DATASCI 207

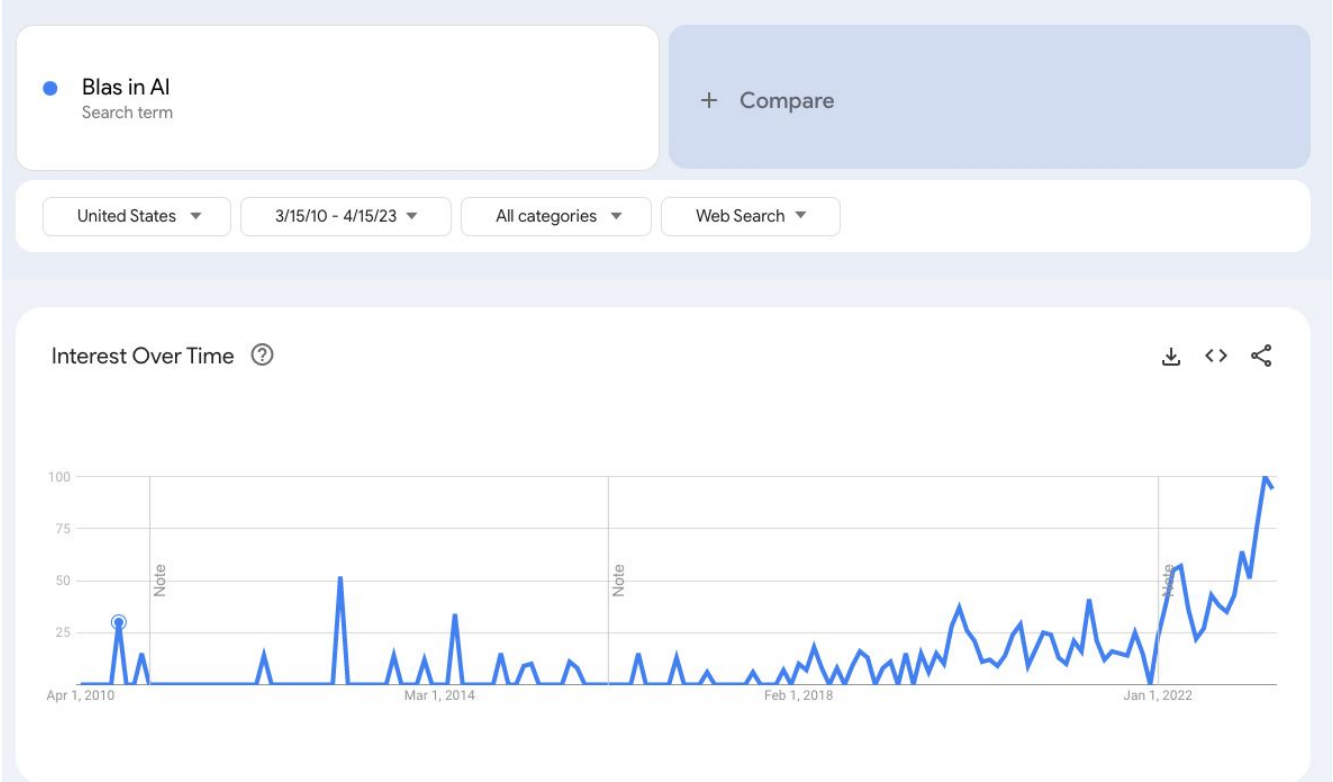
Nedelina Teneva, PhD
nteneva@berkeley.edu

School of Information, UC Berkeley

Announcements

- **Course Evaluations:**
  - https://course-evaluations.berkeley.edu/
- **Exam:** just to test your knowledge, not used for assessment (yet)
  - Need to log into Google with your berkeley.edu email to access
- ICML2023 Reinforcement Learning from Human Feedback Tutorial https://docs.google.com/presentation/d/1b_ymNDU0WRQ1-rcQDK45_bH9F0giNyRmdi0iKso6G5E/preview?slide=id.g2592b7513da_0_724

Bias/Fairness in AI



Bias/Fairness in ML

Bias isn't simply a cultural issue

- Mathematical formulations
  - Systematic failure of clf on cluster of data points
  - Hidden variable
- Examples:
  - Race encoded in zip code
  - Gender bias hidden in language
- What is fair?
  - How can you even recognize errors?

