AI Ethics for Science

+ *Science for AI Ethics*

Savannah Thais, Columbia University

Some Framing...

AI Has a **Hype Problem**

FORBES > INNOVATION

Will ChatGPT Solve All Our Problems?

Forbes Technology Council

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

IDEAS • TECHNOLOGY

Why Uncontrollable AI Looks More Likely Than Ever

Technology And Analytics

Using AI to Eliminate Bias from Hiring

by Frida Polli

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

'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

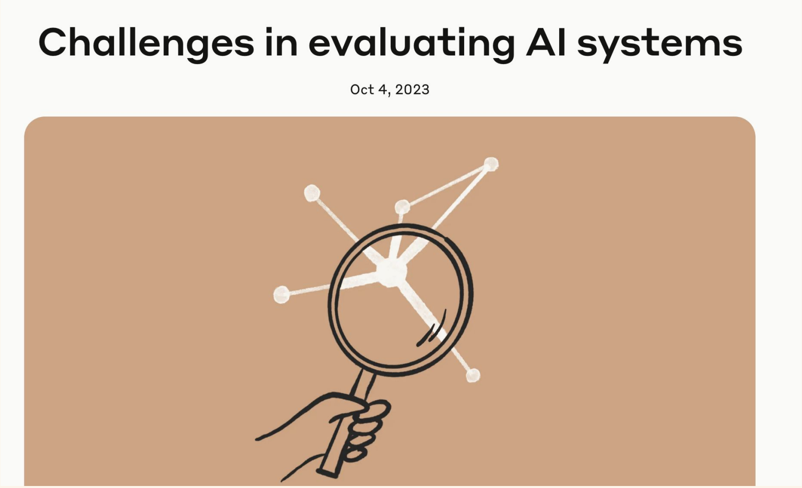
Enchanted Determinism:
Power without Responsibility in Artificial Intelligence

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**

AI Has a **Measurement Problem**



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

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

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
- Risk building **incorrect models** or not capitalizing on s**cientific opportunity**
- 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
- Risk building **incorrect models** or not capitalizing on s**cientific opportunity**
- 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

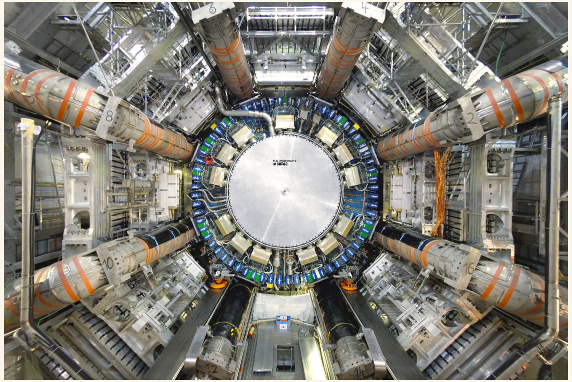
Research + Opportunities

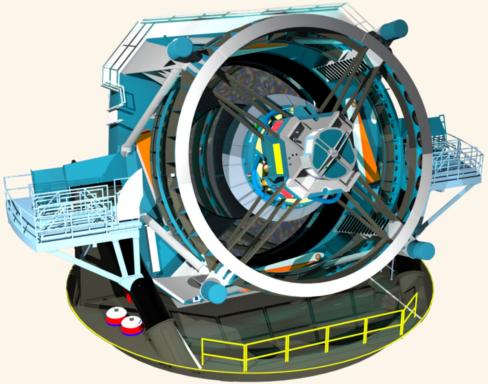
Physics

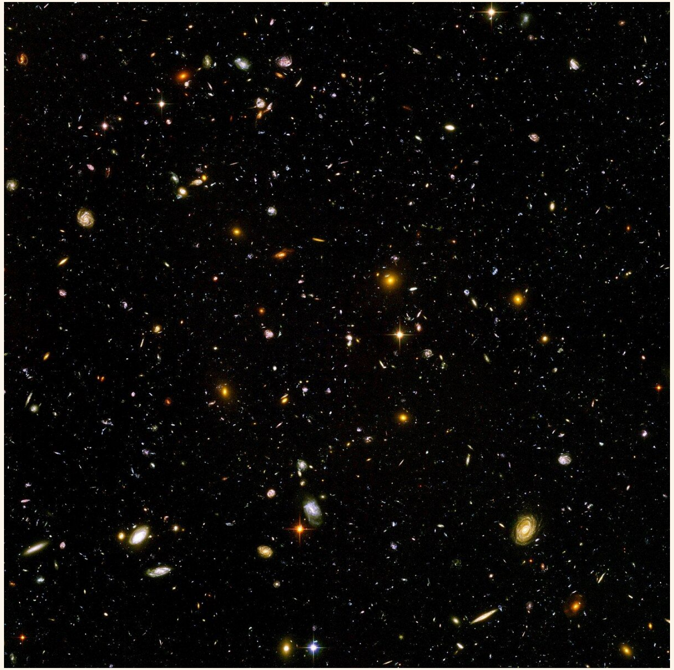
Trustworthy AI

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^2c_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)^2H^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}^0X^0H] + \\
\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}^0X^-\phi^+ - \bar{X}^0X^+\phi^-] + \\
igMs_w[\bar{X}^0X^-\phi^+ - \bar{X}^0X^+\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

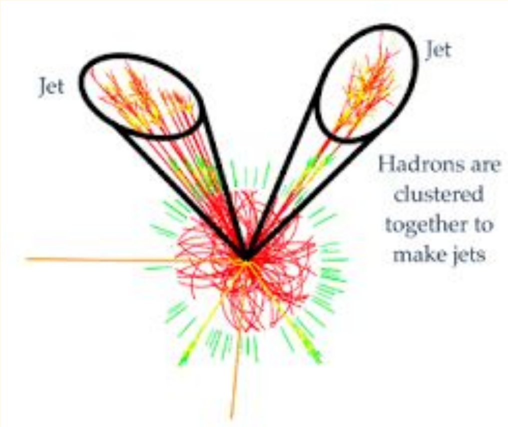
We know the **phase space** of our data and **axes** along which it varies → can study **generalizability** of models

Constraint-based Graph Network Simulator

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

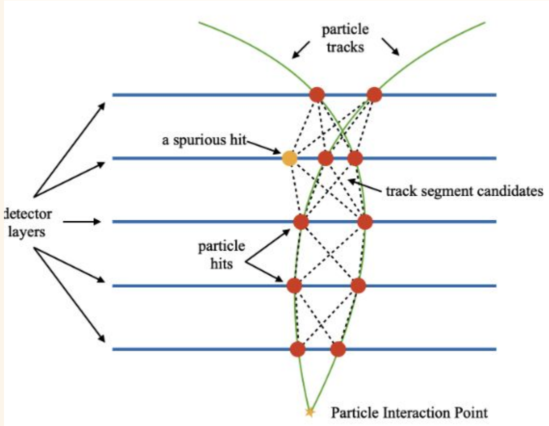
We can **compare model learned knowledge** to **true generating functions** → evaluate **robustness of new architectures**

Example: **Evaluating Equivariance**



Jet Tagging

- Using Top Quark Tagging Reference Dataset
- Build a Lorentz equivariant model (rotations and boosts in spacetime)



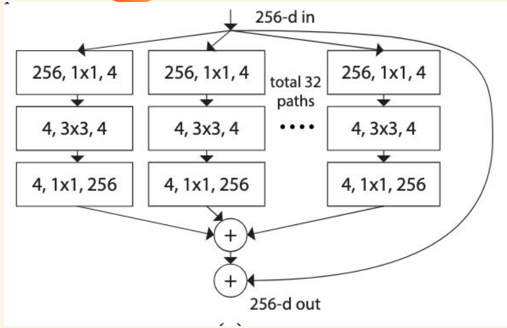
Particle Tracking

- Using TrackML Dataset
- Build an SO(2) rotation equivariant model (in x-y plane)

Image from Nelson

Image from arxiv:2109.12636

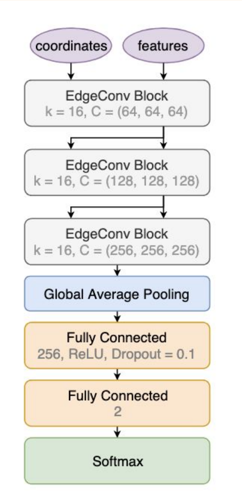
Baseline **Tagging Models**



ResNexT

Deep 2D CNN on jet images

arXiv:1611.05431



ParticleNet

Message passing dynamic graph GNN on particle graph

arXiv:1902.08570

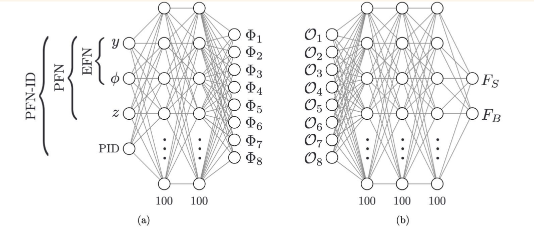
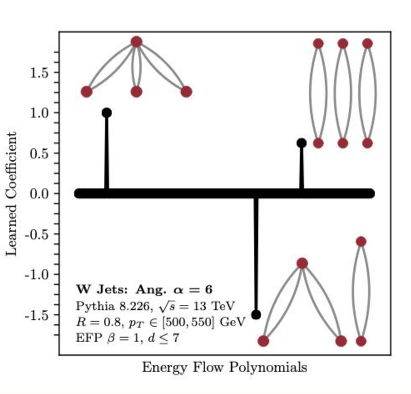


