiments

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

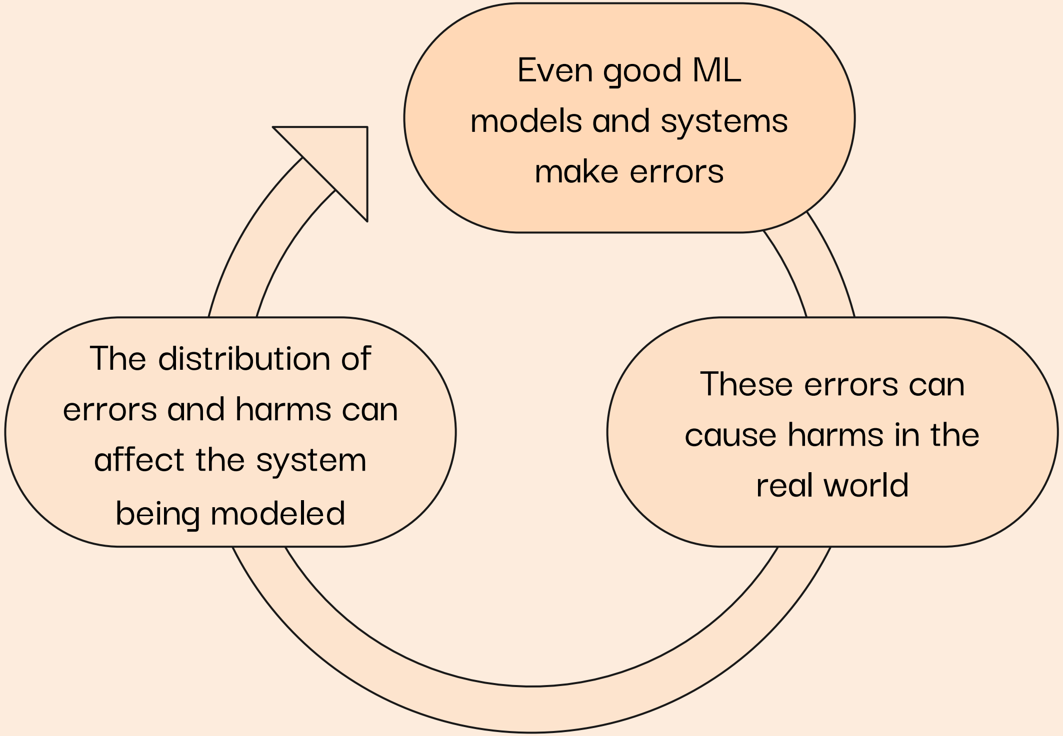
07

Refine + Repeat

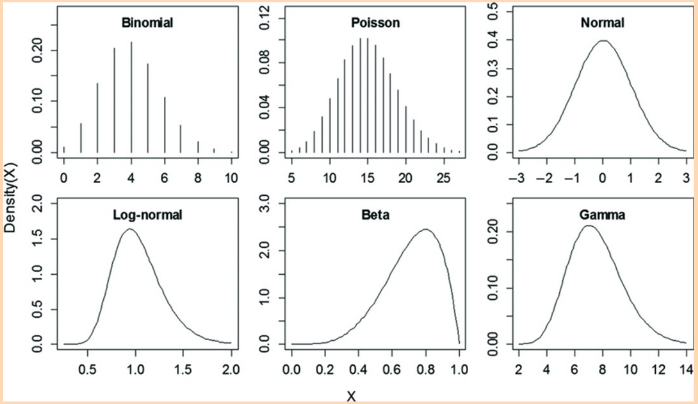
Momentum of decay products may be informative OR another architecture may work better

A Scientific Framework for AI Impacts

- Errors and harms distributed differently (though harm usually depends on error)
- **Understanding these distributions (across different variables) is then key to much of AI ethics and responsible data science work**
- Additional complexity: like in physics these systems and models are often stacked



Characterizing the Distributions



So we get similar many similar questions to those we have in a physics analysis:

- What data selections or biases contributed to the distribution?
- How do we ensure the distribution is modeled correctly? What kinds of tests can we use?
- What should the distribution look like (or what do we want it to look like given certain assumptions/goals)?
- If anything is incorrect or needs to be changed, what knobs can we turn?

In physics these questions are often hard but answerable in some other applications they may not be!

