OTHER TOPICS IN AI ETHICS

Dr. Savannah Thais

Columbia University

LSSTC

03/03/2023

What Does AI Ethics Mean to You?

- What’s the most interesting or surprising thing you’ve learned today?
- How will you translate this into your work or actions?

Other Quantitative Topics

Data Stewardship

Your model is only as good as your training data

How much data is available and does each entry have the same information?

Do you have examples of all data classes/ranges?

How expensive is it to create/collect more data or labels?

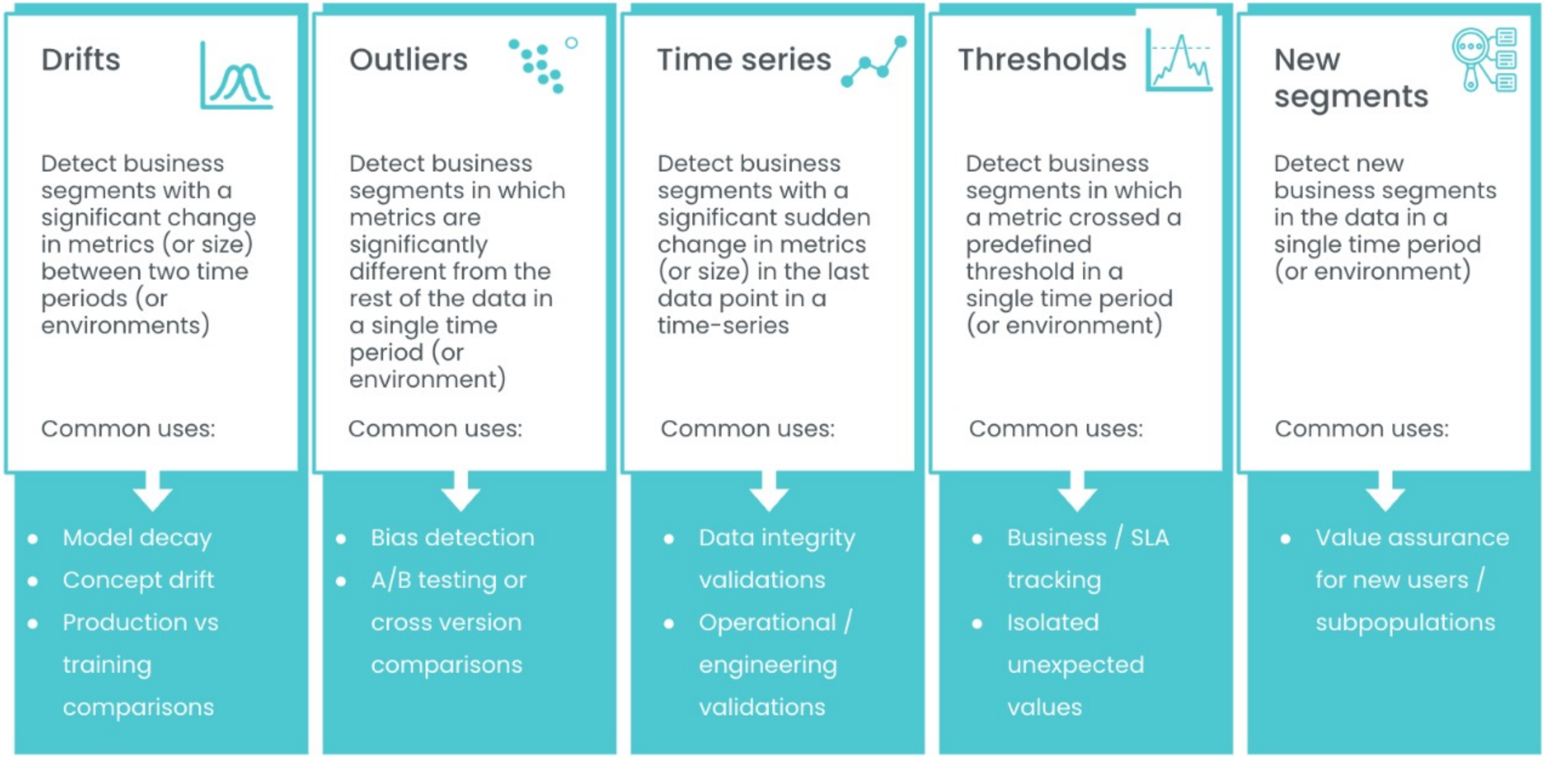
Are the available labels related to the decision you want to make?

Is there noise in your label creation or distribution?

Are classes and inputs balanced and normalized?

Are there patterns in your data you don’t want the model to exploit?

Data and Performance Modeling

- Unlike physics, real world systems aren’t static
  - Behavior, and therefore data, can shift over time
  - Models or measurements can effect the system you’re modeling
- It’s critical to continuously evaluate model performance
  - 

Drifts

Detect business segments with a significant change in metrics (or size) between two time periods (or environments)

Common uses:

- Model decay
- Concept drift
- Production vs training comparisons

Outliers

Detect business segments in which metrics are significantly different from the rest of the data in a single time period (or environment)

Common uses:

- Bias detection
- A/B testing or cross version comparisons

Time series

Detect business segments with a significant sudden change in metrics (or size) in the last data point in a time-series

Common uses:

- Data integrity validations
- Operational / engineering validations

Thresholds

Detect business segments in which a metric crossed a predefined threshold in a single time period (or environment)

Common uses:

- Business / SLA tracking
- Isolated unexpected values

New segments

Detect new business segments in the data in a single time period (or environment)