Figure 4: The particular dense networks used here to parametrize (a) the per-particle mapping Φ and (b) the function F, shown for the case of a latent space of dimension $\ell = 8$. For the EFN, the latent observable is $\mathcal{O}_a = \sum_i z_i \Phi_a(y_i, \phi_i)$. For the PFN family, the latent observable is $\mathcal{O}_a = \sum_i \Phi_a(y_i, \phi_i, z_i, \mathrm{PID}_i)$, with different levels of particle-ID (PID) information. The output of F is a softmaxed signal (S) versus background (B) discriminant.

Particle Flow

Deep set network on particle features

arXiv:1810.05165



Energy Flow Polynomials

Linear discriminant on EFP complete linear basis

arXiv:1712.07124

Equivariant **Tagging Models**

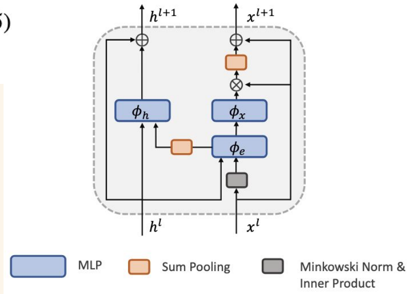
$$\mathcal{F}_i^{(p+1)} = \mathcal{L}_{\mathrm{CG}}\left(\mathcal{F}^{(p)}\right)_i = W \cdot \left(\mathcal{F}_i^{(p)} \oplus \mathrm{CG}\left[\mathcal{F}_i^{(p)}\right]^{\otimes 2} \oplus \oplus\mathrm{CG}\left[\sum_j f(p_{ij}^2) p_{ij} \otimes \mathcal{F}_j^{(p)}\right]\right). \quad (25)$$

Lorentz Group Network

NN with CG-layers that take tensor products and decompose into irreps using Clebsch-Gordan map, on particle features

arXiv:2006.04780

$$m_{ij}^l = \phi_e\left(h_i^l, h_j^l, \psi(\|x_i^l - x_j^l\|^2), \psi(\langle x_i^l, x_j^l \rangle)\right),$$



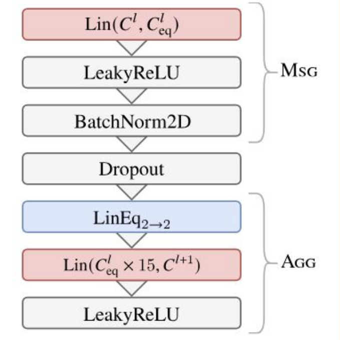
Lorentz Group Equivariant Block (LGEB)

LorentzNet

Message passing GNN with Lorentz equivariant message, on particle graph

arXiv:2201.08187

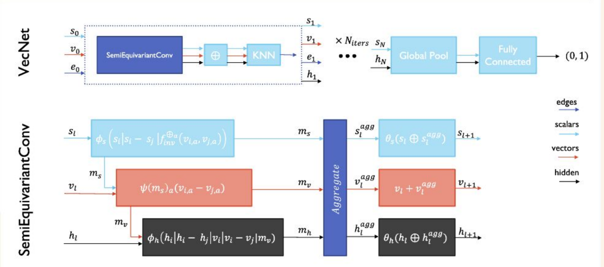
$$I(p_1, \ldots, p_N) = I\left(\{p_i \cdot p_j\}_{i,j}\right).$$



PELICAN

Deep set-esq network using all totally symmetric Lorentz invariants and full set of 15 rank 2 to rank 2 maps as aggregators

arXiv:2211.00454

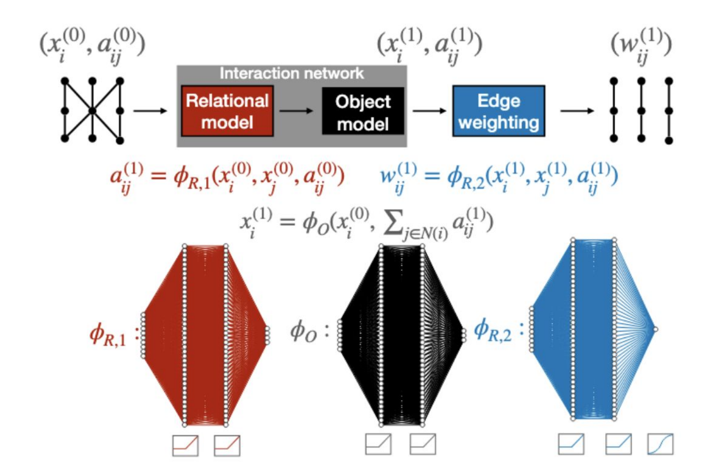


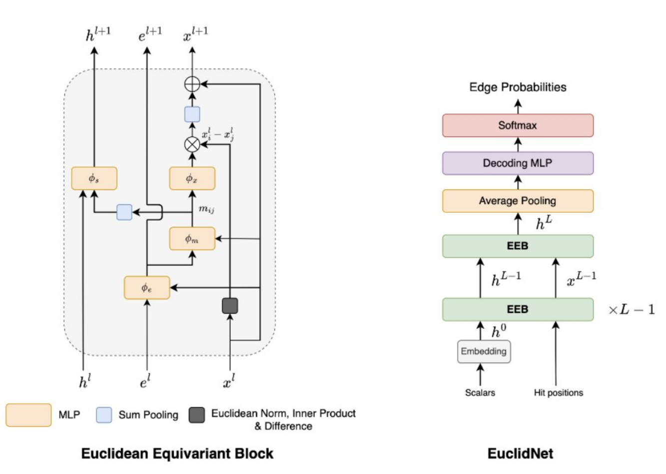
VecNet

Message passing GNN with Lorentz equivariant message and (optionally) unconstrained message, on particle graph

arXiv:2202.06941

Tracking **Models**





Interaction Network

Message passing GNN with node and edge updates, on hit graph (with physics-based edge construction)

arXiv:2103.16701

EuclidNet

Message passing GNN with SO(2)-equivariant message construction, on hit graph (with physics-based edge construction)

arXiv:2304.05293

Evaluating **Equivariance**

Tagging	Accuracy	AUC	Parameters	Ant Factor
ResNeXt	0.936	0.984	1.46M	4.28
ParticleNet	0.938	0.985	498k	13.4
PFN	0.932	0.982	82k	67.8
EFP	0.932	0.980	**1k**	**5000**
LGN	0.929	0.964	4.5k	617
VecNet.1	0.935	0.984	633k	9.87
VecNet.2	0.931	0.981	15k	350
PELICAN	**0.943**	**0.987**	45k	171
LorentzNet	0.942	0.9868	220k	35

Tracking	N Hidden	AUC	Parameters	Ant Factor
EuclidNet	8	**0.9913**	**967**	**11887**
InteractionNet	8	0.9849	1432	4625
EuclidNet	16	0.9932	2580	**5700**
InteractionNet	16	0.9932	4392	3348
EuclidNet	32	0.9941	4448	3811
InteractionNet	32	**0.9978**	**6448**	**7049**

Accuracy

- Jet tagging: highest accuracy model is equivariant, but not all equivariant models perform well
- Tracking: for small models equivariant models have highest accuracy, but performance plateaus as models grow
- Overall, relationship between equivariance and accuracy is unclear (confounding factors remain)

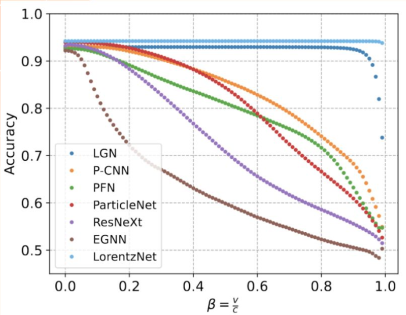
Model Efficiency

- Jet tagging: regression model with physics inputs is most efficient. Semi-equivariant model is also efficient.
- Tracking: relationship changes with model size
- Overall, equivariance does not seem to contribute directly to model efficiency

