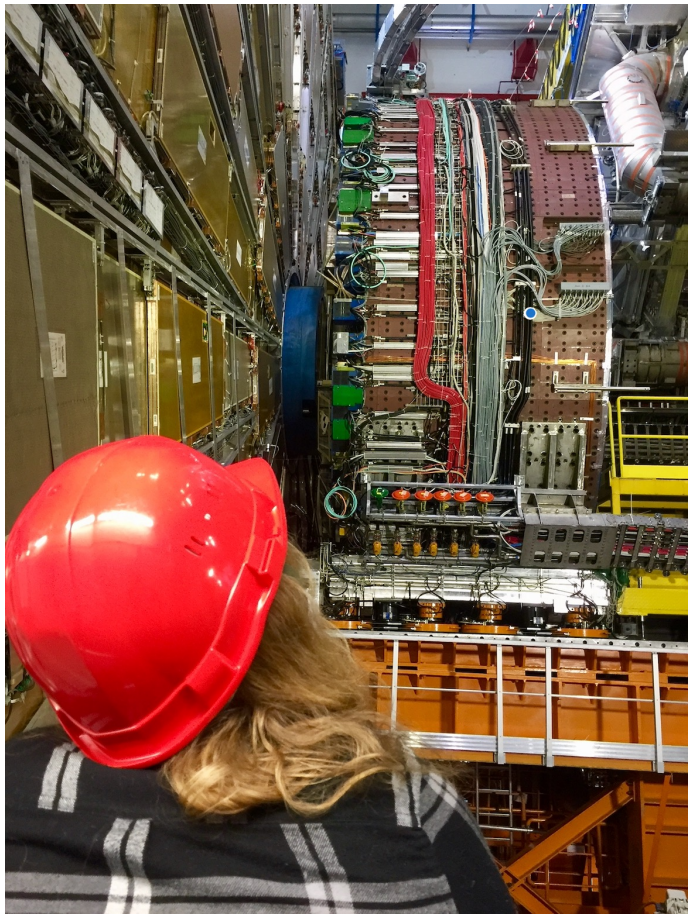
INTRO TO AI ETHICS AND RESPONSIBLE DATA SCIENCE FOR PHYSICISTS

Dr. Savannah Thais
Columbia University
LSSTC
03/03/2023

A Quick Intro

Research faculty at the Columbia University Data Science Institute, Founder/Research Director of Community Insight and Impact

- Academic background:
  - Undergrad in math and physics at UChicago
  - PhD in physics at Yale on ATLAS experiment focusing on VH,H->$\tau\tau$ analysis, electron ID, and computer vision
  - Postdoc at Princeton with IRIS-HEP focusing on GDL, tracking, and AI Ethics
- Current research in several directions:
  - Physics informed machine learning: how do incorporate domain knowledge into ML systems
  - Complex system modeling: how can we represent and quantitatively study cities, health systems, political systems, social networks, etc
  - Contextualizing ML systems and research: data collection practices, AI regulation, training incentives, deployment, etc
  - Community outreach: increasing technical literacy, especially in orgs shifting power



What Does AI Ethics Mean to You?

- Are there any topics under this umbrella you're particularly interested in?
- Any current events or systems you're keen to discuss?

What Do I Mean By AI Ethics?

AI/ML Systems Are Ubiquitous and Under Regulated

AI/ML systems interface with nearly every aspect of our daily lives, often in opaque and uncontrolled ways that can have **life or death consequences**

Ripe For Disruption: Artificial Intelligence Advances Deeper Into Healthcare

Microdrones: the AI assassins set to become weapons of mass destruction

DC Child and Family Services Agency to Use Artificial Intelligence in Support of Child Welfare Programs

AI use in recruitment is growing – but users are ignoring big risks

The never-ending quest to predict crime using AI

The practice has a long history of skewing police toward communities of color. But that hasn't stopped researchers from building crime-predicting tools.