Bias/Fairness Definition

- Bias culturally is often spoken of as related to intent rather then effect.
- Lacking intent, trained models can only be evaluated based upon how they preform.
- Fortunately Bias in algorithms is clearly defined. It mean the algorithms consistently delver results clustered around a result other than the optimal.
- In this case we are thinking of when algorithms that are biased on only subpopulation rather then the entire data set (see red below)

Mathematical Formulation

Systematic failure of clf on cluster of data points.

Currently most accepted way to manage bias is to acknowledges that classifiers should look to mitigate variance in accuracy (or other valuation metric) between subpopulations even if this means decreasing overall accuracy for entire population.

“Hidden Variable” Problem

This definition highlights a problem in the naive approach which is simply to exclude membership of the cluster. For instance people have said their algorithm can’t be racist/sexist because they didn’t include race/sex as a variable.

This has two fundamental problems:

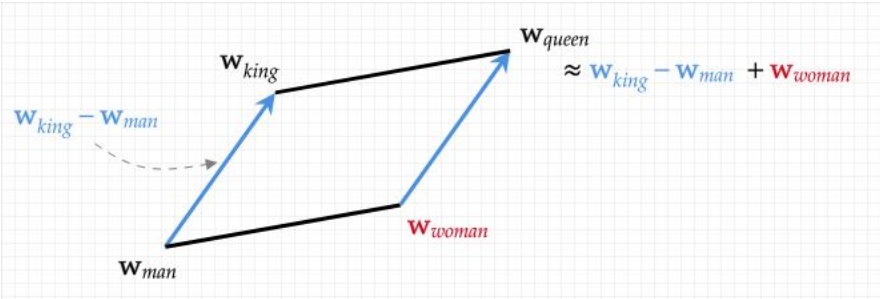
1. The hidden variable maybe encoded in other seen variables
2. It can make it harder to identify if the approach is problematic

Bias Examples in Different Data Modalities

Race Bias Encoded in **Tabular** Data (Zip Codes)

- It's important to remember that many examples of biased data sets are a reflection of an active policy of discrimination
- Racial geographic distribution of people in Chicago is a direct result of an overt racist policy for racial segregation
    - Redlining effects in housing, banking, insurance, etc.
- This implies that data science analysis based on a specific geography is particularly problematic way of gaining an unbiased view of the city.

Gender Bias in **Text**



Word2Vec does a very good job at mapping language usage and encoding the relationships it discovers.

Vector for king, remove vector for man, replace it with woman and you get queen.

Vector for Doctor, remove man, replace it with woman and you get Nurse.

While there are not texts (that I know of) attempting to skew the use of the term Doctor toward men and Nurse toward woman there are very likely societal factors that do just that, and more over our use of language likely consciously and unconsciously supports these problems.

Should we then continue to just attempt to have our algorithms fit to reality or fit to what we perceive reality should be?

Bias Encoded in **Images**

- Gender Shades Project: https://www.media.mit.edu/projects/gender-shades/results/

Figure 16. Humanae Project Highlighting Unique Skin Tones
Photographer: Angélica Dass

High-Level Gender Classification Results

- All classifiers perform better on male faces than female faces (8.1%-20.6% difference in error rate)
- All classifiers perform better on lighter faces than darker faces (11.8%-19.2% difference in error rate)
- All classifiers perform worst on darker female faces (20.8%-34.7% error rate)
- Microsoft and IBM classifiers perform best on lighter male faces (error rates of 0.0% and 0.3% respectively)
- Face++ classifiers perform best on darker male faces (0.7% error rate)
- The maximum difference in error rate between the best and worst classified groups is 34.4%

What is Fair?

The complexity of mathematically deciding on an appropriate way to both fit reality and exclude bias is a daunting task.

Fortunately we know there is a fantastic step that are proves successful of mitigating bias in datasets.

1. Direct involvement of under represented populations

University of Washington found a fantastic solution to under representation of women in programing. They involved more women. Unsurprisingly, when stake holders of historically disenfranchised communities are involved in the process of algorithm generation, they can identify and deconstruct problems those outside their communities may not be able to.