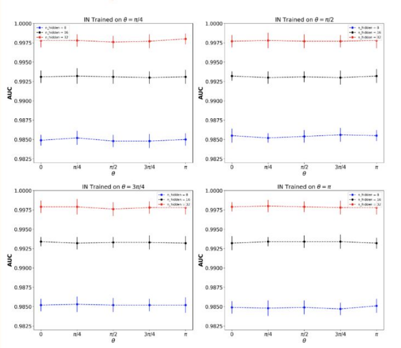
Ant factor = 10^5/[(1-AUC)*N_p]

Evaluating **Equivariance**

Tagging



Tracking



Tagging	Training %	Accuracy	AUC
LorentzNet	0.5%	**0.932**	**0.9793**
ParticleNet	0.5%	0.913	0.9687
LorentzNet	1%	**0.932**	**0.9812**
ParticleNet	1%	0.919	0.9734
LorentzNet	5%	**0.937**	**0.9839**
ParticleNet	5%	0.931	0.9839

Generalizability

- Jet tagging: equivariant models generalize, but not all to the same extent
- Tracking: both equivariant and sufficiently large non-equivariant models generalize
- Overall, equivariance provides a good amount of generalization, but other models can too (tradeoffs)

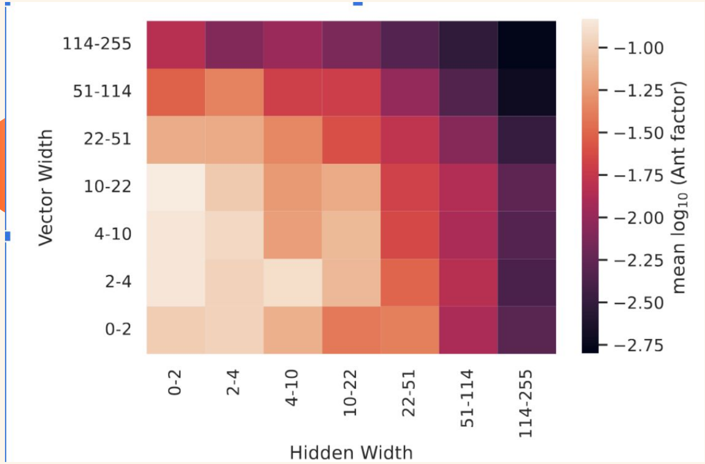
Data Efficiency

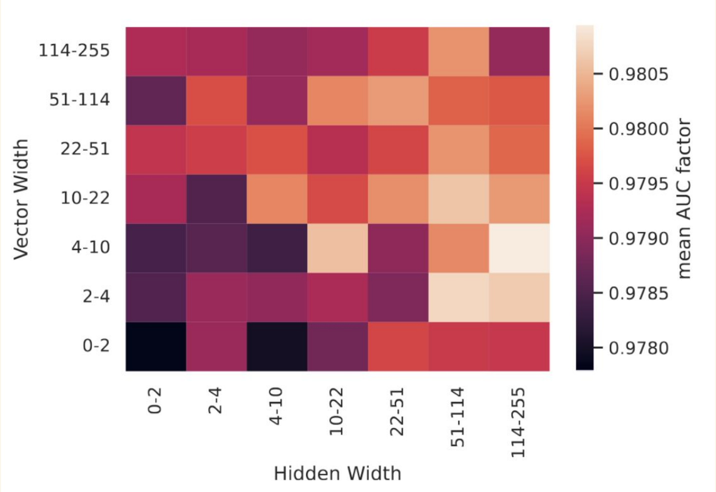
- Jet tagging: clear benefit from equivariance in very small data regimes: achieves 99% of full accuracy with just 0.5% of training dataset
    - Compared to 97% for non-equivariant model
- Overall, seems to be the most replicable benefit of equivariance. This is demonstrated in other papers, such as NequIP

Over-**constraint?**

Is full equivariance the right approach for HEP tasks?

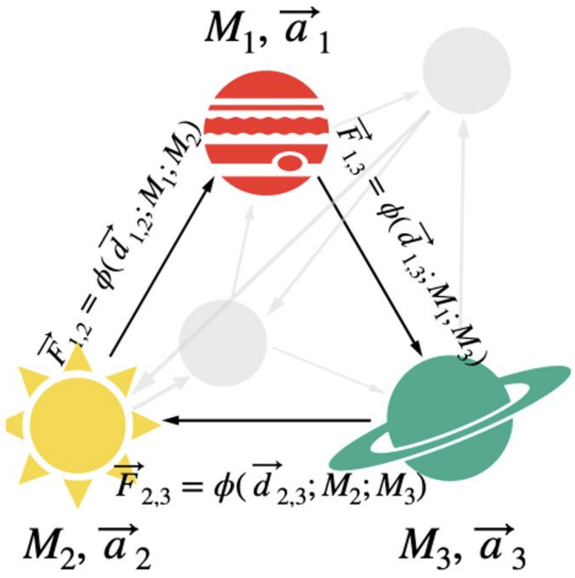
- Unconstrained models can learn to generalize under symmetry transformations
- VecNet studies show optimal accuracy and model efficiency are achieved with mixed equivariant and non-equivariant information
- While the underlying physics is obeys symmetries, observed data is likely NOT fully symmetric





Example: **Testing Explainability**

1. Our inputs are the positions of the bodies
2. They are converted into pairwise distances
3. **Our model tries to guess a mass for each body**
4. **It then also guesses a force, that is a function of distance and masses**
5. **Using Newton's laws of motion ($\sum \vec{F} = M\vec{a}$) it converts the forces into accelerations**
6. **Finally, it compares this predicted acceleration, with the true acceleration from the data**