Common uses:

- Value assurance for new users / subpopulations

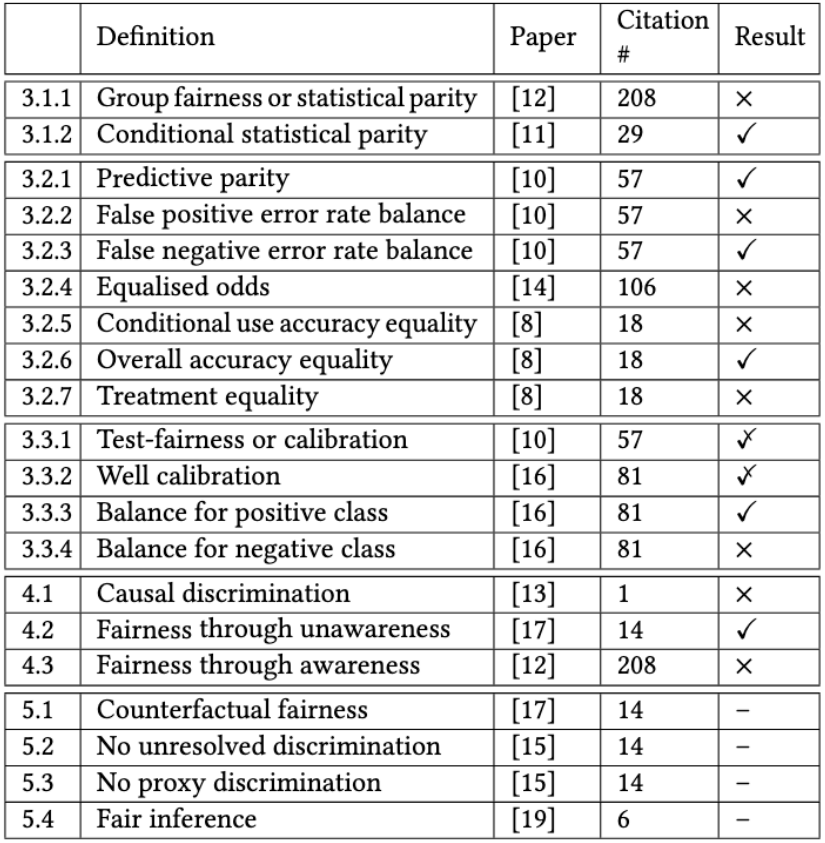
Further reading

source

Quantitative Fairness

- There is no universal definition of fairness
- Commonly used definitions
  - Individual fairness: any two individuals who are similar with respect to a task should receive similar outcomes
  - Group fairness: demographics of the individuals receiving any outcome - positive or negative - should be the same as demographics of the underlying population
- Definitions don't always account for pre-existing societal conditions or the fact that model predictions != model impacts