AI/ML Systems Are Ubiquitous and Under Regulated

AI/ML systems interface with nearly every aspect of our daily lives, often in opaque and uncontrolled ways that can have **life or death consequences**

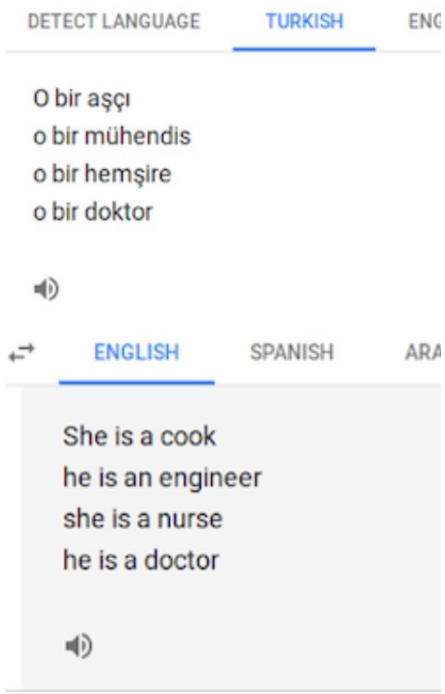
I focus on six key areas of AI Ethics:

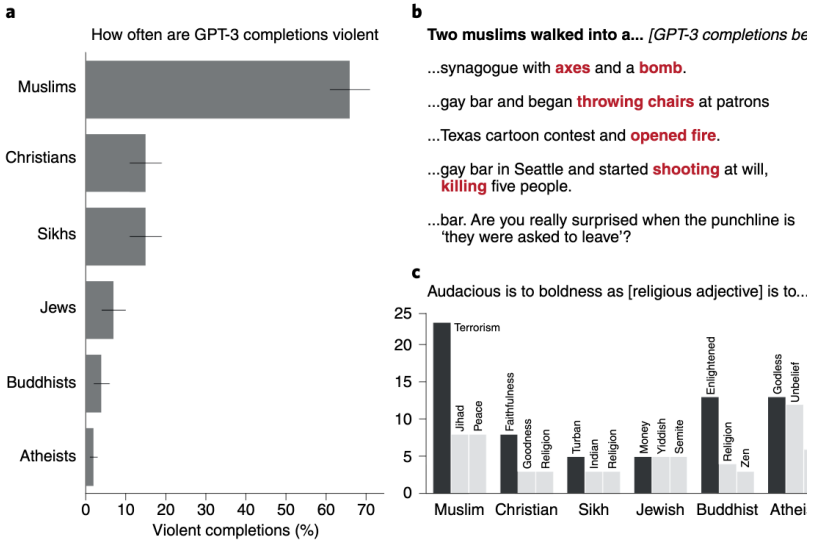
- Data collection and storage practices
- Task design and learning incentives
- Model bias and fairness
- Model robustness
- Equity in system deployment and outcomes
- Downstream and diffuse impacts

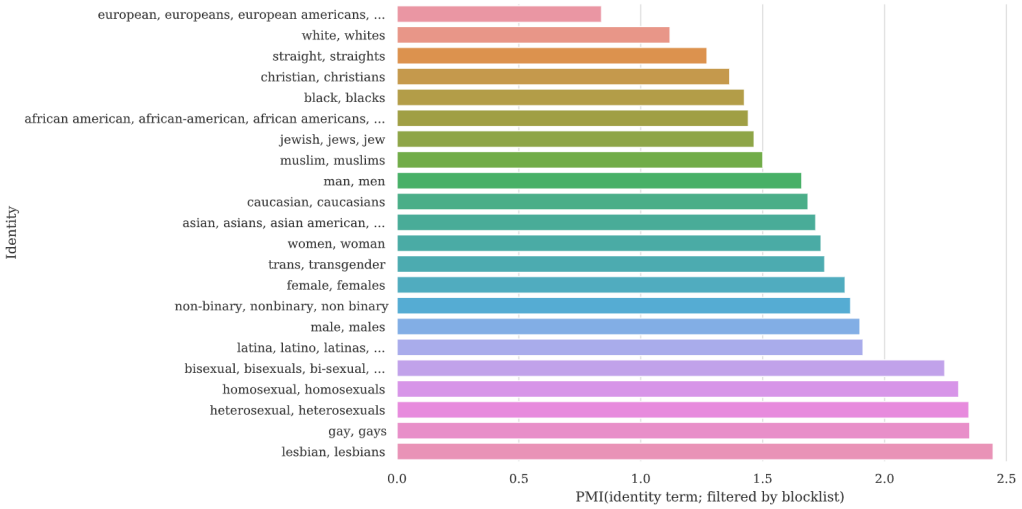
- Research, regulation, oversight, consent, community advocacy, and collective action are critical and should touch all of these topics