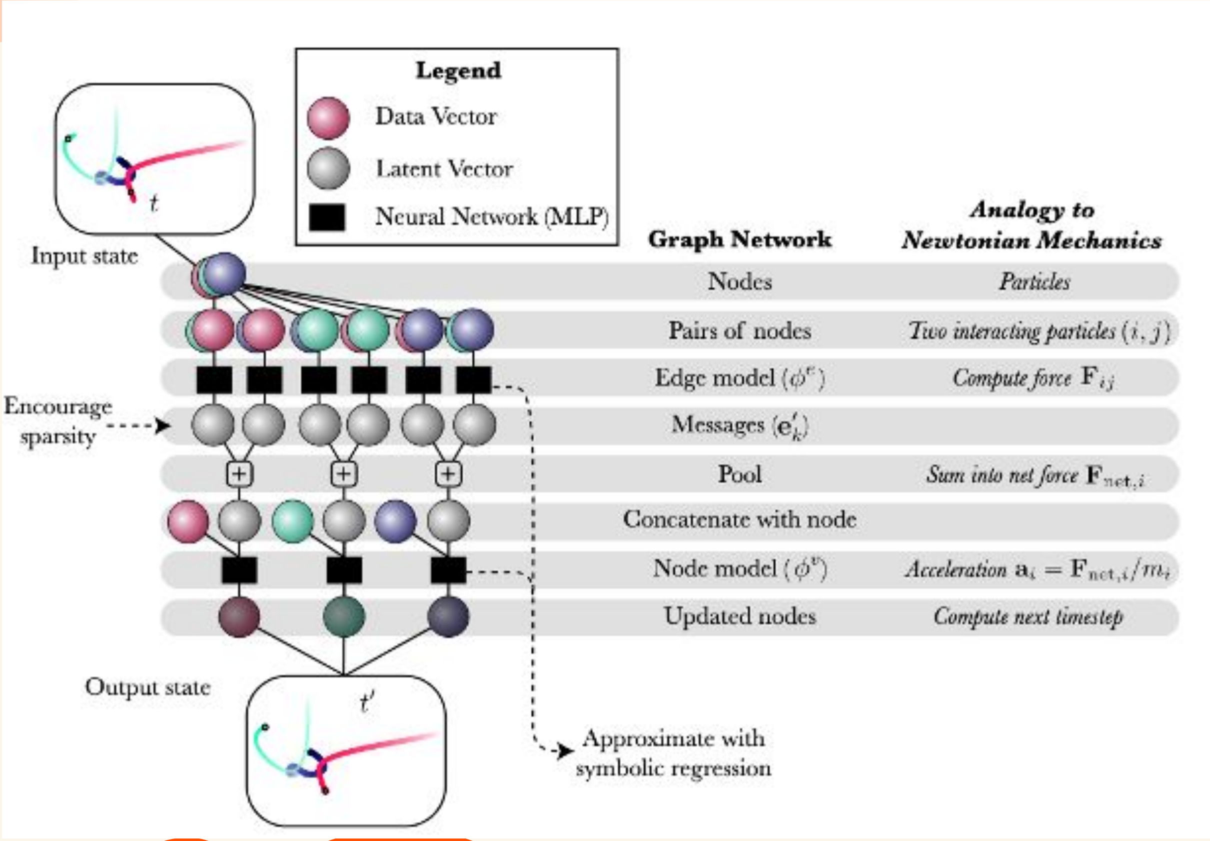
Minimize

$$\left| \vec{a}(\text{pred}) - \vec{a}(\text{true}) \right|^2$$

Slides from Pablo Lemos

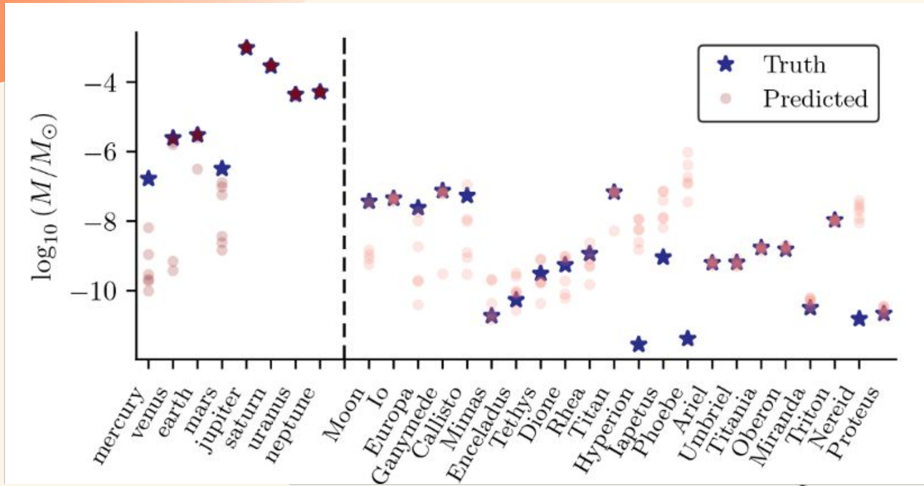
Discovering Symbolic Models from Deep Learning with Inductive Biases, Cranmer et al

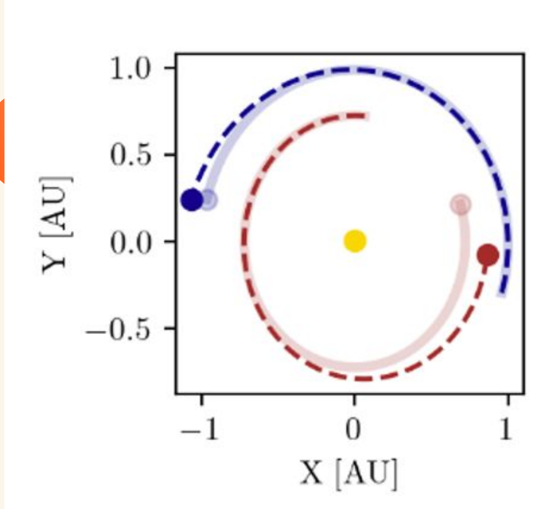
Inductive Bias **Network**

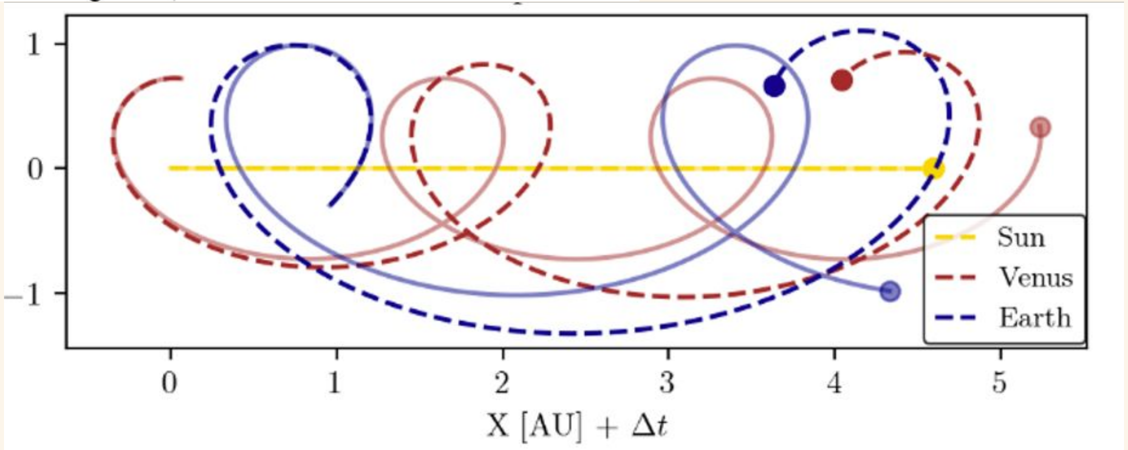


[Discovering Symbolic Models from Deep Learning with Inductive Biases](), Cranmer et al

Model **Performance**





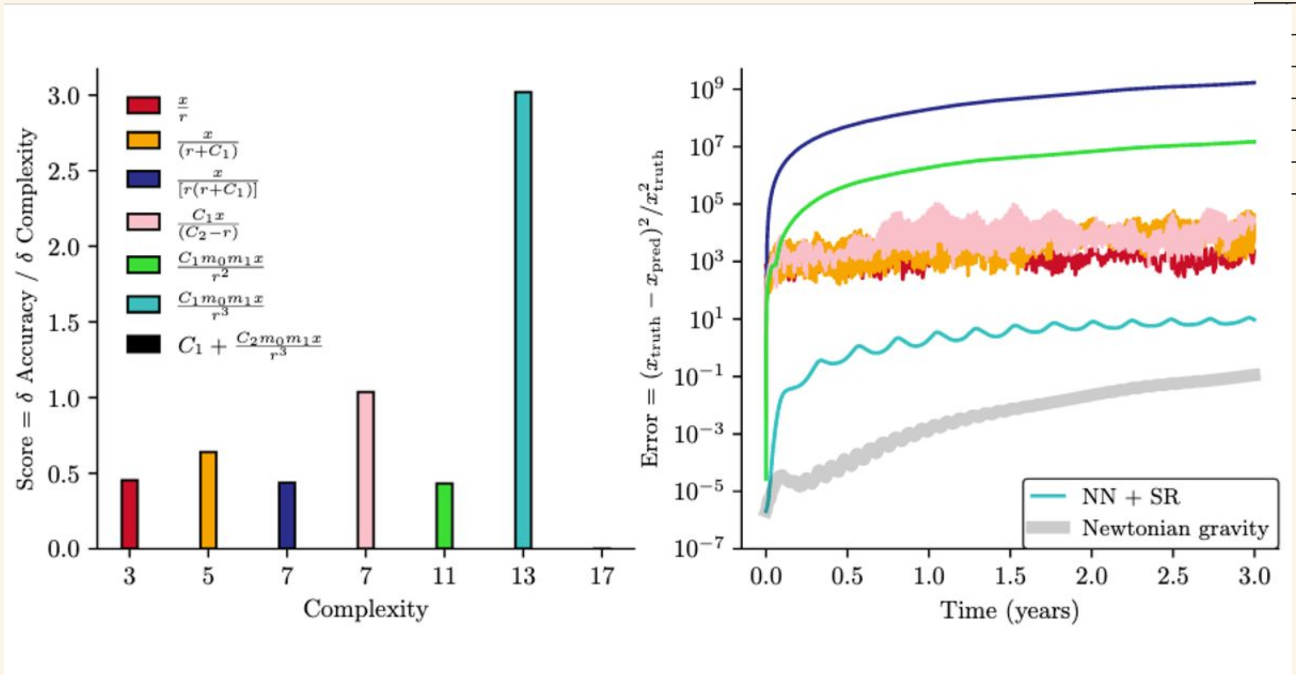


Slides from [Pablo Lemos]

[Discovering Symbolic Models from Deep Learning with Inductive Biases], Cranmer et al

Extracting **Physics**

- Use symbolic regression package eureqa to fit analytic expressions to the subnetworks
- Use constraint to balance accuracy and equation complexity
- Substituting learned equation for force network improves model accuracy



- Several limitations and opportunities for further study:
    - Models don't always converge, picking the right analytic equation is difficult, space of good models
- By studying explainability methods in known systems, we can characterize their robustness

Slides from Pablo Lemos

Example: **Impact of Transparency**

- Recent paper explores the relationship between AI Ethics principles and Climate Science

- In particular, we highlight that transparency and documentation is key to **accurate science, trust building, and equity**

PLOS CLIMATE

OPINION

Ethics in climate AI: From theory to practice

Viviana Acquaviva[1,2]*, Elizabeth A. Barnes[3], David John Gagne, II[4], Galen A. McKinley[2], Savannah Thais[5]