source

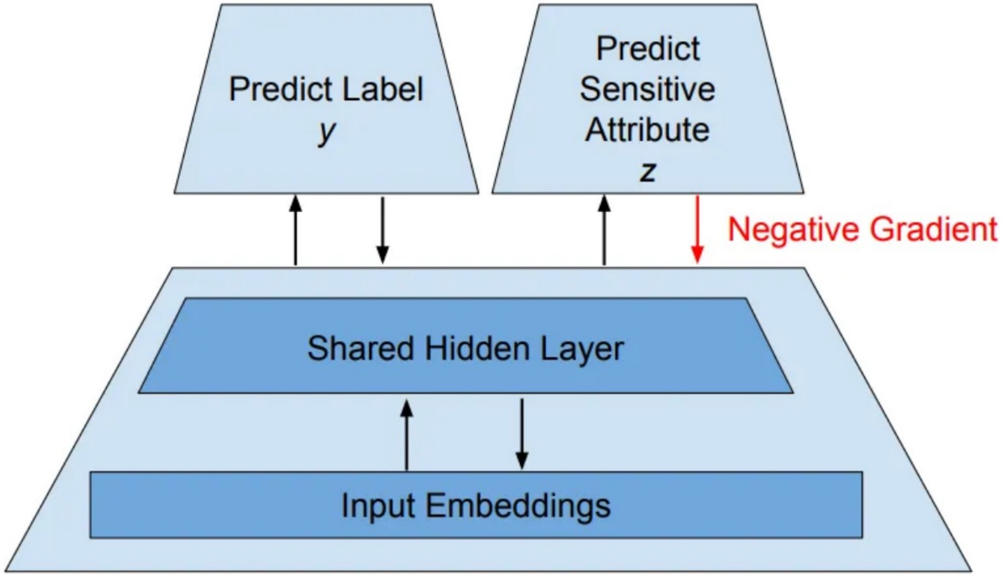


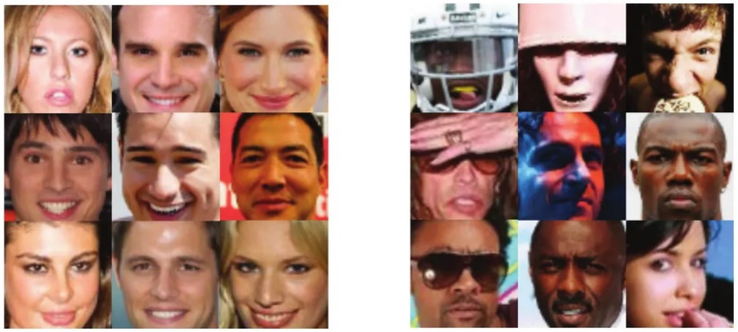
	Definition	Paper	Citation #	Result
3.1.1	Group fairness or statistical parity	[12]	208	×
3.1.2	Conditional statistical parity	[11]	29	✓
3.2.1	Predictive parity	[10]	57	✓
3.2.2	False positive error rate balance	[10]	57	×
3.2.3	False negative error rate balance	[10]	57	✓
3.2.4	Equalised odds	[14]	106	×
3.2.5	Conditional use accuracy equality	[8]	18	×
3.2.6	Overall accuracy equality	[8]	18	✓
3.2.7	Treatment equality	[8]	18	×
3.3.1	Test-fairness or calibration	[10]	57	✓✗
3.3.2	Well calibration	[16]	81	✓✗
3.3.3	Balance for positive class	[16]	81	✓
3.3.4	Balance for negative class	[16]	81	×
4.1	Causal discrimination	[13]	1	×
4.2	Fairness through unawareness	[17]	14	✓
4.3	Fairness through awareness	[12]	208	×
5.1	Counterfactual fairness	[17]	14	–
5.2	No unresolved discrimination	[15]	14	–
5.3	No proxy discrimination	[15]	14	–
5.4	Fair inference	[19]	6	–

Further reading

Debiasing

- If you are already aware of statistical bias in your data there may be ways to address it
  - Reweighting: adjust representation in dataset, add regularization terms
  - Adversarial debiasing: co-train an outcome prediction model and protected class prediction model
  - Synthetic data: Construct a more representative dataset for training





Random Batch Sampling During Standard Face Detection Training

Batch Sampling During Training with Learned Debiasing

Homogenous skin color, pose

Mean Sample Prob: 7.57×10^{-6}

Diverse skin color, pose, illumination

Mean Sample Prob: 1.03×10^{-4}

Further reading

Privacy

- ML models can memorize training data and make it possible for attackers to reverse engineer sensitive information
- Differential privacy: adds randomization to individual inputs in a way that the overall predictive distributions change minimally
- Federated learning: multiple sources download model, train locally, and summarize + reupload updated model

Did you go out drinking over the weekend?

let's call this property **P** (Truth=Yes) and estimate **p**, the fraction of the class for whom **P** holds

source

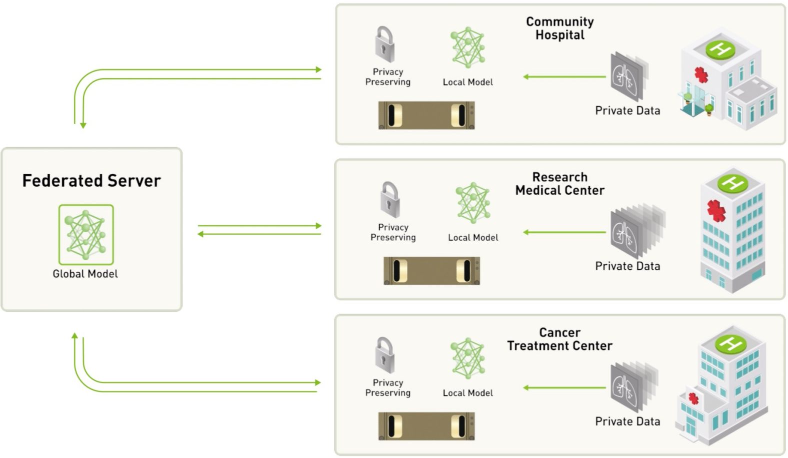
1. flip a coin **C1**
   1. if **C1** is tails, then **respond truthfully**
   2. if **C1** is heads, then flip another coin **C2**
      1. if **C2** is heads then **Yes**
      2. else **C2** is tails then respond **No**

randomization - adding noise - is what gives plausible deniability a process privacy method

the expected number of **Yes** answers is:

$$A=\frac{3}{4}p+\frac{1}{4}(1-p)=\frac{1}{4}+\frac{p}{2}$$

privacy comes from plausible deniability



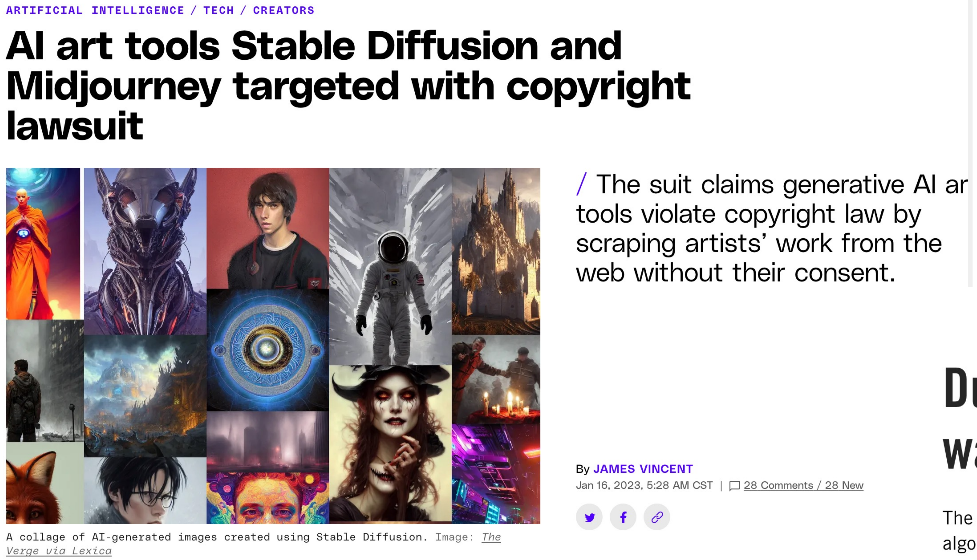
Further reading

Further reading

Other Non-Quantitative Topics

Regulation

- Lot’s of discussion around regulating AI but currently piecemeal approach in the US with many non-binding suggestions
  - NIST AI Risk Management Framework
  - GAO Accountability Framework for Federal Agencies and other Entities
  - DOD Ethical Principles for AI
  - OSTP Blueprint for AI Bill of Rights
- Unfortunately this makes regulation and action reactionary