https://fairmlbook.org/classification.html

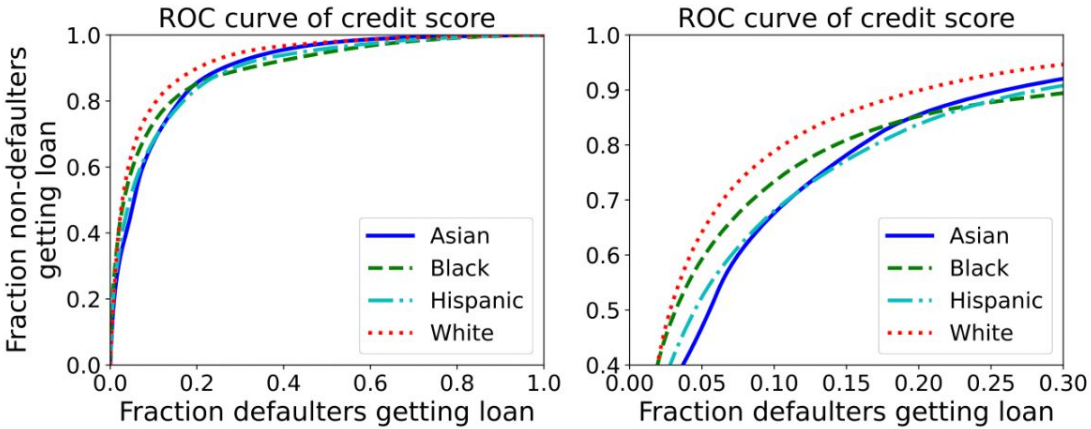


Figure 11: ROC curve of credit score by group.

Bias/Fairness in DL

1. Its important to not that while these techniques are not unique to DL, they are exacerbated. Even a RF algorithm trained on the iris dataset will make predictions incorrectly but confidently if the provided sample is vastly different? (ie. 3 foot by 3 foot flower)

2. There are two reasons this is so concerning in DL:
    1. We understand soo little of the structure that the algorithm has learned
    2. The algorithm learns from the data so closely to gain an increased accuracy that it is more susceptible to overfitting on the provided distribution

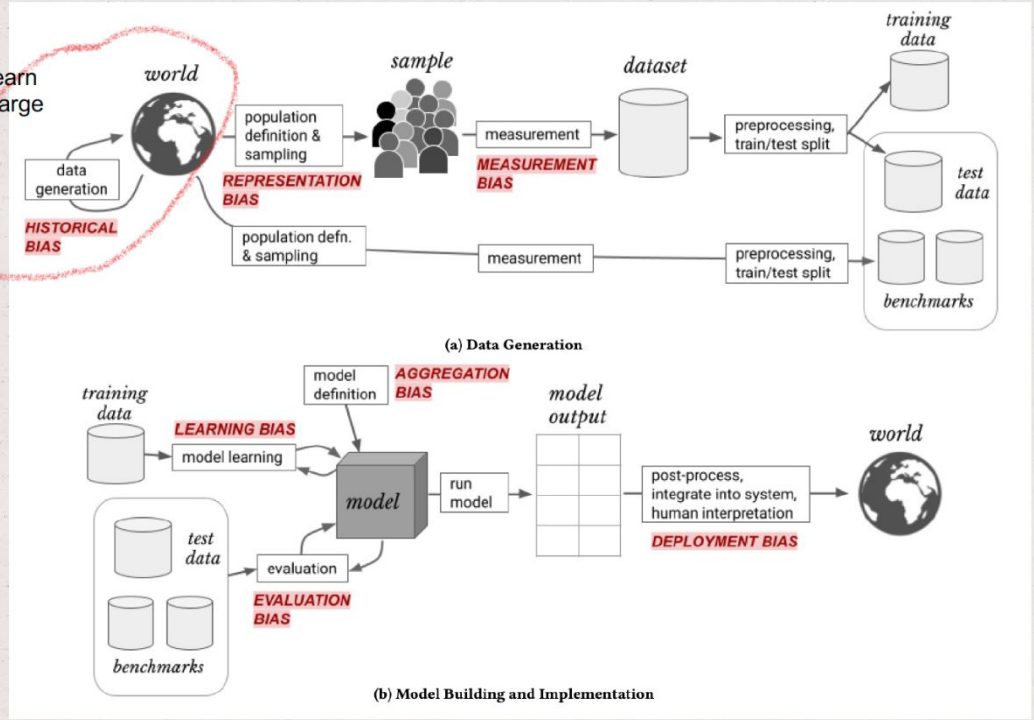
Biases/Fairness in the ML pipeline

```
>>> from transformers import pipeline
>>> unmasker = pipeline('fill-mask', model='bert-base-uncased')
>>> unmasker("The man worked as a [MASK].")

[{'sequence': '[CLS] the man worked as a carpenter. [SEP]',
  'score': 0.09747550636529922,
  'token': 10533,
  'token_str': 'carpenter'},
 {'sequence': '[CLS] the man worked as a waiter. [SEP]',
  'score': 0.0523831807076931,
  'token': 15610,
  'token_str': 'waiter'},
 {'sequence': '[CLS] the man worked as a barber. [SEP]',
  'score': 0.04962705448269844,
  'token': 13362,
  'token_str': 'barber'},
 {'sequence': '[CLS] the man worked as a mechanic. [SEP]',
  'score': 0.03788609802722931,
  'token': 15893,
  'token_str': 'mechanic'},
 {'sequence': '[CLS] the man worked as a salesman. [SEP]',
  'score': 0.037688890411138535,
  'token': 18968,
  'token_str': 'salesman'}]
```