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

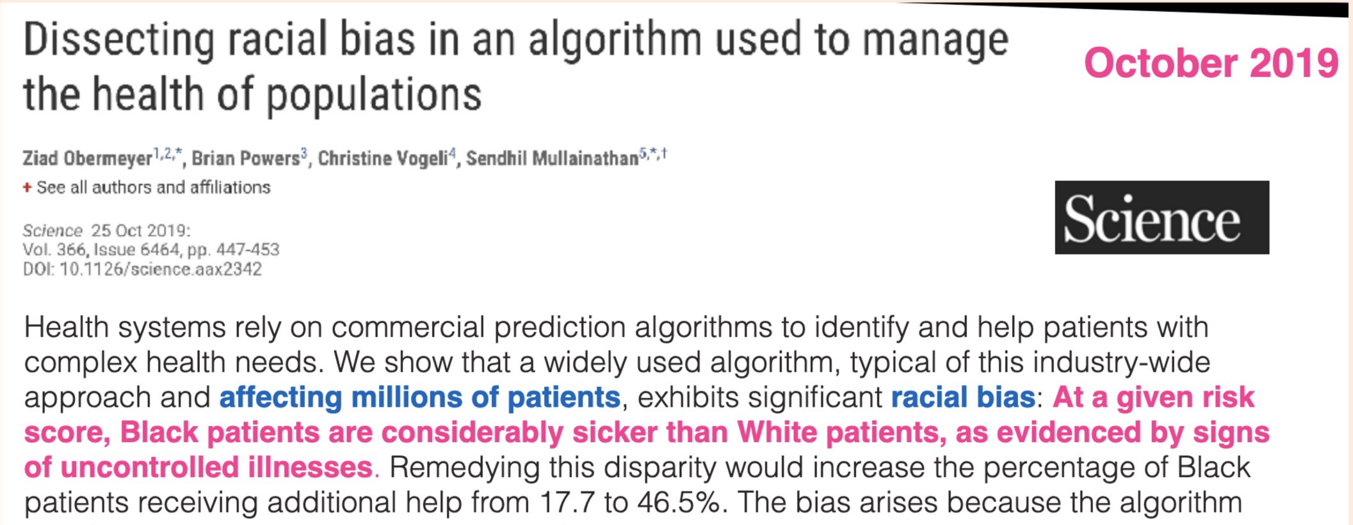
Testing

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

So let's take a mindset of system/model auditors

Pretend we're reviewing a physics paper…we want to make sure that the analysis is rigorous and statistically sound (or understand where it went wrong)

Racial Bias in Healthcare Risk Assessment



Dissecting racial bias in an algorithm used to manage the health of populations

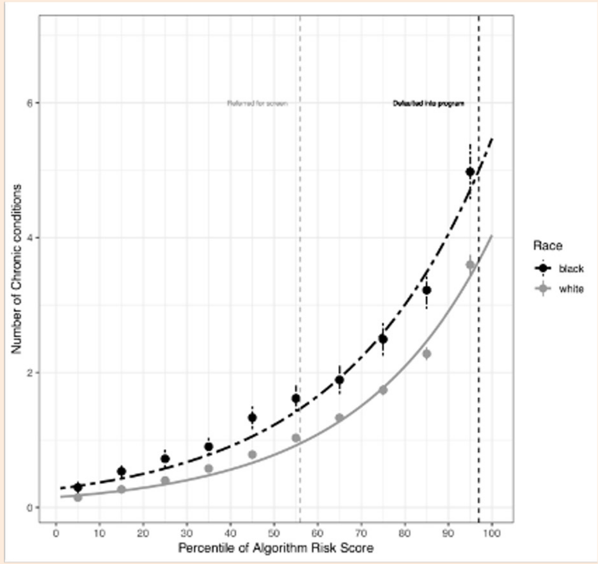
October 2019

Ziad Obermeyer[1,2,*], Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5,*,†]
+ See all authors and affiliations

Science

Science 25 Oct 2019:
Vol. 366, Issue 6464, pp. 447-453
DOI: 10.1126/science.aax2342

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and **affecting millions of patients**, exhibits significant **racial bias**: **At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses**. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm



Understanding the Distributions

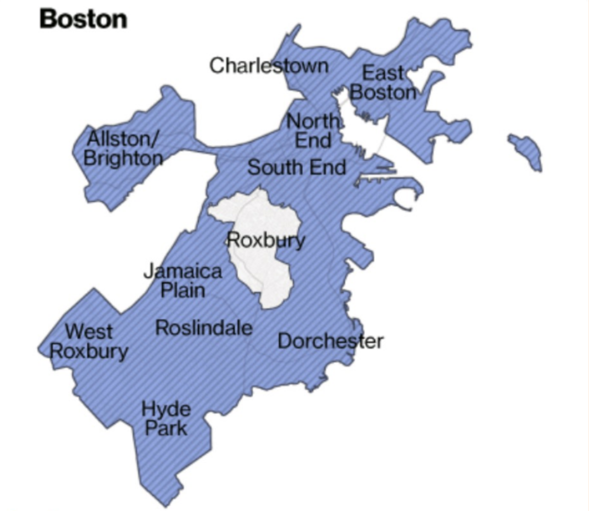
1. Determine an appropriate measure of model validity
2. Check the performance (closure) of the model across all areas where it will be applied (each race)
3. See that it underperforms in some phase spaces…why could that be?
4. Interrogate model variables and labels to look for statistical sources of bias
5. Fix it…in this case by defining a better training label and potentially de-biasing historical data and/or accounting for uncertainty

Racial Bias in Same Day Delivery

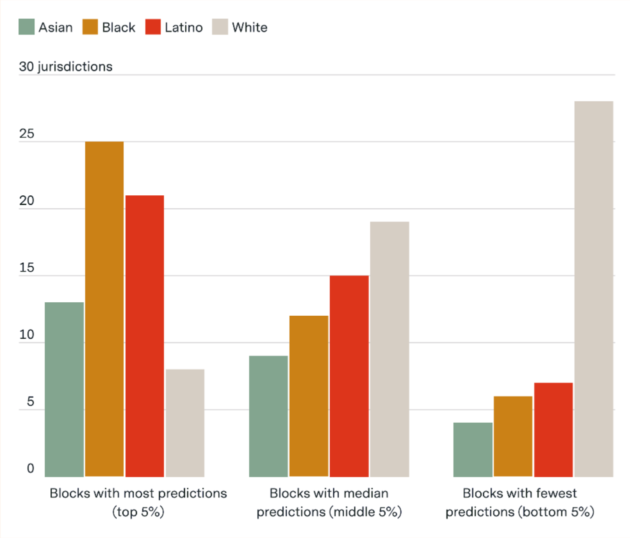
Amazon Doesn't Consider the Race of Its Customers. Should It?

"… In six major same-day delivery cities, however, **the service area excludes predominantly black ZIP codes** to varying degrees, according to a Bloomberg analysis that compared Amazon same-day delivery areas with U.S. Census Bureau data."