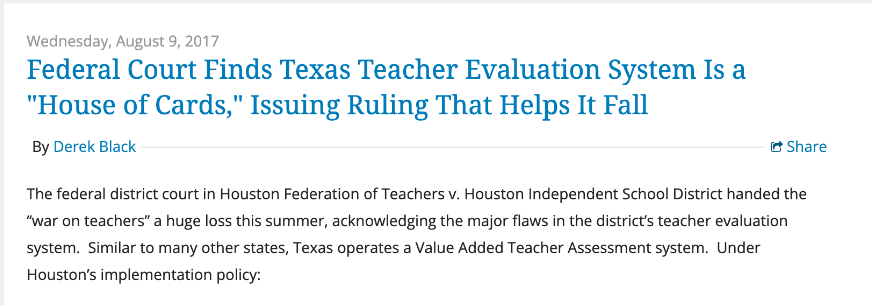
ARTIFICIAL INTELLIGENCE / TECH / CREATORS

AI art tools Stable Diffusion and Midjourney targeted with copyright lawsuit

/ The suit claims generative AI ar tools violate copyright law by scraping artists’ work from the web without their consent.

By JAMES VINCENT
Jan 16, 2023, 5:28 AM CST | 28 Comments / 28 New

A collage of AI-generated images created using Stable Diffusion. Image: The Verge via Lexica

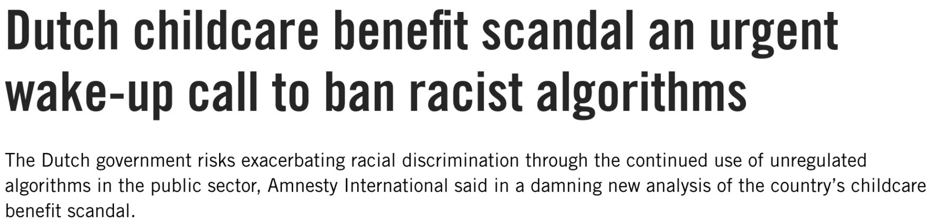


Wednesday, August 9, 2017

Federal Court Finds Texas Teacher Evaluation System Is a "House of Cards," Issuing Ruling That Helps It Fall

By Derek Black Share

The federal district court in Houston Federation of Teachers v. Houston Independent School District handed the “war on teachers” a huge loss this summer, acknowledging the major flaws in the district’s teacher evaluation system. Similar to many other states, Texas operates a Value Added Teacher Assessment system. Under Houston’s implementation policy:

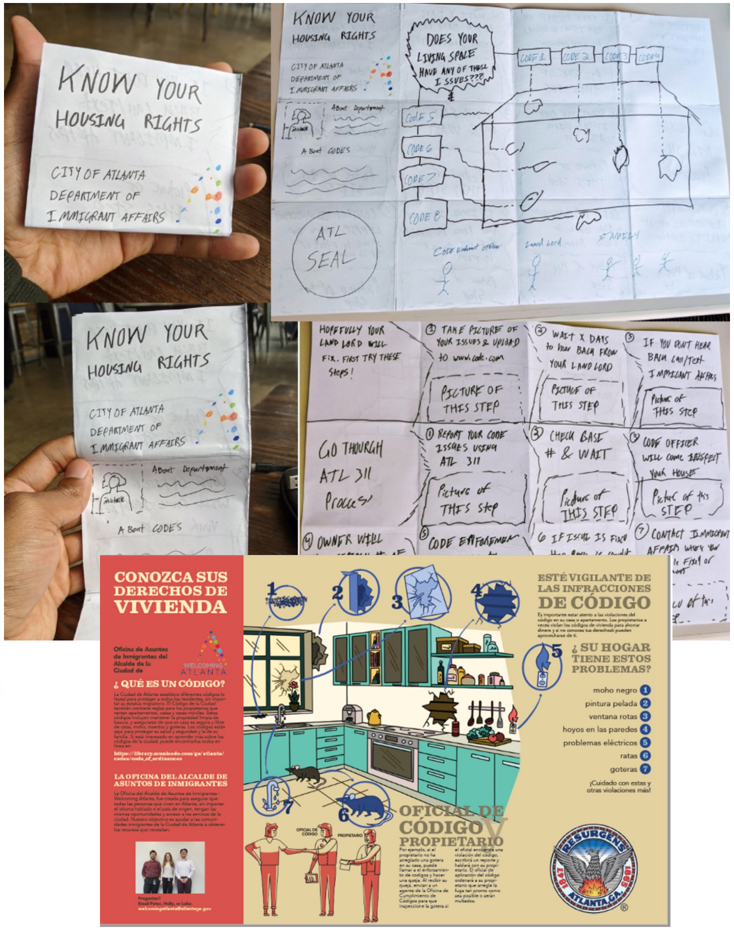


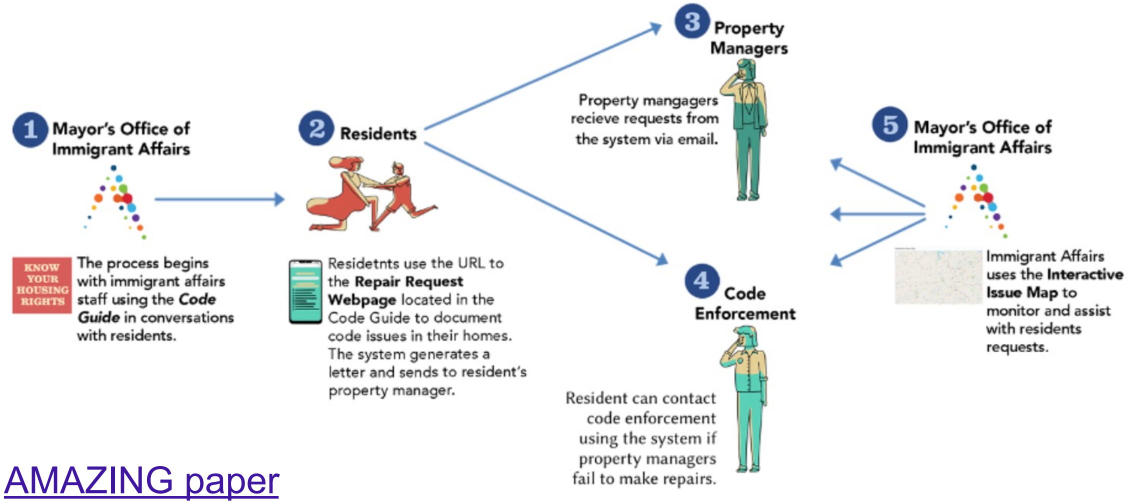
Dutch childcare benefit scandal an urgent wake-up call to ban racist algorithms

The Dutch government risks exacerbating racial discrimination through the continued use of unregulated algorithms in the public sector, Amnesty International said in a damning new analysis of the country’s childcare benefit scandal.

Participatory Design

- A democratic process of design of technology based on the argument that people should be involved in design of the systems they will be using (or subjected to)
- Can have huge benefits:
  - Understanding what stakeholders need
  - Identifying issues with planned technology
  - Fostering trust and adoption
  - Considering alternative solutions


KNOW YOUR HOUSING RIGHTS
CITY OF ATLANTA DEPARTMENT OF IMMIGRANT AFFAIRS
DOES YOUR LIVING SPACE HAVE ANY OF THESE ISSUES???
ATL SEAL
GO THROUGH ATL 311 PROCESS
CONOZCA SUS DERECHOS DE VIVIENDA
¿QUÉ ES UN CÓDIGO?
ESTÉ VIGILANTE DE LAS INFRACCIONES DE CÓDIGO
¿SU HOGAR TIENE ESTOS PROBLEMAS?
OFICIAL DE CÓDIGO
PROPIETARIO



1 Mayor's Office of Immigrant Affairs

The process begins with immigrant affairs staff using the ***Code Guide*** in conversations with residents.

2 Residents

Residetnts use the URL to the **Repair Request Webpage** located in the Code Guide to document code issues in their homes. The system generates a letter and sends to resident's property manager.

3 Property Managers

Property mangagers recieve requests from the system via email.

4 Code Enforcement

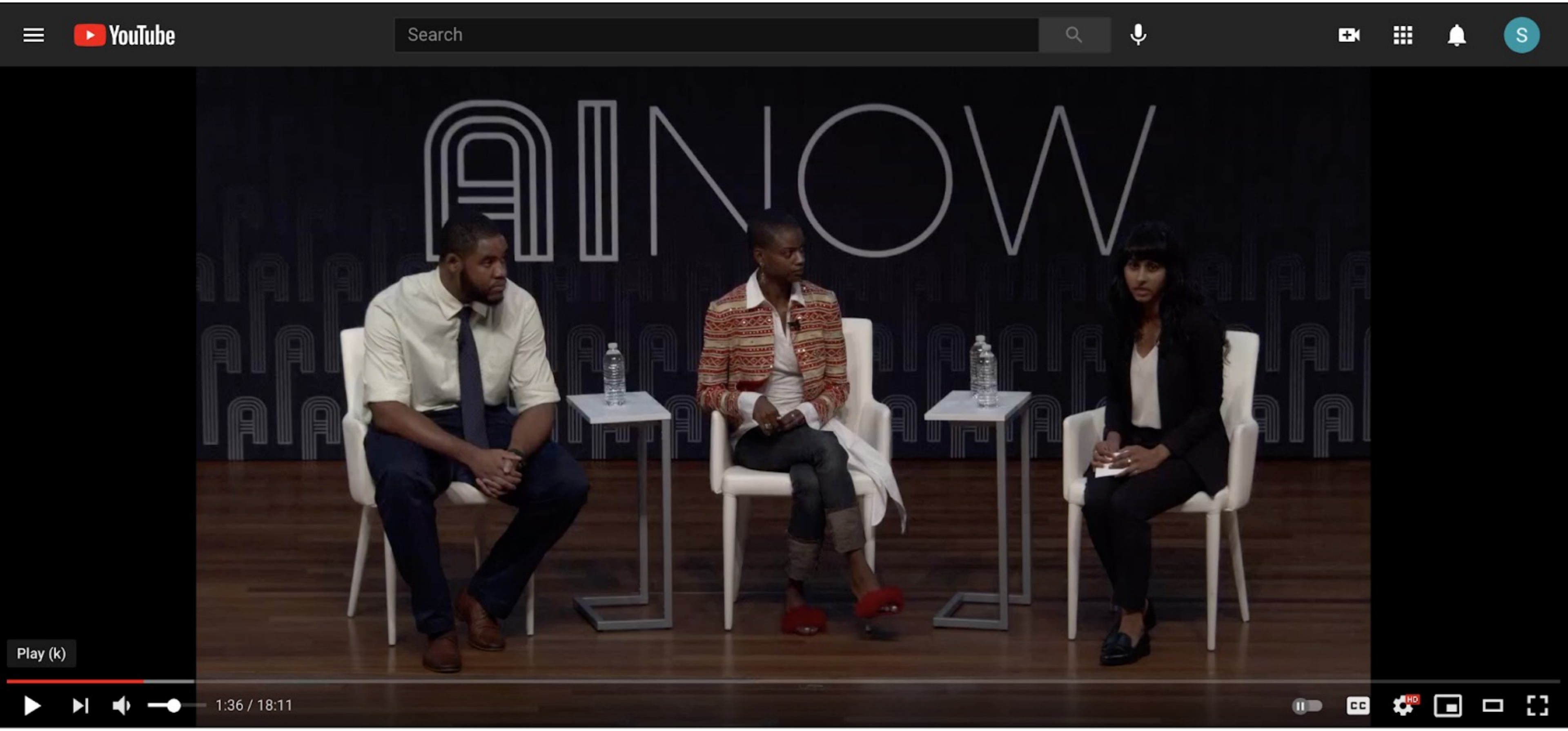
Resident can contact code enforcement using the system if property managers fail to make repairs.

