AI Ethics + *Fundamental Physics*

Savannah Thais, Columbia University

Some Framing...

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?

IDEAS • TECHNOLOGY

Why Uncontrollable AI Looks More Likely Than Ever

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

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.

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**

Danger of Treating AI as **Magic vs Science**

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**

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

Danger of Treating AI as **Magic vs Science**

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**

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

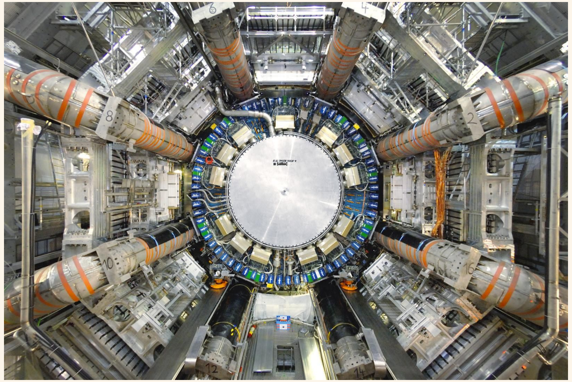
Research: Opportunities

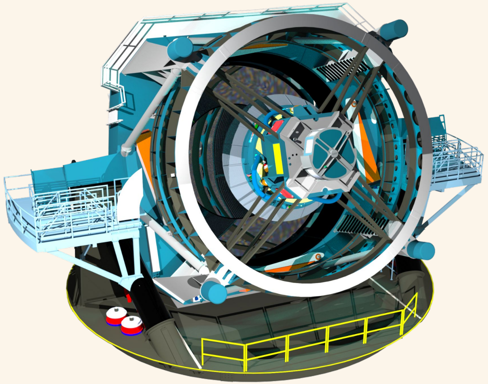
Physics

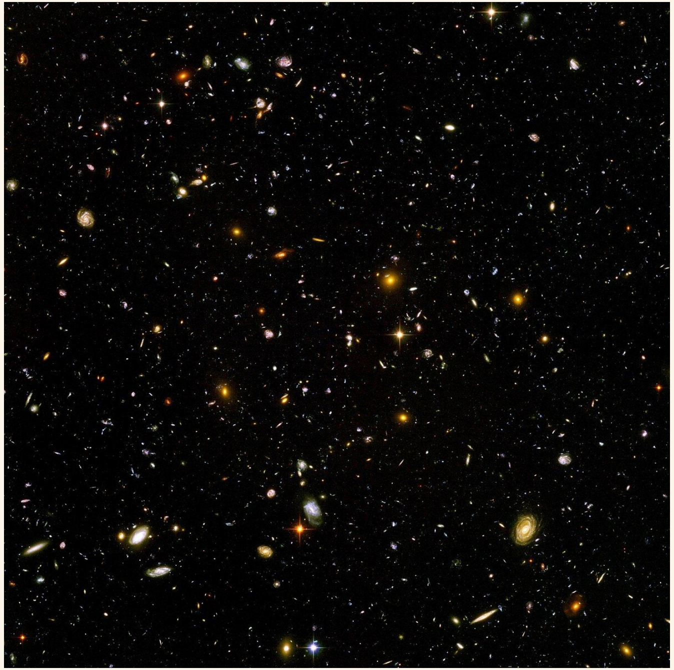
Trustworthy AI

Physics as a **Sandbox**

$$
\begin{gathered}
\mathcal{L}_{\mathrm{StandardModel}} = \\
-\tfrac{1}{2}\partial_\nu g^a_\mu \partial_\nu g^a_\mu - g_s f^{abc}\partial_\mu g^a_\nu g^b_\mu g^c_\nu - \tfrac{1}{4}g_s^2 f^{abc}f^{ade}g^b_\mu g^c_\nu g^d_\mu g^e_\nu + \\
\tfrac{1}{2}ig_s^2(\bar{q}^\sigma_i\gamma^\mu q^\sigma_j)g^a_\mu + \bar{G}^a\partial^2 G^a + g_s f^{abc}\partial_\mu\bar{G}^a G^b g^c_\mu - \partial_\nu W^+_\mu\partial_\nu W^-_\mu - \\
M^2 W^+_\mu W^-_\mu - \tfrac{1}{2}\partial_\nu Z^0_\mu\partial_\nu Z^0_\mu - \tfrac{1}{2c_w^2}M^2 Z^0_\mu Z^0_\mu - \tfrac{1}{2}\partial_\mu A_\nu\partial_\mu A_\nu - \tfrac{1}{2}\partial_\mu H\partial_\mu H - \\
\tfrac{1}{2}m_h^2 H^2 - \partial_\mu\phi^+\partial_\mu\phi^- - M^2\phi^+\phi^- - \tfrac{1}{2}\partial_\mu\phi^0\partial_\mu\phi^0 - \tfrac{1}{2c_w^2}M\phi^0\phi^0 - \beta_h[\tfrac{2M^2}{g^2} + \\
\tfrac{2M}{g}H + \tfrac{1}{2}(H^2 + \phi^0\phi^0 + 2\phi^+\phi^-)] + \tfrac{2M^4}{g^2}\alpha_h - igc_w[\partial_\nu Z^0_\mu(W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - Z^0_\nu(W^+_\mu\partial_\nu W^-_\mu - W^-_\mu\partial_\nu W^+_\mu) + Z^0_\mu(W^+_\nu\partial_\nu W^-_\mu - \\
W^-_\nu\partial_\nu W^+_\mu)] - igs_w[\partial_\nu A_\mu(W^+_\mu W^-_\nu - W^+_\nu W^-_\mu) - A_\nu(W^+_\mu\partial_\nu W^-_\mu - \\
W^-_\mu\partial_\nu W^+_\mu) + A_\mu(W^+_\nu\partial_\nu W^-_\mu - W^-_\nu\partial_\nu W^+_\mu)] - \tfrac{1}{2}g^2 W^+_\mu W^-_\mu W^+_\nu W^-_\nu + \\