https://www.bloomberg.com/graphics/2016-amazon-same-day/

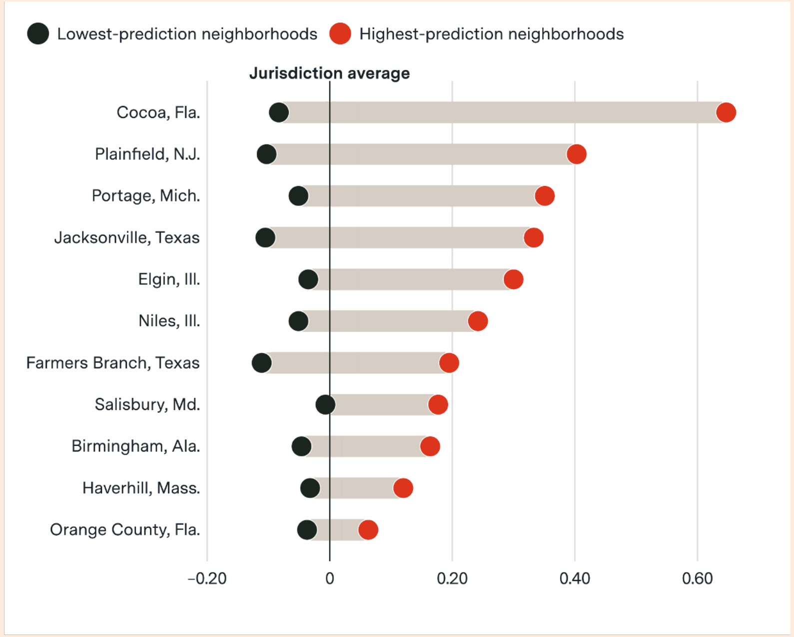


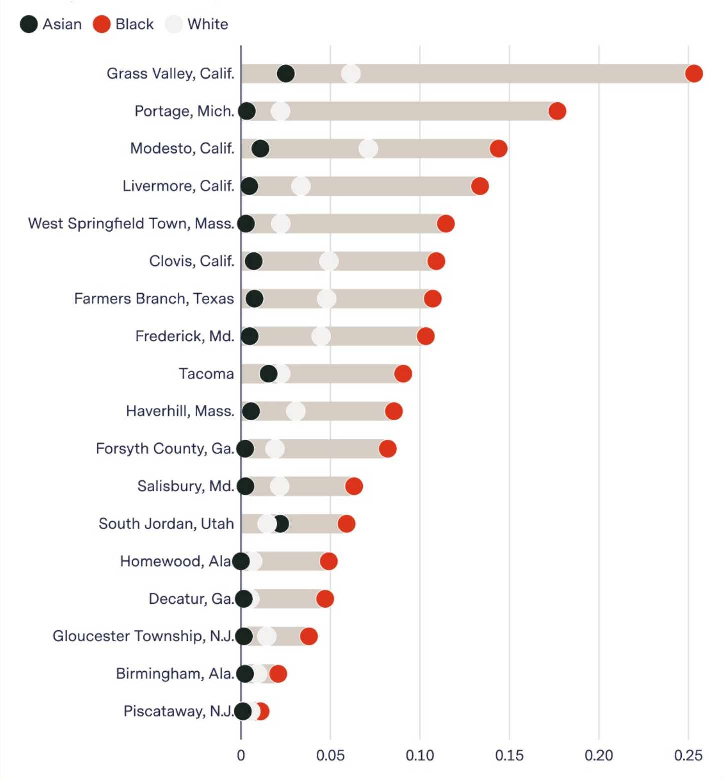
Predictive Policing



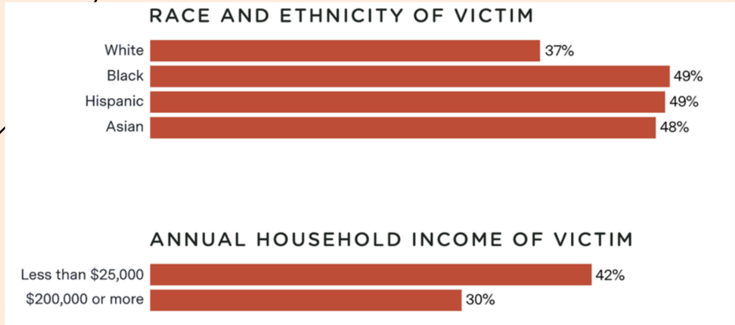
From excellent Gizmodo report

Predictive Policing





Jurisdiction	Type of Agreement (Year)	Unlawful/Biased Practices	Timeframe of Unlawful/Biased Practices	Predictive Policing System (Year)	Evidence of Biased Data in Predictive Policing
Maricopa County (AZ)	Federal District Court Order (2013),[154] Federal District Court Order (2016),[155] DOJ Settlement Agreement (2015)[156]	Unconstitutional and racially biased stops, searches, and arrests; unconstitutional lengthening of stops; unlawful stops; unlawful retaliation against people who made complaints or criticized MCSO	2007–2011; 2014–2017	PredPol (Mesa, 2016–present); HunchLab (Peoria, 2015–present); RTMDx (Glendale, 2012 pilot); BJA-funded pilot (Tempe, 2014)	Unclear. Mesa shares data with MCSO, and PredPol uses crime data.



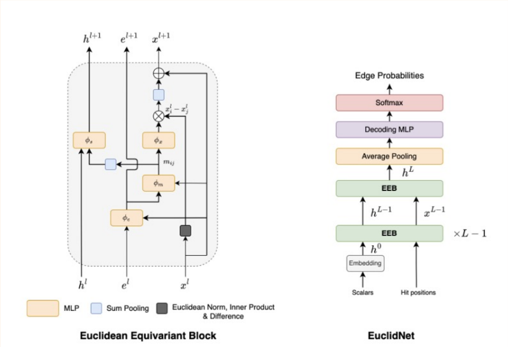
Understanding the Distributions

- What is an appropriate measure of performance here? (not all crimes are reported)
- Where could the apparent bias in prediction arise from? How could we understand if it's accurate?
- How does the system itself affect the phase space the developers are trying to measure
- How could we start to address these concerns?