5 Mayor's Office of Immigrant Affairs

Immigrant Affairs uses the **Interactive Issue Map** to monitor and assist with residents requests.

AMAZING paper

Collective Action

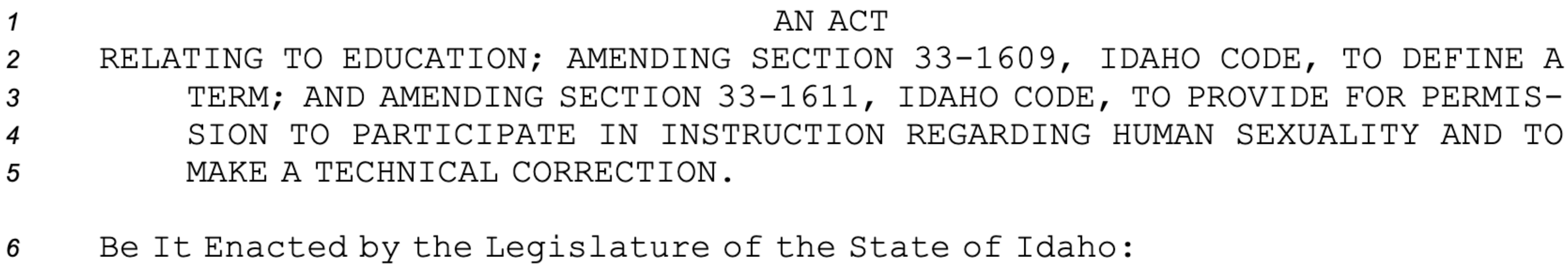


NYT Article

Some of My Current Work

New Types of Data

- Current line of work is investigating how to extract quantitative features from policy text
  - Cybersecurity recommendations to create a database to inform policy makers and understand state of the field
  - State level SexEd policies to inform causal modeling of key health outcomes
- Interesting features: specific actions, policy stringency, similarities, specific target groups
  - Challenge: policy text has a specific probability distribution that may differ from typical NLP corpuses
- Central question: how do you design evaluation metrics?
  - Partial answer: work with domain experts
  - How do you determine when your model is trustworthy enough?

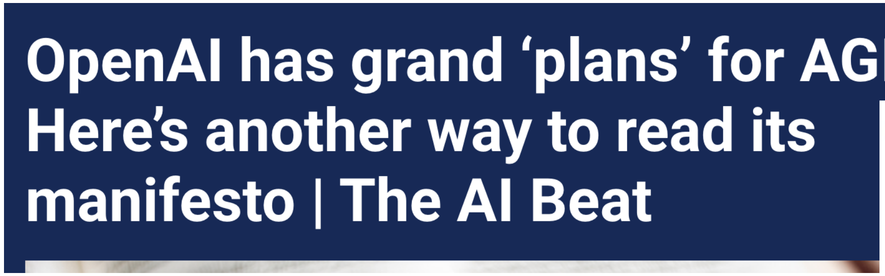


AN ACT

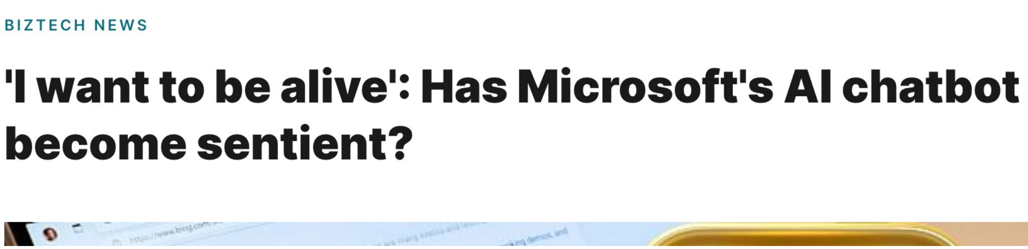
RELATING TO EDUCATION; AMENDING SECTION 33-1609, IDAHO CODE, TO DEFINE A TERM; AND AMENDING SECTION 33-1611, IDAHO CODE, TO PROVIDE FOR PERMISSION TO PARTICIPATE IN INSTRUCTION REGARDING HUMAN SEXUALITY AND TO MAKE A TECHNICAL CORRECTION.