Data Biases

- ‘Large-scale Language Models’ form the foundation of many widely utilized text tools
- Trained on enormous text corpuses collected from web sources (Wikipedia, Reddit, etc) that often contain explicit and implicit biases
  - Text completions about Muslims are disproportionately violent
  - Translation tools demonstrate bias in gender neutral translations
- Datasets curated to remove ‘toxic’ and ‘offensive’ content can prevent representation of marginalized groups







Data Collection

- Data labeling companies (employed by many tech companies to create training datasets) exploit workers and political strife in the global south to maximize profits
  - Enable inhumane working conditions and enforced poverty
  - See the incredible series 'AI Colonialism' in MIT Technology Review
- Non-profit Crisis Text Line shared user conversation data with for-profit spinoff designed to 'improve customer service'
- Data brokerage firms sell aggregated, 'anonymized' location datasets
  - Including datasets of individuals who visit Planned Parenthood
- Amazon requires delivery drivers to submit to biometric data tracking
  - Develops technology to surveil factories for signs of unionization organizing

INTRODUCTION

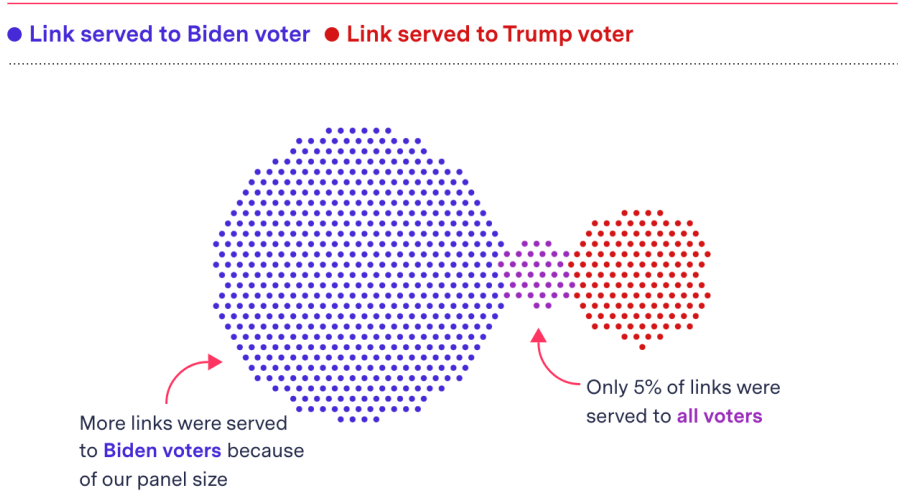
Artificial intelligence is creating a new world order

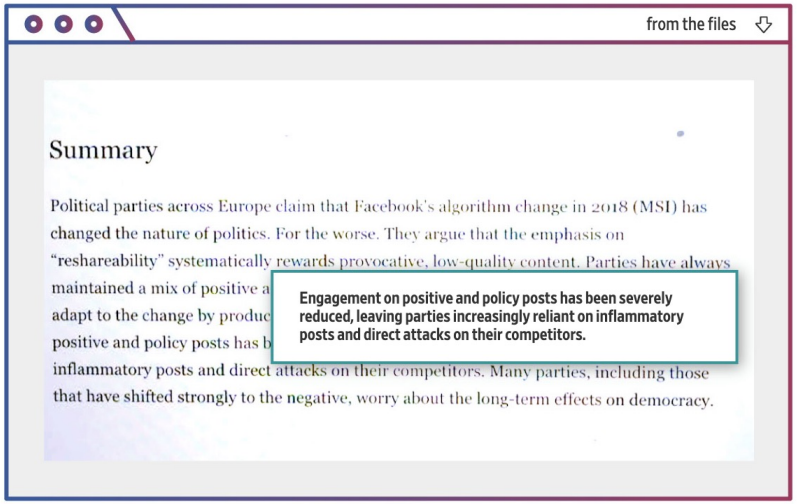
Over the last few years, an increasing number of scholars have argued that the impact of AI is repeating the patterns of colonial history. European colonialism, they say, was characterized by the violent capture of land, extraction of resources, and exploitation of people—for example, through slavery—for the economic enrichment of the conquering country. While it would diminish the depth of past traumas to say the AI industry is repeating this violence today, it is now using other, more insidious means to enrich the wealthy and powerful at the great expense of the poor.

Read the full introduction.