\tfrac{1}{2}g^2 W^+_\mu W^-_\nu W^+_\mu W^-_\nu + g^2 c_w^2(Z^0_\mu W^+_\mu Z^0_\nu W^-_\nu - Z^0_\mu Z^0_\mu W^+_\nu W^-_\nu) + \\
g^2 s_w^2(A_\mu W^+_\mu A_\nu W^-_\nu - A_\mu A_\mu W^+_\nu W^-_\nu) + g^2 s_w c_w[A_\mu Z^0_\nu(W^+_\mu W^-_\nu - \\
W^+_\nu W^-_\mu) - 2A_\mu Z^0_\mu W^+_\nu W^-_\nu] - g\alpha[H^3 + H\phi^0\phi^0 + 2H\phi^+\phi^-] - \\
\tfrac{1}{8}g^2\alpha_h[H^4 + (\phi^0)^4 + 4(\phi^+\phi^-)^2 + 4(\phi^0)^2\phi^+\phi^- + 4H^2\phi^+\phi^- + 2(\phi^0)^2 H^2] - \\
gMW^+_\mu W^-_\mu H - \tfrac{1}{2}g\tfrac{M}{c_w^2}Z^0_\mu Z^0_\mu H - \tfrac{1}{2}ig[W^+_\mu(\phi^0\partial_\mu\phi^- - \phi^-\partial_\mu\phi^0) - \\
W^-_\mu(\phi^0\partial_\mu\phi^+ - \phi^+\partial_\mu\phi^0)] + \tfrac{1}{2}g[W^+_\mu(H\partial_\mu\phi^- - \phi^-\partial_\mu H) - W^-_\mu(H\partial_\mu\phi^+ - \\
\phi^+\partial_\mu H)] + \tfrac{1}{2}g\tfrac{1}{c_w}(Z^0_\mu(H\partial_\mu\phi^0 - \phi^0\partial_\mu H) - ig\tfrac{s_w^2}{c_w}MZ^0_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) + \\
igs_w MA_\mu(W^+_\mu\phi^- - W^-_\mu\phi^+) - ig\tfrac{1-2c_w^2}{2c_w}Z^0_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) + \\
igs_w A_\mu(\phi^+\partial_\mu\phi^- - \phi^-\partial_\mu\phi^+) - \tfrac{1}{4}g^2 W^+_\mu W^-_\mu[H^2 + (\phi^0)^2 + 2\phi^+\phi^-] - \\
\tfrac{1}{4}g^2\tfrac{1}{c_w^2}Z^0_\mu Z^0_\mu[H^2 + (\phi^0)^2 + 2(2s_w^2 - 1)^2\phi^+\phi^-] - \tfrac{1}{2}g^2\tfrac{s_w^2}{c_w}Z^0_\mu\phi^0(W^+_\mu\phi^- + \\
W^-_\mu\phi^+) - \tfrac{1}{2}ig^2\tfrac{s_w^2}{c_w}Z^0_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) + \tfrac{1}{2}g^2 s_w A_\mu\phi^0(W^+_\mu\phi^- + \\
W^-_\mu\phi^+) + \tfrac{1}{2}ig^2 s_w A_\mu H(W^+_\mu\phi^- - W^-_\mu\phi^+) - g^2\tfrac{s_w}{c_w}(2c_w^2 - 1)Z^0_\mu A_\mu\phi^+\phi^- - \\
g^1 s_w^2 A_\mu A_\mu\phi^+\phi^- - \bar{e}^\lambda(\gamma\partial + m_e^\lambda)e^\lambda - \bar{\nu}^\lambda\gamma\partial\nu^\lambda - \bar{u}^\lambda_j(\gamma\partial + m_u^\lambda)u^\lambda_j - \\
\bar{d}^\lambda_j(\gamma\partial + m_d^\lambda)d^\lambda_j + igs_w A_\mu[-(\bar{e}^\lambda\gamma^\mu e^\lambda) + \tfrac{2}{3}(\bar{u}^\lambda_j\gamma^\mu u^\lambda_j) - \tfrac{1}{3}(\bar{d}^\lambda_j\gamma^\mu d^\lambda_j)] + \\
\tfrac{ig}{4c_w}Z^0_\mu[(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{e}^\lambda\gamma^\mu(4s_w^2 - 1 - \gamma^5)e^\lambda) + (\bar{u}^\lambda_j\gamma^\mu(\tfrac{4}{3}s_w^2 - \\