Be It Enacted by the Legislature of the State of Idaho:

AI Has a Hype Problem

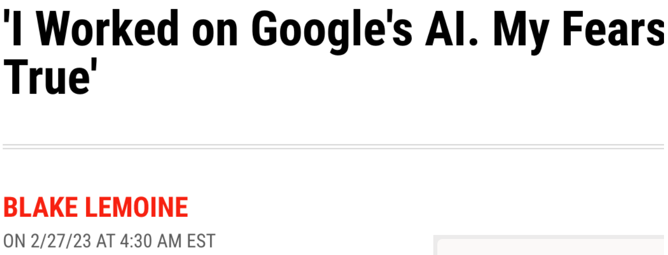


OpenAI has grand ‘plans’ for AGI. Here’s another way to read its manifesto | The AI Beat



BIZTECH NEWS

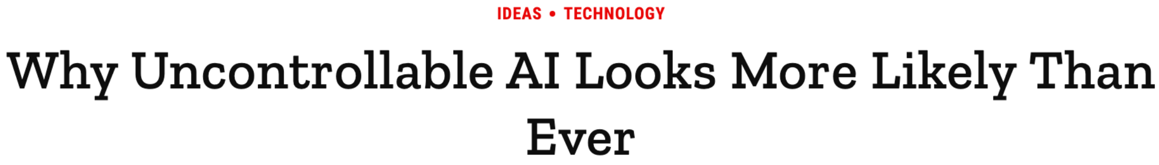
'I want to be alive': Has Microsoft's AI chatbot become sentient?



'I Worked on Google's AI. My Fears True'

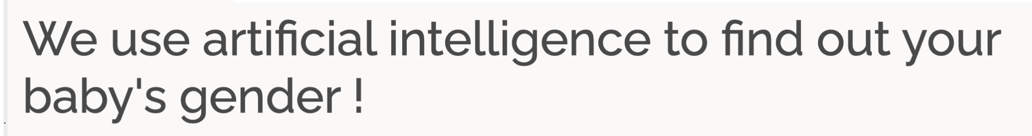
BLAKE LEMOINE

ON 2/27/23 AT 4:30 AM EST



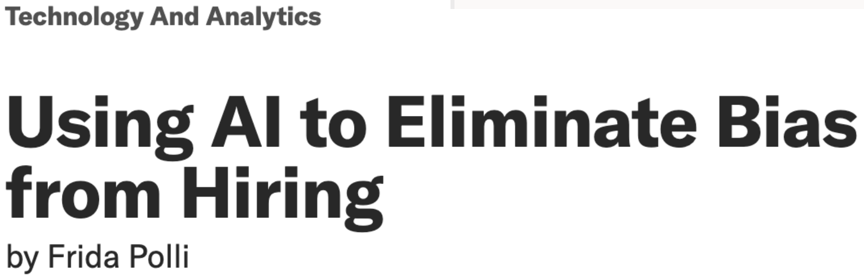
IDEAS • TECHNOLOGY

Why Uncontrollable AI Looks More Likely Than Ever



We use artificial intelligence to find out your baby's gender !

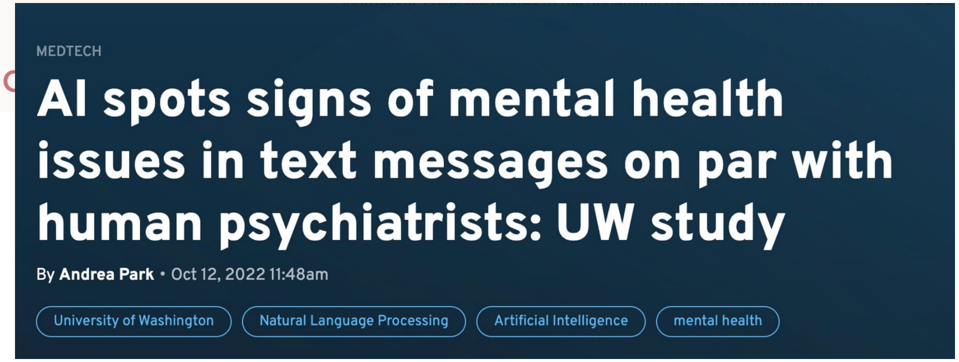
Today, artificial intelligenc



Technology And Analytics

Using AI to Eliminate Bias from Hiring

by Frida Polli



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

The Danger to Treating Math as Magic

- Hype prevents us from interrogating imagined futures
  - We define problems by presenting technology as a solution
- Allows us to subject people to unscientific and inaccurate sociotechnical systems
  - Can have tangible life altering consequences
  - Can rapidly entrench or enhance existing biases
  - Can push responsibility for harm onto the user who inherently has less control
- Artificially limits research directions
  - Focus on scale approaches, reduced interest in theory
  - Belief we have solved problems we haven’t

Scientific ML Beyond Science

Reframing how we think about ML and treating it more scientifically can help address some of these issues

- What function are we trying to approximate?
  - Can we mathematically define the outcome we're interested in?
  - How well does the function the algorithm is learning approximating the function we're interested in? Can we even measure this?
  - Are there impacts, correlations, or data points we want to avoid?
- Where does the training data come from?
  - How is it stored? Can it be reused? Who owns it?
- How are these systems designed and built?
  - Do users/impactees opt in?
  - Who decides the function to be learned and what is 'good enough'?
  - How are they deployed? Who reaps the benefits and who is negatively impacted?