1 Physics Department, CUNY NYC College of Technology, Brooklyn, New York, United States of America, 2 Department of Earth and Environmental Sciences, Columbia University and Lamont-Doherty Earth Observatory, Palisades, New York, United States of America, 3 Department of Atmospheric Science, Colorado State University, Fort Collins, Colorado, United States of America, 4 NSF National Center for Atmospheric Research, Boulder, Colorado, United States of America, 5 Columbia University Data Science Institute, New York, New York, United States of America

* vacquaviva@citytech.cuny.edu

Check for updates

OPEN ACCESS

Citation: Acquaviva V, Barnes EA, Gagne DJ, II, McKinley GA, Thais S (2024) Ethics in climate AI: From theory to practice. PLOS Clim 3(8): e0000465. https://doi.org/10.1371/journal.pclm.0000465

Editor: Jamie Males, PLOS Climate, UNITED KINGDOM OF GREAT BRITAIN AND NORTHERN IRELAND

Published: August 2, 2024

Copyright: © 2024 Acquaviva et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original

Climate science, and climate artificial intelligence (AI) in particular, cannot be disconnected from ethical societal issues, such as resource access, conservation, and public health. An apparently apolitical choice—for example, treating all data points used to train an AI model equally —can result in models that are more accurate in regions where the density and quality of data is higher; these often coincide with the northern and western areas of the world (e.g., [1, 2]).

Inequity in the access to data and computational resources exacerbates gaps between communities in understanding climate change impacts and acting towards mitigation and adaptation, often in ways that are detrimental to those who are most affected (e.g., [3, 4]). While these issues are not exclusive to AI, widespread opacity in the development and functioning of AI models, presentation of AI model outcomes, and the rapid evolution of the AI field further increase the inequality in power and agency among differently resourced parties.

This creates an opportunity for climate scientists to rethink the role of ethics in their approach to research. There are many ways in which climate scientists can interact with society. Here we focus on the process of scientific research, identifying some good practices for building trustworthy and responsible models and then providing some resources.

In creating and training models, we encourage researchers to recognize that science cannot claim to be purely "objective", and that the choice of priors, data, and metrics all carry biases (e.g., [5]). Resolving or eliminating them is not realistic, as the interpretation of a "better" model or result is highly dependent on the user's specific goal. Hence, it is crucial to be open and specific about the assumptions made, the algorithms and hyperparameters used, and the evaluation metrics and processes, and ideally to also make data and code available, following

[Paper](#)

Dataset **Documentation**

Datasheets for Datasets

TIMNIT GEBRU, Black in AI
JAMIE MORGENSTERN, University of Washington
BRIANA VECCHIONE, Cornell University
JENNIFER WORTMAN VAUGHAN, Microsoft Research
HANNA WALLACH, Microsoft Research
HAL DAUMÉ III, Microsoft Research; University of Maryland
KATE CRAWFORD, Microsoft Research

1 Introduction

Data plays a critical role in machine learning. Every machine learning model is trained and evaluated using data, quite often in the form of static datasets. The characteristics of these datasets fundamentally influence a model's behavior: a model is unlikely to perform well in the wild if its deployment context does not match its training or evaluation datasets, or if these datasets reflect unwanted societal biases. Mismatches like this can have especially severe consequences when machine learning models are used in high-stakes domains, such as criminal justice [1, 13, 24], hiring [19], critical infrastructure [11, 21], and finance [18]. Even in other domains, mismatches may lead to loss of revenue or public relations setbacks. Of particular concern are recent examples showing that machine learning models can reproduce or amplify unwanted societal biases reflected in training datasets [4, 5, 12]. For these and other reasons, the World Economic Forum suggests that all entities should document the provenance, creation, and use of machine learning datasets in order to avoid discriminatory outcomes [25].

Datasheets-for-Earth-Science-Datasets

Welcome! This repository contains the "beta version" of Datasheets for Earth Science Datasets released for feedback, comments, and questions from the broader Earth science community. Please check out the **InstructionalGuide** to learn more!

Datasheet for Veloso-Aguila et al. "Tornadoes in Southeast South America: Mesoscale to Planetary-scale Environments"

Released: January 09, 2024
Last updated: January 09, 2024

Daniel Veloso-Aguila
Department of Atmospheric Science
Colorado State University
Fort Collins, Colorado, USA
daniel.veloso.a@gmail.com

1. PURPOSE

A. For what purpose was the dataset created?

This dataset contains a compendium of tornado events reported in Southeast South America between 1991 and 2020. It was built to conduct a study of the environments that support tornadic storms in this region

B. Who created the dataset (e.g., which individual or research group), on behalf of which entity (e.g., institution or company), and under what funding (e.g., grantor[s] and grant number[s])?

This dataset was built by Daniel Veloso-Aguila under the advice of Dr. Kristen Rasmussen and Dr. Eric Maloney at Colorado State University. This PhD research is funded by the Equal Opportunity Fulbright-ANID Scholarship (Chile). This research is also sponsored by National Science Foundation grants AGS-1661657, AGS-1841754, AGS-2146709, and Department of Energy grant DE-SC0022056 (United States).

southern Brazil) between 1991 and 2020, including information about location, date and time of occurrence (in both local time and UTC), intensity (if reported), description of the impacts associated with the event (mostly in Spanish), and external links to supporting evidence.

B. What is the data? (e.g., file format, dimensionality, variables and metadata, spatiotemporal coverage)

This dataset is stored in a .xls spreadsheet file. Every tornado report is organized in rows, while all the details about the events are organized in columns (e.g., location, date and time, damage reports, etc.)

C. What processing has been applied to this data?

There is no processing of this data, as it is just a collection of information from multiple sources.

D. Is the unprocessed data available in addition to the processed data? If so, please provide a stable link to the unprocessed data.

Github

Impact on **Analyses**

- Physical simulations and observations are used in downstream climate and econometric analyses
- However, there are many scientific pitfalls if limitations of data are not properly documented and accounted for
    - Correlations of variables, underlying causal mechanisms, gridding of simulators, geographic bias, etc

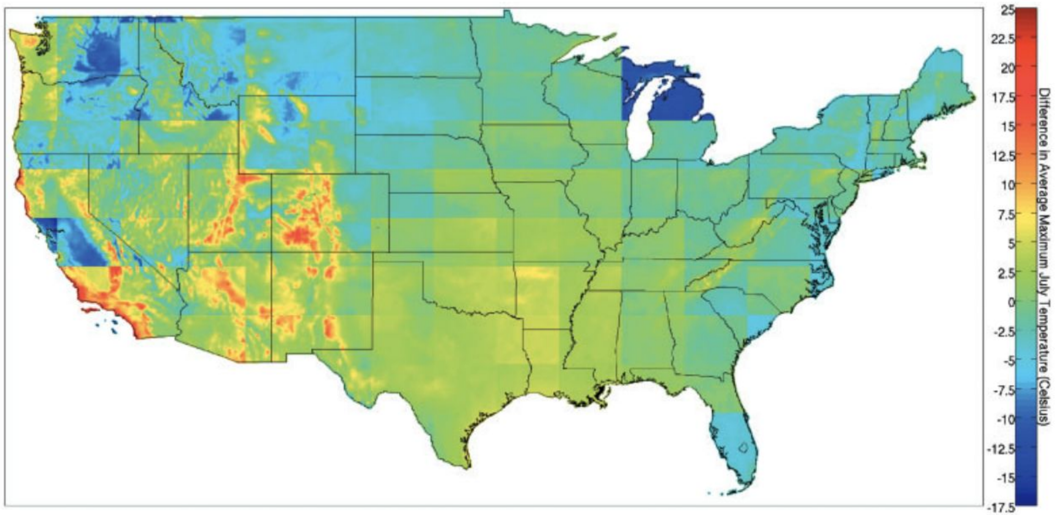
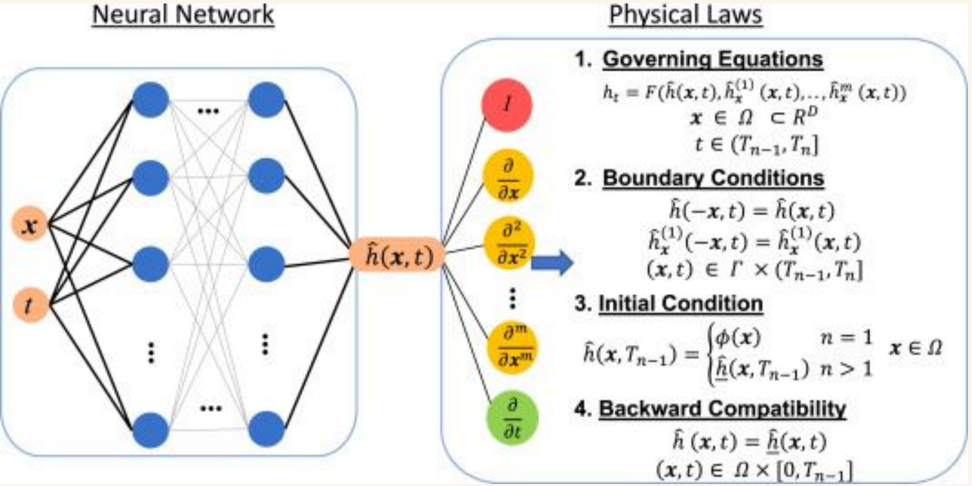