1 - \gamma^5)u^\lambda_j) + (\bar{d}^\lambda_j\gamma^\mu(1 - \tfrac{8}{3}s_w^2 - \gamma^5)d^\lambda_j)] + \tfrac{ig}{2\sqrt{2}}W^+_\mu[(\bar{\nu}^\lambda\gamma^\mu(1+\gamma^5)e^\lambda) + \\
(\bar{u}^\lambda_j\gamma^\mu(1+\gamma^5)C_{\lambda\kappa}d^\kappa_j)] + \tfrac{ig}{2\sqrt{2}}W^-_\mu[(\bar{e}^\lambda\gamma^\mu(1+\gamma^5)\nu^\lambda) + (\bar{d}^\kappa_j C^\dagger_{\lambda\kappa}\gamma^\mu(1 + \\
\gamma^5)u^\lambda_j)] + \tfrac{ig}{2\sqrt{2}}\tfrac{m_e^\lambda}{M}[-\phi^+(\bar{\nu}^\lambda(1-\gamma^5)e^\lambda) + \phi^-(\bar{e}^\lambda(1+\gamma^5)\nu^\lambda)] - \\
\tfrac{g}{2}\tfrac{m_e^\lambda}{M}[H(\bar{e}^\lambda e^\lambda) + i\phi^0(\bar{e}^\lambda\gamma^5 e^\lambda)] + \tfrac{ig}{2M\sqrt{2}}\phi^+[-m_d^\kappa(\bar{u}^\lambda_j C_{\lambda\kappa}(1-\gamma^5)d^\kappa_j) + \\
m_u^\lambda(\bar{u}^\lambda_j C_{\lambda\kappa}(1+\gamma^5)d^\kappa_j] + \tfrac{ig}{2M\sqrt{2}}\phi^-[m_d^\lambda(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1+\gamma^5)u^\kappa_j) - m_u^\kappa(\bar{d}^\lambda_j C^\dagger_{\lambda\kappa}(1 - \\
\gamma^5)u^\kappa_j] - \tfrac{g}{2}\tfrac{m_u^\lambda}{M}H(\bar{u}^\lambda_j u^\lambda_j) - \tfrac{g}{2}\tfrac{m_d^\lambda}{M}H(\bar{d}^\lambda_j d^\lambda_j) + \tfrac{ig}{2}\tfrac{m_u^\lambda}{M}\phi^0(\bar{u}^\lambda_j\gamma^5 u^\lambda_j) - \\
\tfrac{ig}{2}\tfrac{m_d^\lambda}{M}\phi^0(\bar{d}^\lambda_j\gamma^5 d^\lambda_j) + \bar{X}^+(\partial^2 - M^2)X^+ + \bar{X}^-(\partial^2 - M^2)X^- + \bar{X}^0(\partial^2 - \\
\tfrac{M^2}{c_w^2})X^0 + \bar{Y}\partial^2 Y + igc_w W^+_\mu(\partial_\mu\bar{X}^0 X^- - \partial_\mu\bar{X}^+ X^0) + igs_w W^+_\mu(\partial_\mu\bar{Y}X^- - \\
\partial_\mu\bar{X}^+ Y) + igc_w W^-_\mu(\partial_\mu\bar{X}^- X^0 - \partial_\mu\bar{X}^0 X^+) + igs_w W^-_\mu(\partial_\mu\bar{X}^- Y - \\
\partial_\mu\bar{Y}X^+) + igc_w Z^0_\mu(\partial_\mu\bar{X}^+ X^+ - \partial_\mu\bar{X}^- X^-) + igs_w A_\mu(\partial_\mu\bar{X}^+ X^+ - \\
\partial_\mu\bar{X}^- X^-) - \tfrac{1}{2}gM[\bar{X}^+ X^+ H + \bar{X}^- X^- H + \tfrac{1}{c_w^2}\bar{X}^0 X^0 H] + \\
\tfrac{1-2c_w^2}{2c_w}igM[\bar{X}^+ X^0\phi^+ - \bar{X}^- X^0\phi^-] + \tfrac{1}{2c_w}igM[\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-] + \\
igMs_w[\bar{X}^0 X^-\phi^+ - \bar{X}^0 X^+\phi^-] + \tfrac{1}{2}igM[\bar{X}^+ X^+\phi^0 - \bar{X}^- X^-\phi^0].
\end{gathered}
$$