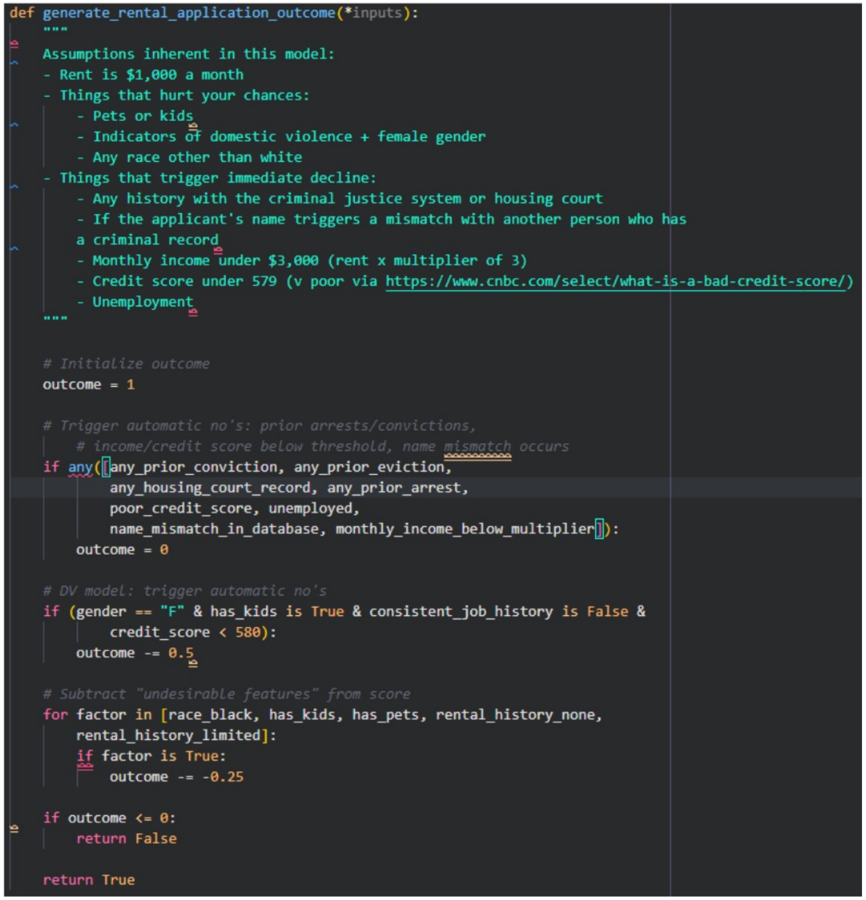
Other Levers for Change

These are sociotechnical issues and cannot be solved by quantitative approaches alone

- Diversified funding and training mechanisms
  - Shift power concentrations in the ML space
- Changes to the research ecosystem
  - Require heavier burden of proof and more rigorous testing
  - Assess stakeholder value before development
  - More representative benchmarks that are more closely related to real world tasks
  - Interdisciplinary collaboration putting different types of knowledge on the same level
- Regulatory actions
  - Require standardized testing for high stakes applications (a la FDA)
  - Improved transparency into training data, model development, and testing (informed consent for participation)
- Improved technical literacy
  - For reporters, advocacy organizations, and the general public

Technical Literacy to Shift Power

- Grassroots activism can be an important tool for developing safer technology
    - How do we effectively teach people about algorithmic systems so that they can meaningfully consent or resist?
- Initial project building interactive story telling around rental screening tools
    - Help build intuition around these tools and identify modes of resistance/recourse
- Developing a survey study with mutual aid networks in Brooklyn to understand needs and interests of activists
    - How can we as technologists be the most useful in supporting community work (participatory design)



```python
def generate_rental_application_outcome(*inputs):
    """
    Assumptions inherent in this model:
    - Rent is $1,000 a month
    - Things that hurt your chances:
        - Pets or kids
        - Indicators of domestic violence + female gender
        - Any race other than white
    - Things that trigger immediate decline:
        - Any history with the criminal justice system or housing court
        - If the applicant's name triggers a mismatch with another person who has
        a criminal record
        - Monthly income under $3,000 (rent x multiplier of 3)
        - Credit score under 579 (v poor via https://www.cnbc.com/select/what-is-a-bad-credit-score/)
        - Unemployment
    """

    # Initialize outcome
    outcome = 1

    # Trigger automatic no's: prior arrests/convictions,
        # income/credit score below threshold, name mismatch occurs
    if any([any_prior_conviction, any_prior_eviction,
            any_housing_court_record, any_prior_arrest,
            poor_credit_score, unemployed,
            name_mismatch_in_database, monthly_income_below_multiplier]):
        outcome = 0

    # DV model: trigger automatic no's
    if (gender == "F" & has_kids is True & consistent_job_history is False &
            credit_score < 580):
        outcome -= 0.5

    # Subtract "undesirable features" from score
    for factor in [race_black, has_kids, has_pets, rental_history_none,
        rental_history_limited]:
        if factor is True:
            outcome -= -0.25

    if outcome <= 0:
        return False

    return True
```

Open Discussion!

✉ st3565@columbia.edu @basicsciencesav