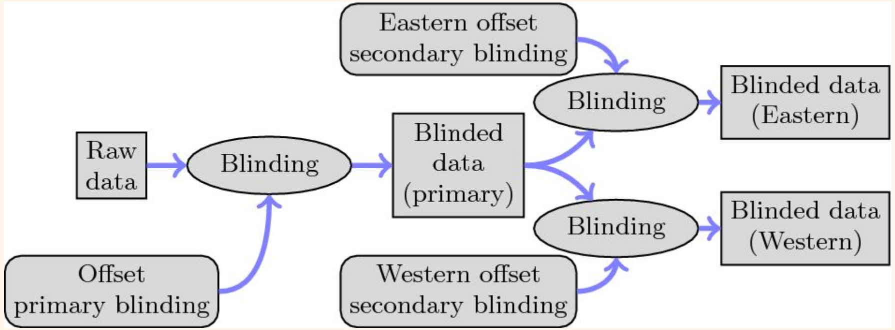
Figure 3 Aggregation bias: Hadley grid averages versus PRISM grid averages in each PRISM grid (1961–1999)

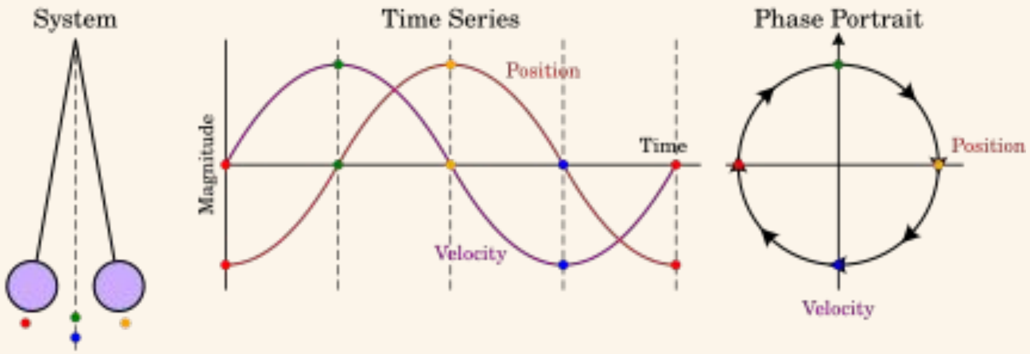
Notes: The figure plots the difference in the average daily maximum temperature in the month of July in the years 1960–1999 between the GCM (Hadley III), which has the coarser resolution, and the fine-scale weather grid (PRISM 2009). A positive number indicates that the GCM grid average exceeds the PRISM average, which is based on interpolated station data.

Using Weather Data and Climate Model Output in Economic Analyses of Climate Change: Auffhammer et al

Physics as **Inspiration**



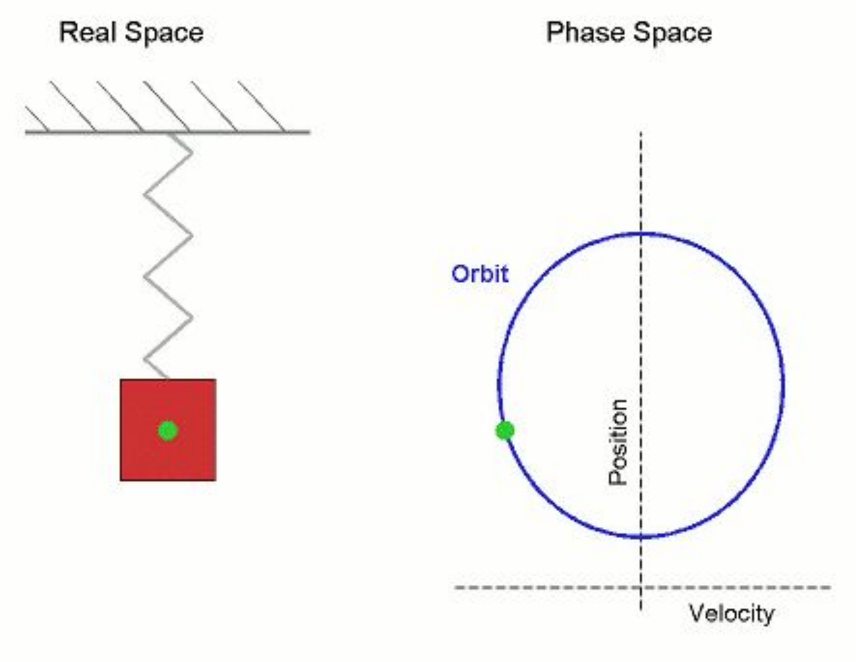




Example: **Phase Space**

- Concept of a space where where all possible ‘states’ of a dynamic system are represented as unique points

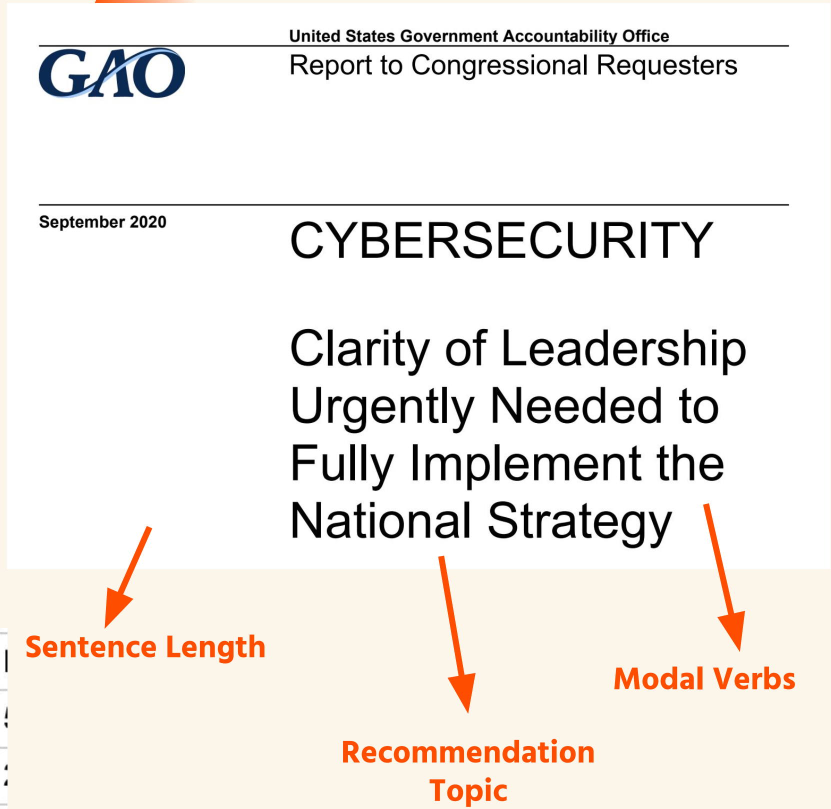
- We can extend this concept to characterize the space in which we expect a model to perform
  - Construct axes that fully (or as fully as possible) describe the different distributions of the performance space



Phase Space of **Policy Research**

- Developing a model to extract and group actionable policy recommendations from large corpus of documents and legal writings
- By characterizing the phase space we can evaluate the robustness of our models
  - Combine statistical analysis and domain expertise to construct phase space
  - In non-physics problems, may not be able to fully characterize the space

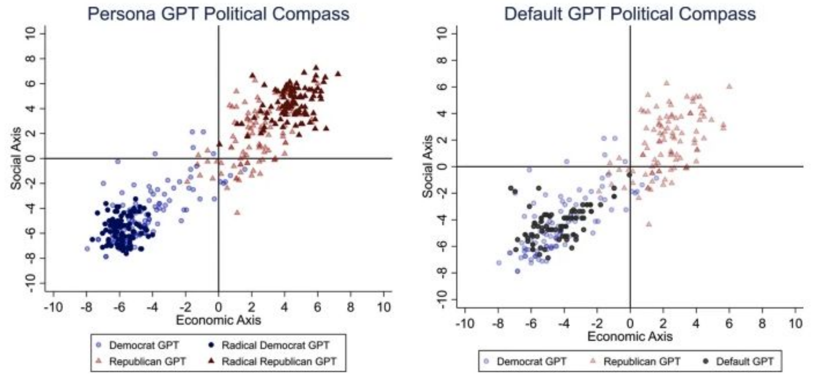
	Precision	Recall	Accuracy
Sentences contain Modal Verb	1	1	1
Sentences don’t contain Modal Verb	0.7273	0.7619	0.8764
all	0.9211	0.9333	0.9267