Physics as a **Sandbox**

Learning to Pivot with Adversarial Networks

Gilles Louppe
New York University
g.louppe@nyu.edu

Michael Kagan
SLAC National Accelerator Laboratory
makagan@slac.stanford.edu

Kyle Cranmer
New York University
kyle.cranmer@nyu.edu

We know many of the **dependencies** in our data and how our experiments/pre-processing **shape the data** → evaluate **de-biasing methods**

Energy flow polynomials:
A complete linear basis for jet substructure

Patrick T. Komiske, Eric M. Metodiev, Jesse Thaler
Center for Theoretical Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
E-mail: pkomiske@mit.edu, metodiev@mit.edu, jthaler@mit.edu

We know some patterns a model should learn and can build **interpretable bases** for some problems → contribute to **mechanistic interpretability**

ATLAS flavour-tagging algorithms for the LHC Run 2 pp collision dataset

The ATLAS Collaboration

We know the **phase space** of our data and **axes** along which it varies → can study **generalizability** of models

Constraint-based Graph Network Simulator

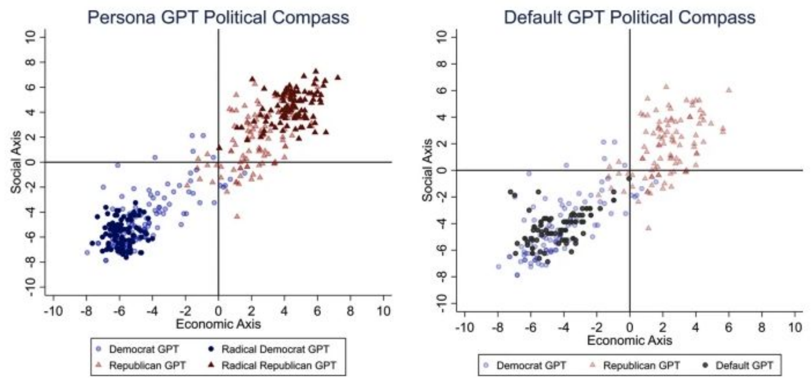
Yulia Rubanova [*1] Alvaro Sanchez-Gonzalez [*1] Tobias Pfaff [1] Peter Battaglia [1]

We can **compare model learned knowledge** to **true generating functions** → evaluate **robustness of new architectures**