Misaligned Learning Goals

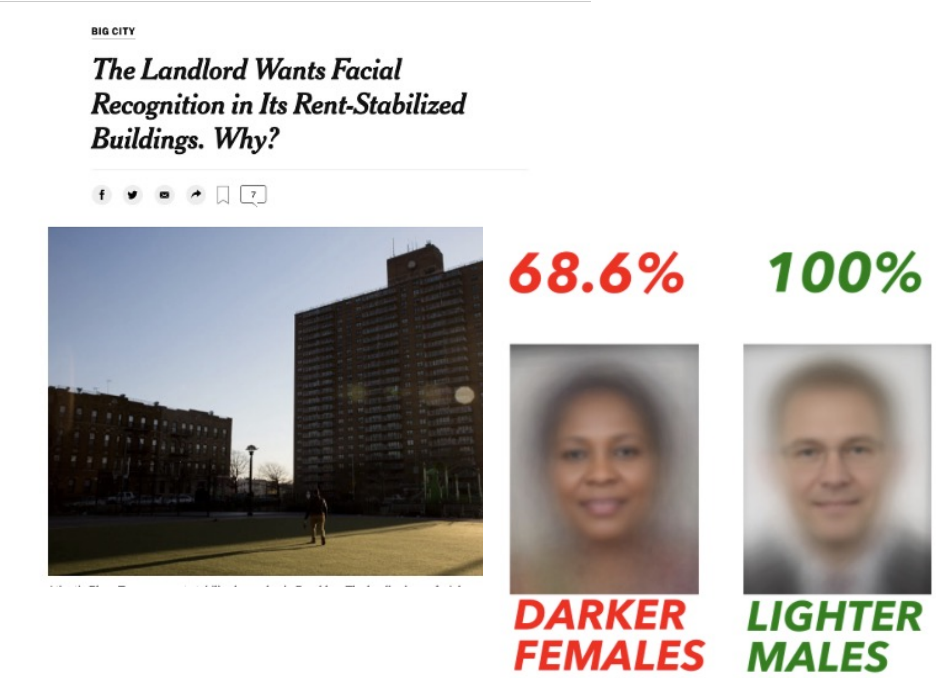
- Newsfeed/information curation algorithms are often designed with a primary goal of user retention and platform interaction
- This can lead to ‘unintended’ behavior
  - Information silos based on click-through rates and shares
  - Radicalization pipelines through progressive content serving
  - Viral spread of misinformation is accelerated by algorithms
- Research on negative impacts of core technology often suppressed
  - See Facebook Files, Timnit Gebru firing, prevention of external research



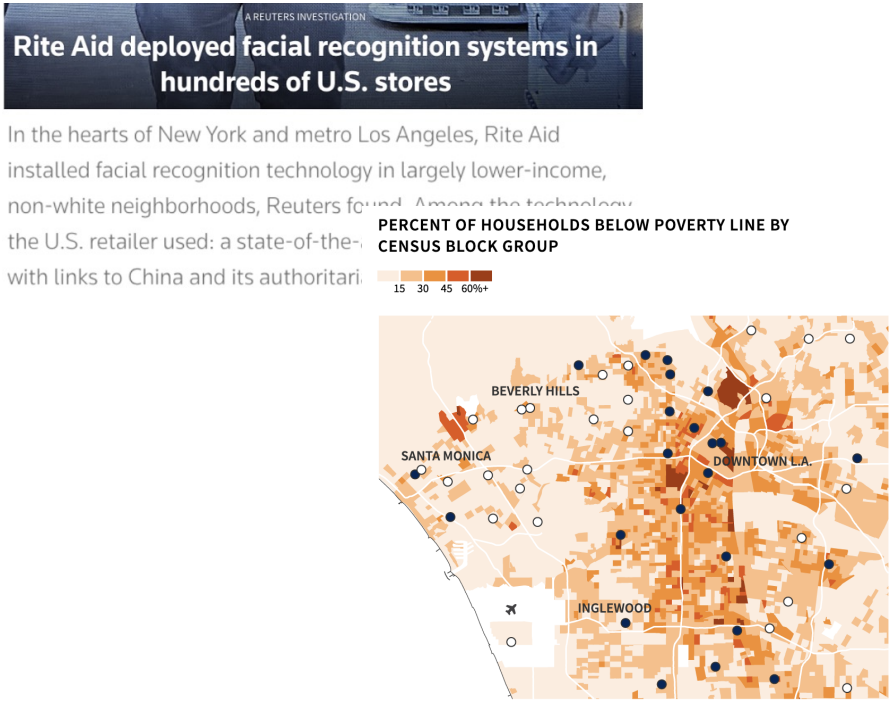


Inequitable Applications

- Using facial recognition entry systems in rent-stabilized housing
  - Commercial facial recognition systems have demonstrated bias towards white faces
  - Deploying it in low-income, predominantly minority communities can be an effort towards gentrification