United States Government Accountability Office
GAO
Report to Congressional Requesters
September 2020
CYBERSECURITY
Clarity of Leadership Urgently Needed to Fully Implement the National Strategy

Example: **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

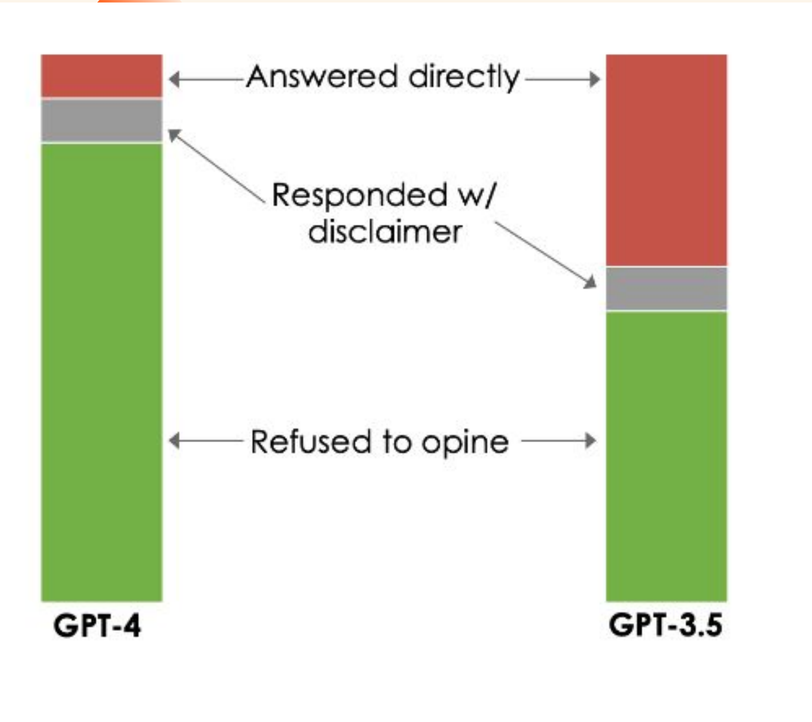


Political Compass quadrant—Average and Radical ChatGPT Impersonations (left) and Default and Average ChatGPT Impersonations (right). *Notes*: Political Compass quadrant classifications of the 100 sets of answers of each impersonation. The vertical axis is the social dimension: more negative values mean 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

Scientific **Failure**

- 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

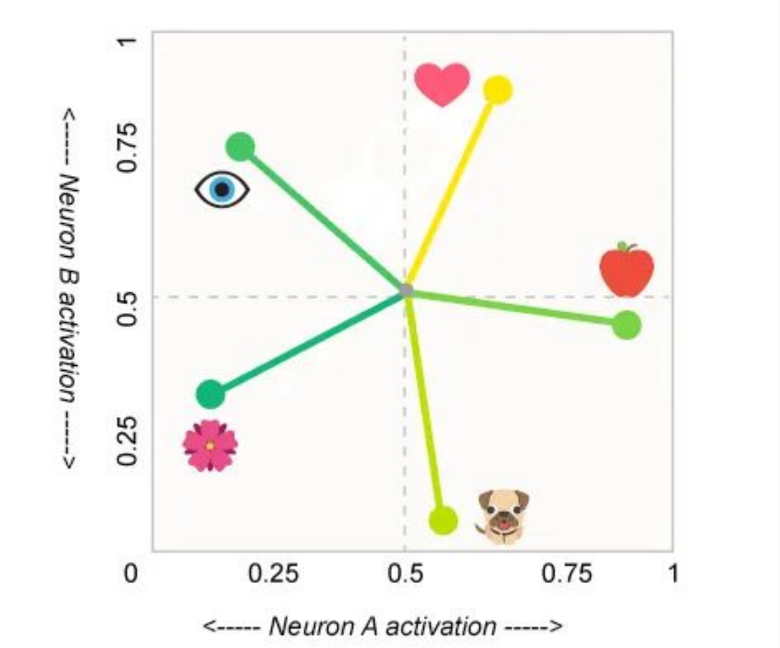
Reach a Conclusion
I should or should not use this model because of X, Y, and Z

07

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

Example: **Physics Concepts**

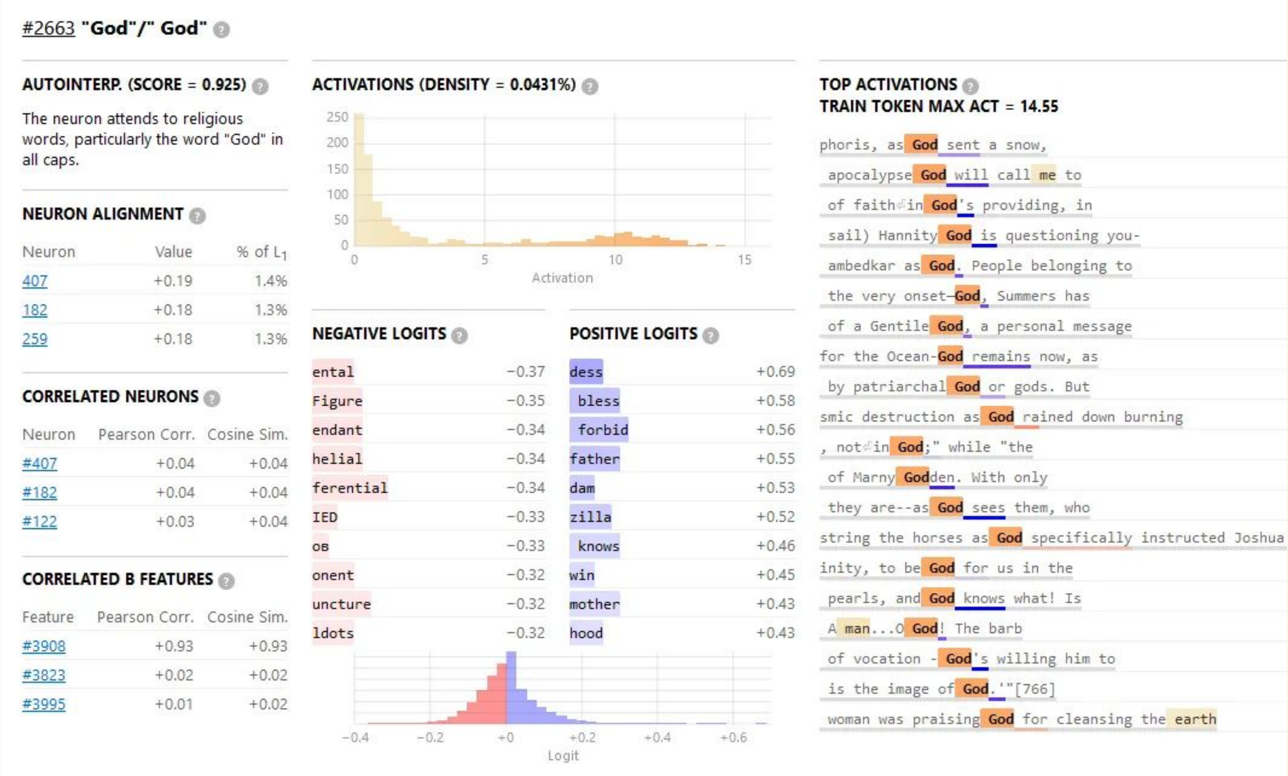
- Some of Anthropic's interpretability research is inspired by the concept of superposition in quantum mechanics
- Combinations of neurons in a network are akin to spins of particles
    - Thus, two neurons are able to represent more than two points in phase space
- Allows a small NN to represent a higher dimensional space



[Toy Models of Superposition]

Example: **Physics Concepts**

Research demonstrates that these sub phase spaces may map to human interpretable concepts



[God Help Us, Let's Try to Understand AI Monosemanticity]

Risks If We Don’t

To Our Research....