Experimental **Design**

- A paper found that RLHF results in ChatGPT having a strong liberal/Democratic bias
- Prompt ChatGPT to respond to political statements while impersonating people from a side of the political spectrum and compare to neutral responses
- Collect answers to the same question 100 times to reduce variability

Fig. 2

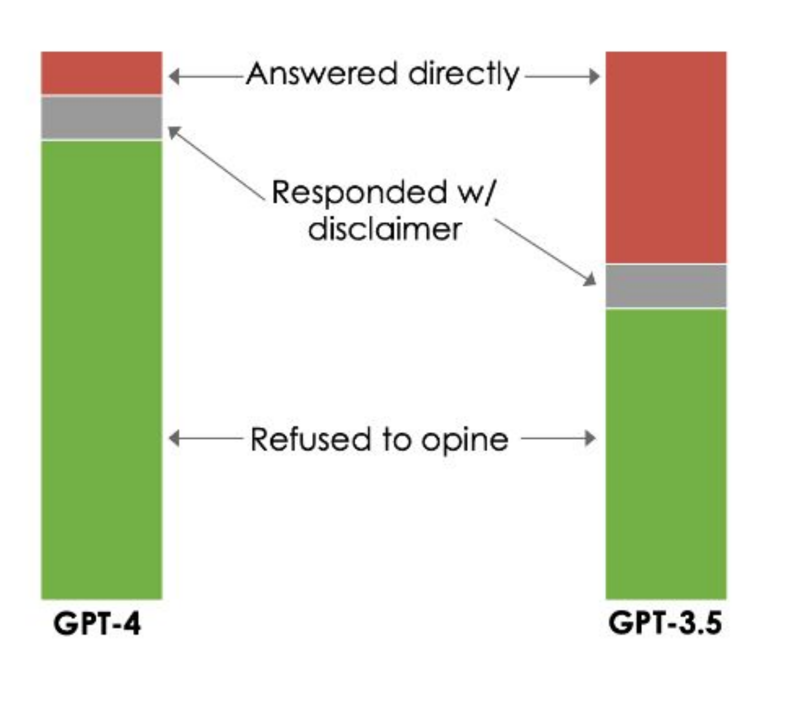


Political Compass quadrant—Average and Radical ChatGPT Impersonations (left) and Default and Average ChatGPT Impersonations (right). *Notes*: Political Compass quadrant classifications of the 100 sets of answers of each impersonation. The vertical axis is the social dimension: more negative values mean more libertarian views, whereas more positive values mean more authoritarian views. On the horizontal axis is the economic dimension: more negative values represent more extreme left views, and more positive values represent more extreme right views

More human than human: measuring ChatGPT political bias: Motoki et al

Experimental **Design**

- The paper had some scientific flaws
- Questions were asked as multiple choice + with prompting to try to force the model to opine (no construct validity)
- Generated politically neutral questions with ChatGPT and asked the model how a democrat or republican would answer
- Results depend on question ordering, and asking all questions in the same session



Does ChatGPT have a liberal bias?: Narayanan and Kapoor

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

Research: Risks

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 **limited fundamental understanding of something we have built**

Equivariance Is Not All You Need: Characterizing the Utility of Equivariant Graph Neural Networks for Particle Physics Tasks

Savannah Thais[1] Daniel Murnane[2]

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.

1. Introduction and Background

Over the past several years, Machine Learning (ML) has been established as a core component of many types of physics research (Carleo et al., 2019; Tanaka et al., 2021; Erdmann et al., 2021). Because physics is governed by

(Reiser et al., 2022). Equivariant GNNs combine several different types of inductive biases. As explained below, GNNs are permutation equivariant by construction and the graph itself (a combination of nodes and connective edges) incorporates an explicit relational or structural inductive bias into the data representation. Equivariant GNNs add an additional symmetry-based inductive bias by requiring that the function learned by the GNN is equivariant under transformations of some specified symmetry group.