Some of My Current Research

Equivariant Graph Neural Networks for Charged Particle Tracking

Daniel Murnane[1], Savannah Thais[2], Ameya Thete[3]
[1] Scientific Data Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA
[2] Data Science Institute, Columbia University, New York, NY 10027, USA
[3] Department of Physics, BITS, Pilani – KK Birla Goa Campus, Zuarinagar, Goa 403726, India
E-mail: dtmurnane@lbl.gov



Euclidean Equivariant Block

EuclidNet

Equivariance Is Not All You Need

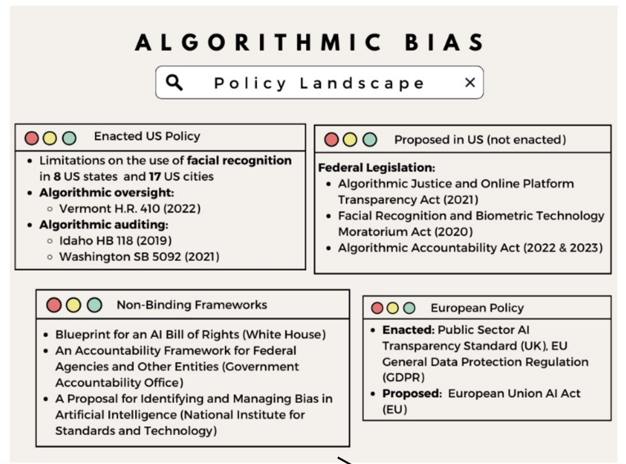
Characterizing the Utility of Equivariant Graph Neural Networks for Particle Physics Tasks

Savannah Thais[1] and Daniel Murnane[2]
1 Data Science Institute, Columbia University
2 Computing Science Research, Lawrence Berkeley National Lab

Abstract: Incorporating inductive biases into ML models is an active area of ML research, especially when ML models are applied to data about the physical world. Equivariant Graph Neural Networks (GNNs) have recently become a popular method for learning from physics data because they directly incorporate the symmetries of the underlying physical system. Drawing from the relevant literature around group equivariant networks, this paper presents a comprehensive evaluation of the proposed benefits of equivariant GNNs by using real-world particle physics reconstruction tasks as an evaluation test-bed. We demonstrate that many of the theoretical benefits generally associated with equivariant networks may not hold for realistic systems and introduce compelling directions for future research that will benefit both the scientific theory of ML and physics applications.

Algorithmic Bias: Looking Beyond Data Bias to Ensure Algorithmic Accountability and Equity

Savannah Thais,[1,*] Hannah Shumway,[2] and Austin Iglesias Saragih[3]



Misrepresented Technological Solutions in Imagined Futures: The Origins and Dangers of AI Hype

Anonymous Author(s)

CYsyphus – The Cybersecurity Policy Recommendation Tool

COLUMBIA | SIPA School of International and Public Affairs
Jason Healey, Jennifer Lake, Savannah Thais, Angela Woodall, Nitansha Bansal, Yunjie Qian, Shubham Kaushal
School of International and Public Affairs at Columbia University
Arnold A. Saltzman Institute of War and Peace Studies at Columbia University
Data Science Institute at Columbia University

CYsyphus (pronounced "SIGH-si-fis") is a **decision-support tool**, that mines the wisdom from past cyber reports and presents them in an easy-to-search online database. With the long-term vision to **capture and code every cybersecurity recommendation made in the English language**, the tool aims **to reduce, by an order of magnitude, the amount of time it takes to ideate and create policy-relevant recommendations.**

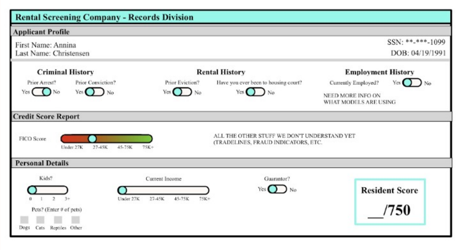
Understanding Historical, Socio-Economic, and Policy Contributions to COVID-19 Health Inequities

Savannah Thais*
Princeton University, Community Insight and Impact
savannah.thais@gmail.com

Allie Saizan
Community Insight and Impact

Shaine Leibowitz
Community Insight and Impact

Ashay Singh
Community Insight and Impact



Resources (Physics 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!
Let’s discuss!

savannah.thais@gmail.com

@basicsciencesav