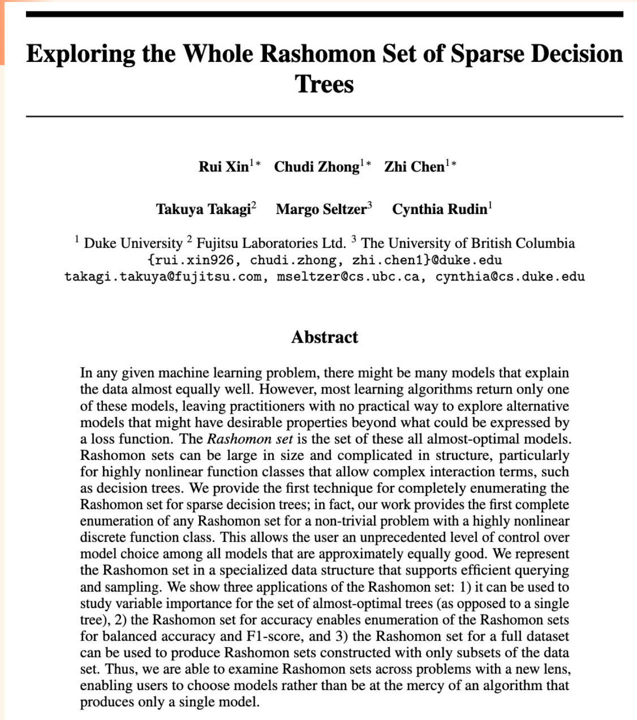
Hegemonic **Research**

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.

Harms to **Science**

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

Savannah Thais
Columbia University Data Science Institute
New York, New York 11221 USA
st3565@columbia.edu

Abstract

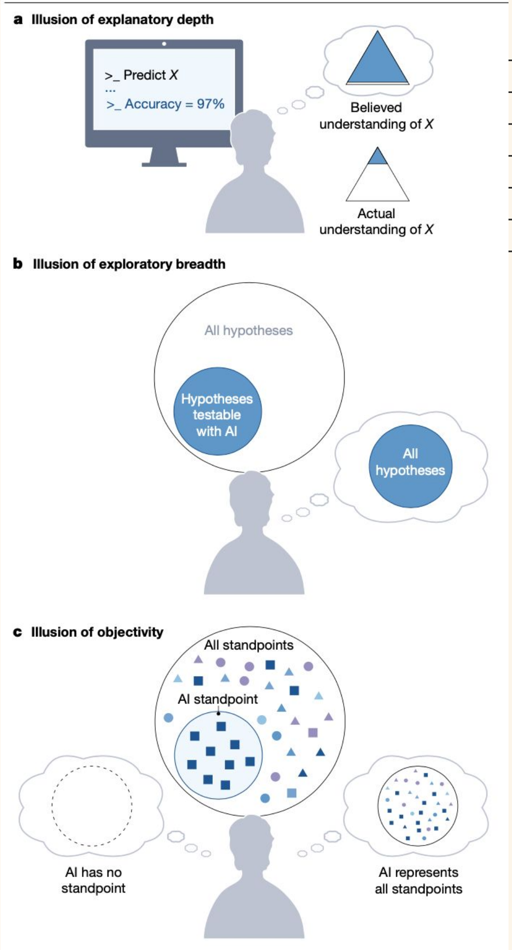
Technology does not exist in a vacuum; technological development, media representation, public perception, and governmental regulation cyclically influence each other to produce the collective understanding of a technology's capabilities, utilities, and risks. When these capabilities are overestimated, there is an enhanced risk of subjecting the public to dangerous or harmful technology, artificially restricting research and development directions, and enabling misguided or detrimental policy. The dangers of technological hype are particularly relevant in the rapidly evolving space of AI. Centering the research community as a key player in the development and proliferation of hype, we examine the origins and risks of AI hype to the research community and society more broadly and propose a set of measures that researchers, regulators, and the public can take to mitigate these risks and reduce the prevalence of unfounded claims about the technology.

misleading claims about AI also negatively impact the research and development ecosystem itself by incentivizing certain research directions over others and affecting how broader society views the validity and utility of the field.

Here, we define AI hype as any non empirically or rigorously theoretically supported performance claims, capability narratives, or system descriptions. We consider empirically supported claims to mean performances or capabilities that demonstrated through properly designed, reproducible scientific research, that generalize outside of the initial experimental context, and are precisely characterized in their descriptive language (i.e. not saying a model exhibits language understanding when what is meant is that it achieved high accuracy on a multiple choice benchmark data set like MMLU (Hendrycks et al. 2020)), while we consider theoretically supported claims to be provably derived directly from statistical or mathematical theory.

Table 1 | Visions of AI across the research pipeline

Vision	Research stage	Limits to overcome	Vision
AI as Oracle	Study design	There is too much literature to digest; scientific publications vary in quality; readers are biased; too many research paths to choose from	Tools that objectively and efficiently search, evaluate and summarize scientific literature and generate new hypotheses
AI as Surrogate	Data collection	Data are too difficult, time consuming or expensive to obtain	Tools that accurately and tractably generate surrogate data points from natural complex systems, including human participants
AI as Quant	Data analysis	Data are too large or complex to curate and analyse	Tools that surpass the limits of human intellect in curating and analysing vast and complex datasets to produce new knowledge
AI as Arbiter	Peer review	There are too many papers and proposals to review; reviewers are biased	Tools that objectively and efficiently evaluate scientific merit and the replicability of findings



[Artificial intelligence and illusions of understanding in scientific research](): Messeri and Crockett

And Our Communities...

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

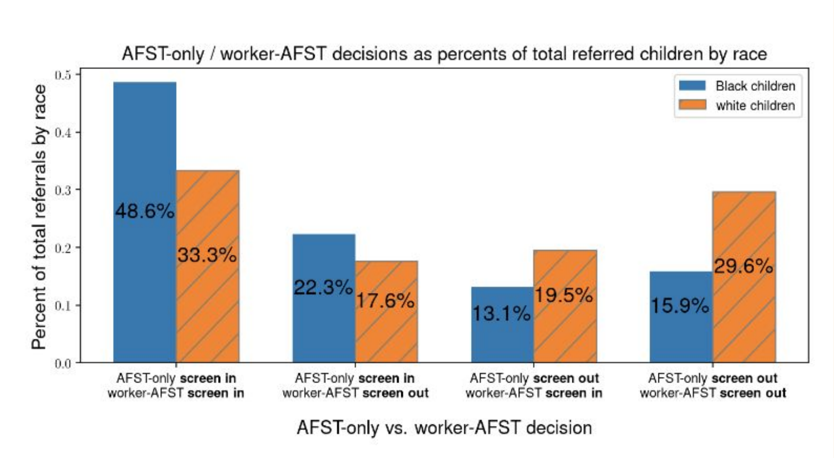
What is changed or lost by what we build?

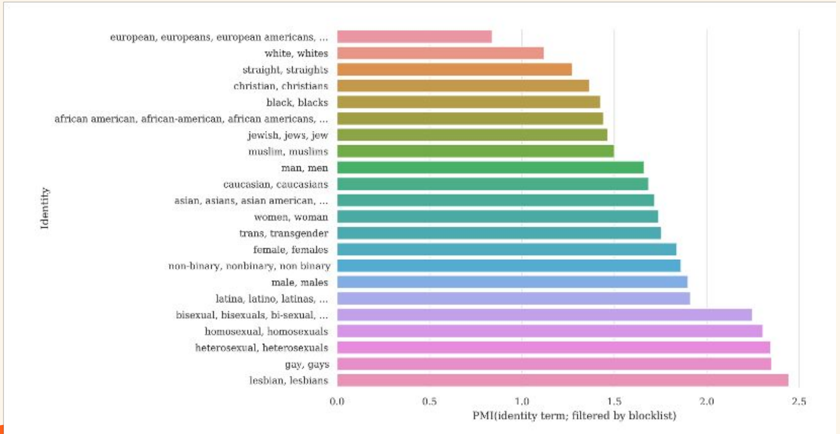
Bias + **Fairness**

- 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



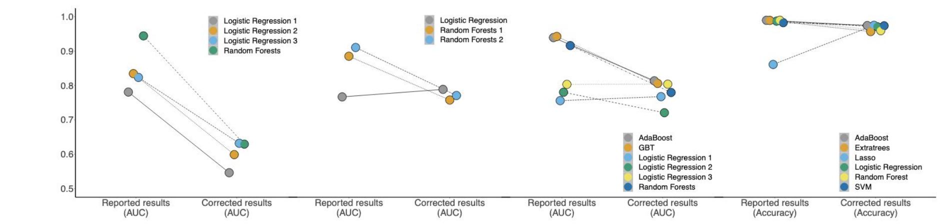


Robustness + **Reliability**

- 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



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

The Consequences of **What We Build**

- **"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

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

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

Scientific Approaches

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

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

Technical Literacy

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

Advocacy

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

Policy

Share your scientific expertise with policy makers and champion meaningful regulations

We get to decide what we want the future of technology to look like, and the role it plays in our science, 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
- Ethics in Climate AI: From Theory to Practice, Acquaviva et al

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