While there are many types of GNNs, we will briefly describe message passing GNNs specifically (Gilmer et al., 2017), as they are the kind used in the example experiments discussed later in this paper. Basic message passing GNNs update the representations of graph nodes by exchanging information between neighboring nodes. In each message passing iteration, nodes aggregate information from their neighbors by applying a learnable function to the features h_j of neighboring nodes x_j (possibly as well as the central node x_i and any features of the connecting edges $e_{i,j}$); this transformed neighborhood information is aggregated by a permutation equivariant function to form the 'message', which is then combined with the central node's current features to produce an updated representation. This process is described mathematically as

$$h_i^{l+1} = \psi(h_i^l, \square_{j \in N(i)} m_{ij}) \qquad (1)$$

Hegemonic **Research**

Certain research approaches dominate publishing venues

- Generally focused on improving performance on benchmark data sets

- Often involves developing new, larger models. Exploiting large data and compute regime

We may neglect other promising avenues of research and the value of null results

Exploring the Whole Rashomon Set of Sparse Decision Trees

Rui Xin[1*] **Chudi Zhong**[1*] **Zhi Chen**[1*]

Takuya Takagi[2] **Margo Seltzer**[3] **Cynthia Rudin**[1]

[1] Duke University [2] Fujitsu Laboratories Ltd. [3] The University of British Columbia
{rui.xin926, chudi.zhong, zhi.chen1}@duke.edu
takagi.takuya@fujitsu.com, mseltzer@cs.ubc.ca, cynthia@cs.duke.edu

Abstract

In any given machine learning problem, there might be many models that explain the data almost equally well. However, most learning algorithms return only one of these models, leaving practitioners with no practical way to explore alternative models that might have desirable properties beyond what could be expressed by a loss function. The *Rashomon set* is the set of these all almost-optimal models. Rashomon sets can be large in size and complicated in structure, particularly for highly nonlinear function classes that allow complex interaction terms, such as decision trees. We provide the first technique for completely enumerating the Rashomon set for sparse decision trees; in fact, our work provides the first complete enumeration of any Rashomon set for a non-trivial problem with a highly nonlinear discrete function class. This allows the user an unprecedented level of control over model choice among all models that are approximately equally good. We represent the Rashomon set in a specialized data structure that supports efficient querying and sampling. We show three applications of the Rashomon set: 1) it can be used to study variable importance for the set of almost-optimal trees (as opposed to a single tree), 2) the Rashomon set for accuracy enables enumeration of the Rashomon sets for balanced accuracy and F1-score, and 3) the Rashomon set for a full dataset can be used to produce Rashomon sets constructed with only subsets of the data set. Thus, we are able to examine Rashomon sets across problems with a new lens, enabling users to choose models rather than be at the mercy of an algorithm that produces only a single model.

Stymied **Progression?**

False Belief

Misaligned research/publishing incentives and flawed scientific design may lead us to believe we have solved problems that we haven't. This risks subjecting real people to damaging or dangerous sytems

Ignoring Problems

Without tackling the challenging questions of model design and evaluation and increasing interdisciplinary collaborations, human-in-the-loop paradigms, and participatory design structures, we risk not making progress on the complicated questions that really matter to society.

Current Society

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?