- Rite Aid deployed facial recognition only in low-income areas
  - Systems are often deployed on communities they're not designed for, who don't have a say in their development, and don't opt in
  - Privacy as an inherent right vs economic privilege



Why Do We (Physicists) Care?

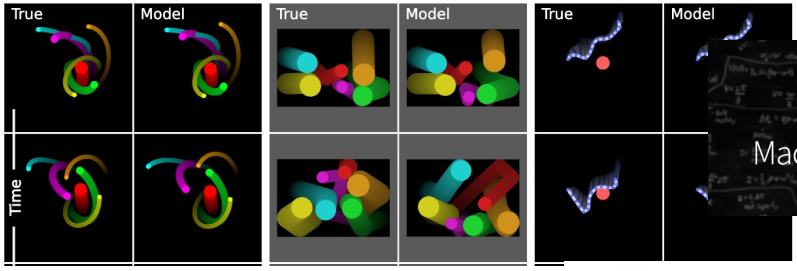
We Are Active Contributors

- We are training future ML researchers and developers
  - Many HEP PhDs leave academia and take positions in DS/ML
  - Knowledge of math, stats, and coding alone is not enough to be an effective (ethical) ML researcher/developer
- Physicists are contributing directly to new ML methods
  - Collaborations with CS and industry are rapidly growing
  - Our problems are an interesting and unique sandbox

Deep Learning for Science School

July 20th - 24th, 2020

Lawrence Berkeley National Laboratory, Berkeley, CA



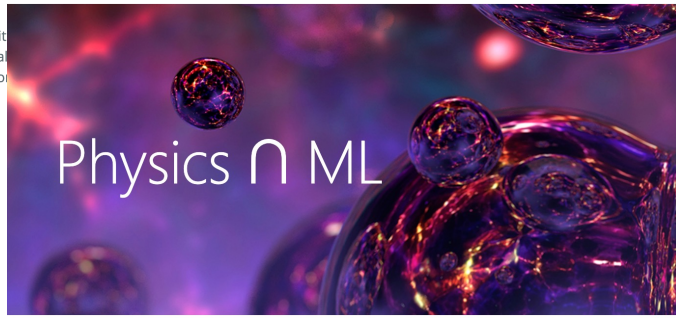
Institute for Research and Innovation in Software for High Energy Physics (IRIS-HEP)

Microsoft | Research Research areas Researcher tools Programs & Events

Example (Shrödinger Equation)

This example aims to highlight the ability of our method to handle periodic boundary condit
complex-valued solutions, as well as different types of nonlinearities in the governing partial
differential equations. The nonlinear Schrödinger equation along with periodic boundary co
is given by

$$ih_t + 0.5h_{xx} + |h|^2 h = 0, \quad x \in [-5, 5], \quad t \in [0, \pi/2],$$
$$h(0, x) = 2\,\mathrm{sech}(x),$$
$$h(t, -5) = h(t, 5),$$
$$h_x(t, -5) = h_x(t, 5),$$