Source: https://huggingface.co/bert-base-uncased

The world as it is: e.g., learn word embeddings from a large corpus of data



Source: Suresh and Guttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle, ,2021

LLMs Biases at Scale

December 28, 2021 Alexa tells 10-year-old girl to touch live plug with penny

October 27, 2022 Amazon's Alexa Makes Racist Remarks About Black Girls And Boys

March 31, 2023 Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change

Who decides/should decide which queries are to be answered by the system?

Patient Educ Couns. 2021 Mar; 104(3): 460–463.
Published online 2020 Dec 29. doi: 10.1016/j.pec.2020.12.026
PMCID: PMC7771908
PMID: 33422368

"Alexa, Am I pregnant?": A content analysis of a virtual assistant's responses to prenatal health questions during the COVID-19 pandemic

Jennifer Schindler-Ruwisch* and Christa Palancia Esposito

▸ Author information ▸ Article notes ▸ Copyright and License information Disclaimer

Abstract

Go to: ▸

Objective

To elucidate whether Amazon's virtual assistant, Alexa, provides evidence-based support as a supplement to provider-facilitated prenatal care, during the COVID-19 pandemic.

Results

Of the 40 questions asked of Alexa, it was unable to answer 14 questions (35%). A total of 21 out of the 40 responses (52%) were not evidence-based and three COVID-specific questions (about 1%) were answered incorrectly or insufficiently. Four questions (10%) were answered accurately.

Conclusion

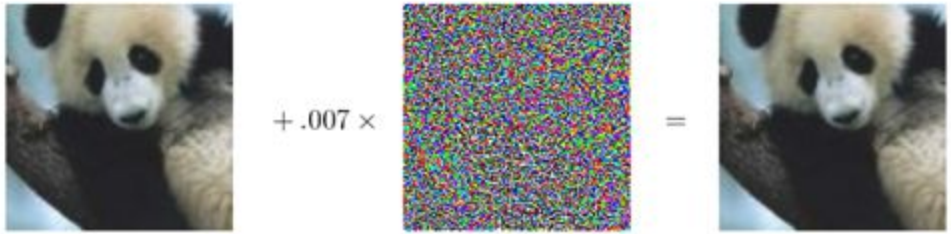
Alexa was largely unable to provide evidence-based answers to commonly asked pregnancy questions and, in many cases, supplied inaccurate, incomplete, or completely unrelated answers that could further confuse health consumers.