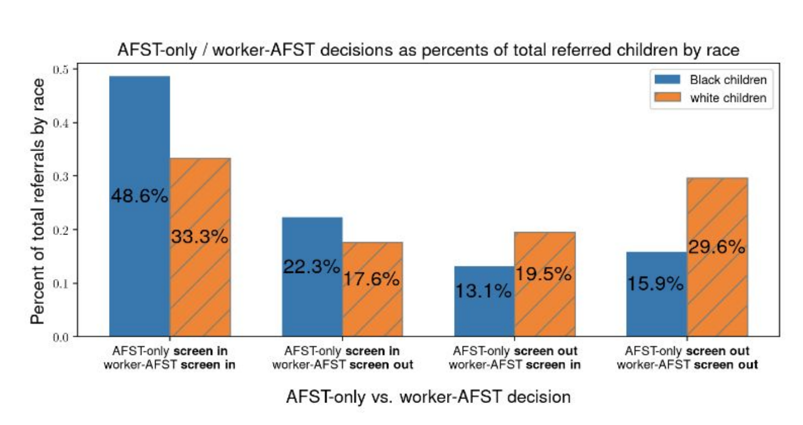
Deployment & Outcomes

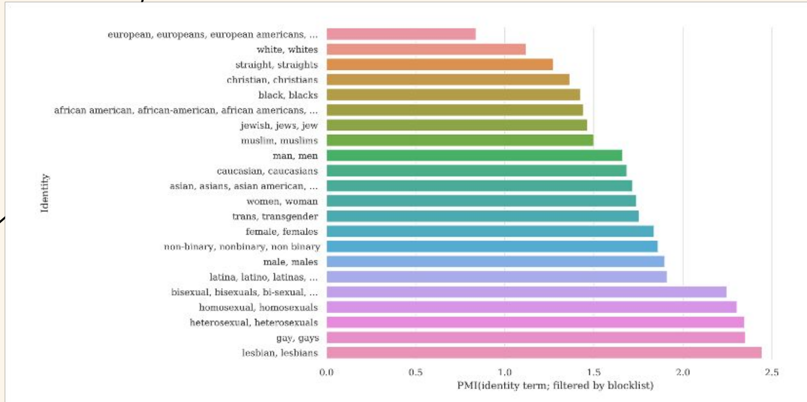
Who is subjected to what, how do we understand impact?

Downstream & Diffuse Impacts

What is changed or lost by what we build?

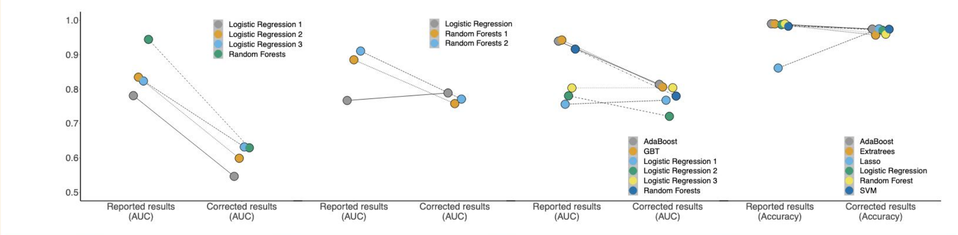
Bias + **Fairness**





- 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

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

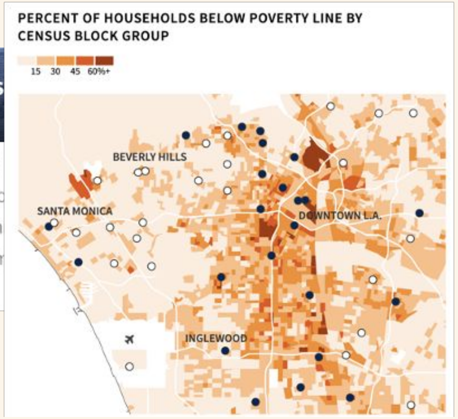
- Scientific mistakes in model construction, training, or evaluation yield unreliable or non-generalizable results
  - E.g. test set not drawn from distribution of interest, illegitimate features, data leakage, sampling bias
- Example: a sepsis prediction tool takes antibiotic use as an input feature, 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

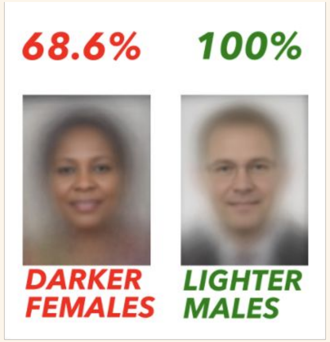
Rite Aid deployed facial recognition systems hundreds of U.S. stores

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

Future Society

The Consequences of **What We Build**

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

Shaping the **Future**

Power Concentration

Concentrating power in the hands of a few corporations with vast compute resources, widening wealth and opportunity inequality gap

Information Ecosystem

Ease of harmful or misleading content, training set contamination, acceleration of mis and disinformation

Climate

Impact of training and inference energy on climate, impact of resource mining for commute resources, relying on AI to solve climate change

Human Value

Devaluing of human elements: creativity, exploration, labor. TESCREAL philosophies.

What Can We Do?

Some **Ideas**

Interdisciplinary Spaces

Cultivate meaningful interdisciplinary spaces and collaborations where contributions are equitably valued

Technical Literacy

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

Scientific Approaches

Treat your model building and evaluation as a science. Draw on scientific methodology and principles

Advocacy

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

Self Interrogation

Consider your personal code of ethics and how it relates to your work and the broader scientific and AI ecosystem. Consider technology transfer

Policy

Share your scientific expertise with policy makers and champion meaningful regulations

Can We Automate Science?

SHOULD

~~Can~~ We Automate Science?

We get to decide what we want the future of technology to look like, and the role it plays in our lives and communities.

We must do so responsibly.

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
- AI Ethics Education for Scientists, Thais

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