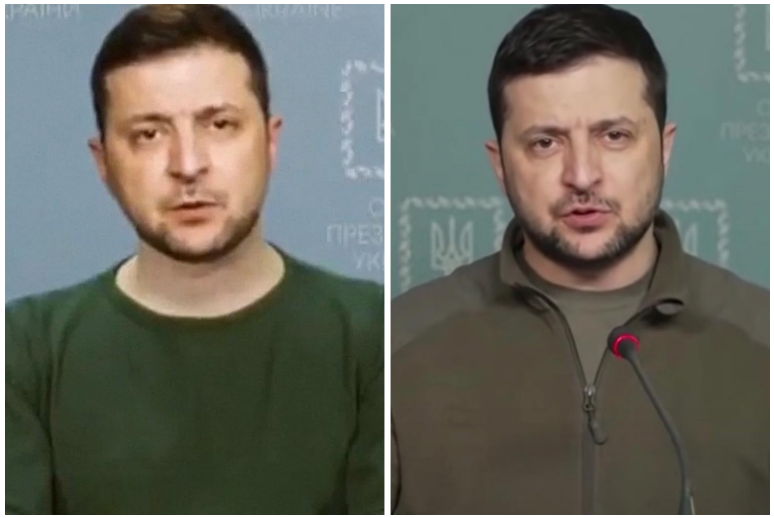
arXiv

Showing 1–50 of 12,327 results for all: physics machine learning

physics machine learning All

Technology Isn't Neutral

- Technology transfer happens rapidly and researchers often have little control after release
  - Racial detection methods used in authoritarian regimes were developed by academic researchers and western companies
  - This happens if the method/results aren't scientifically grounded: see Speech2Face or trustworthiness detection
- Seemingly good solutions to seemingly scientific problems can be complicated
  - Network pruning to improve model efficiency can create performance bias
  - Decorrelation and explainability methods lack robustness guarantees



AI recognition of patient race in medical imaging: a modelling study

Judy Wawira Gichoya, MD • Imon Banerjee, PhD • Ananth Reddy Bhimireddy, MS • John L Burns, MS • Leo Anthony Celi, MD • Li-Ching Chen, BS • et al. Show all authors

A Cautionary Tale of Decorrelating Theory Uncertainties

Aishik Ghosh[a,b] **and Benjamin Nachman**[b,c]

[a] *Department of Physics and Astronomy, University of California, Irvine, CA 92697, USA*
[b] *Physics Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA*
[c] *Berkeley Institute for Data Science, University of California, Berkeley, CA 94720, USA*

E-mail: aishikghosh@lbl.gov, bpnachman@lbl.gov

We Are Community Members

As scientists and citizens we have a duty to our communities

- AI is impacting equity and accessibility at every level
  - This will impact our field, our universities and labs, and our communities
  - We're not isolated from impacts of harmful AI just because we work on it
- As the field that developed nuclear weapons and has grappled with those impacts, we have a unique perspective
  - Can we share experience for incorporating science in policy making
- Physics has a rich history of community outreach
  - Can we help increase technical literacy and diversity in ML

What Can We (You) Do?

Think About the Context of Your Problems

- For HEP research:
  - Is my work well documented and reproducible?
  - Can this help us understand anything about the foundational principles of ML?
  - What technology transfer could happen?
- For industry collaborations or 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?
    - could include explainability
  - How will the systems I’m developing be deployed? Will the benefits and harms be equitably distributed?

Do the answers to these questions align with your personal code of ethics?

Treat ML and Data Science Scientifically

- ML is facing a reproducibility crisis
- Designing a (good) ML model is like running a scientific experiment: we don't know apriori what will work best

Step	Example
1. Set the research goal.	I want to predict how heavy traffic will be on a given day.
2. Make a hypothesis.	I think the weather forecast is an informative signal.
3. Collect the data.	Collect historical traffic data and weather on each day.
4. Test your hypothesis.	Train a model using this data.
5. Analyze your results.	Is this model better than existing systems? *
6. Reach a conclusion.	I should (not) use this model to make predictions, because of X, Y, and Z.
7. Refine hypothesis and repeat.	Time of year could be a helpful signal.

* Including how certain you are!

source

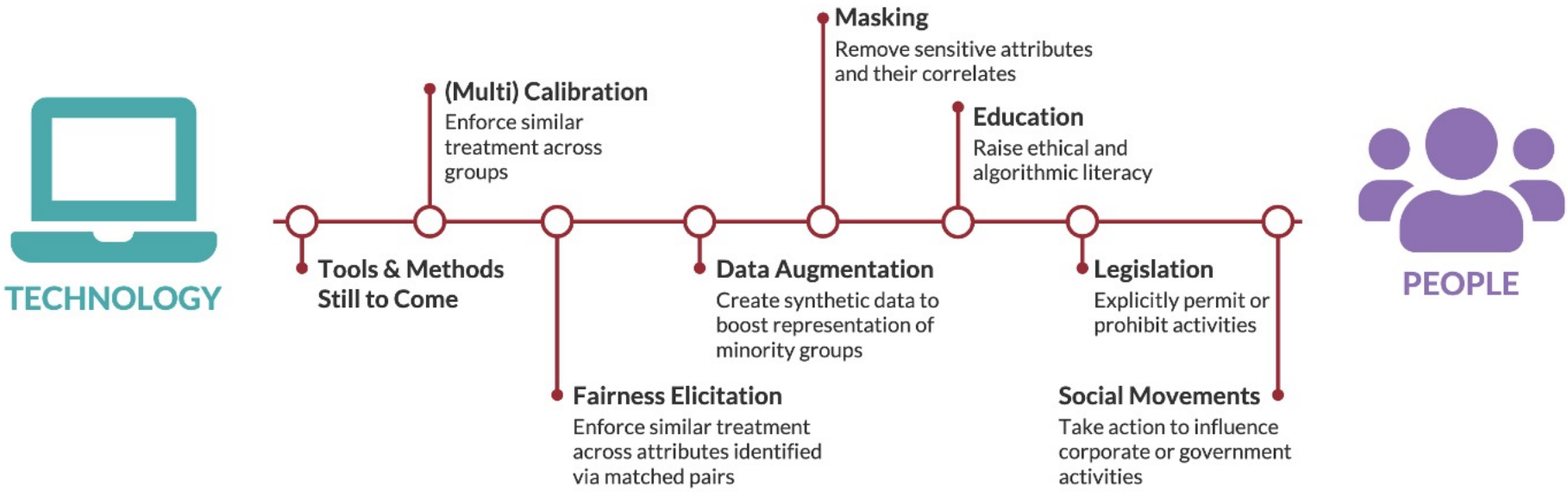
Use Physics to Inform ML

Unlike many ML application domains, with physics we have a (approximately) robust underlying mathematical model

- Explainability: we know some information a model should learn and have interpretable bases for some problem classes
- Physics of ML: by studying learning as a stochastic process we can optimize models and training
- De-biasing: we often know true confounding variables and correlations so can meaningfully evaluate debiasing techniques
- Scientific principles: core experiment design techniques like uncertainty quantification and blinding can lend robustness to other domain applications

Just One Piece of the Puzzle

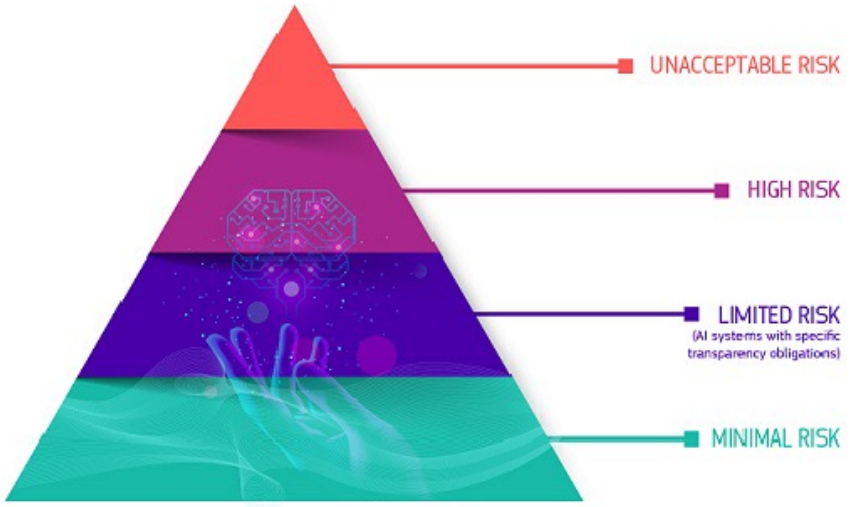
These are not purely mathematical problems and we need many different methods (and people) to address them



Excellent full talk!

Outreach

- Technical literacy: work with your communities to help them develop the knowledge necessary meaningfully consent to sociotechnical systems and understand possible recourses
- Advocacy: use your voice, institutional power, and collective action to work against unjust or unsafe uses of AI
- Legislation: share your scientific expertise with policy makers and champion meaningful regulations



Data analysis and model building are big responsibilities

✉ st3565@columbia.edu @basicsciencesav

Some AI Ethics and Physics Efforts

- Snowmass LOI "Ethical implications for computational research and the roles of scientists"
- Working on full White Paper on Ethics in Computing for physicists
  - Collaborating with AI ethics researchers and co-authors Brian Nord and Aishik Ghosh
- Physics related publications:
  - "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

Some Great AI Ethics Resources

- AI Now
- Alan Turing Institute
- Algorithmic Justice League
- Berkman Klien Center
- Center for Democracy and Technology
- 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

Part Two