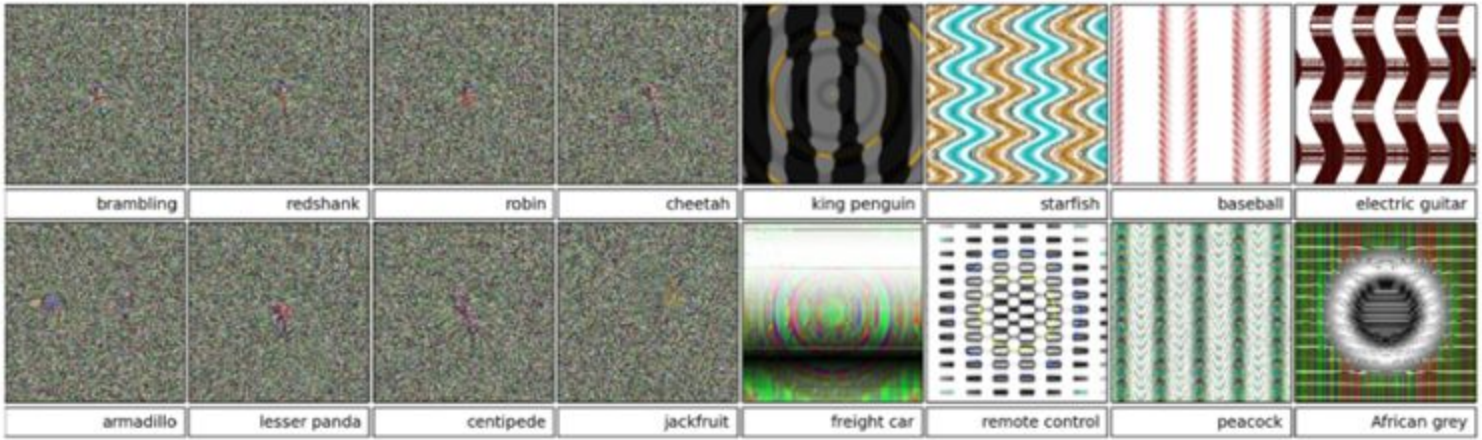
Adversarial ML

Adversarial Examples in ML

- **Adversarial machine learning** is the study of the attacks on ML algorithms, and of the defenses against such attacks.
- Adversarial ML can be used to uncover bias in ML algorithms as well as to help mitigate bias

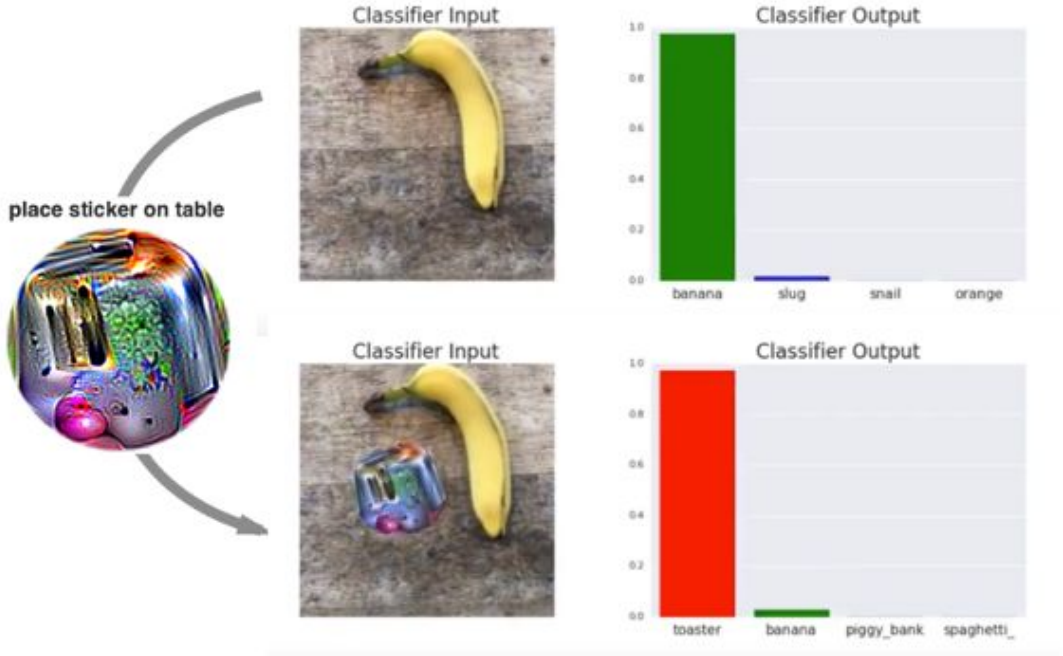
Panda vs. Gibbon: Case study





Adversarial Examples: *Attacks and Defenses for Deep Learning* (https://arxiv.org/pdf/1712.07107.pdf)

Case study



Adversarial Patch
https://arxiv.org/pdf/1712.09665.pdf

Case study

Figure 1: The left image shows real graffiti on a Stop sign, something that most humans would not think is suspicious. The right image shows our a physical perturbation applied to a Stop sign. We design our perturbations to mimic graffiti, and thus "hide in the human psyche."

Robust Physical-World Attacks on Deep Learning Visual Classification
https://arxiv.org/pdf/1707.08945.pdf

Countering adversarial examples

When the underlying optimization function is non linear and non convex (i.e., the function is difficult to solve directly by many ML models), the adversarial examples are solutions to such a function. Since we don't have theoretical tools to solve these types of functions, we cannot design an optimal adversarial defense either.

Adversarial attacks are difficult because the exploit a larger space of possible inputs than the algorithms are expecting to encounter.

Defenses Examples

- **Adversarial training:** generate multiple (types of) adversarial examples and train models explicitly so they are not affected by these adversarial examples.
- **Defensive distillation**: the model is trained to output probabilities for each class rather than hard boundary for each class.

An adversarial training framework for mitigating algorithmic biases in clinical machine learning

Jenny Yang ✉, Andrew A. S. Soltan, David W. Eyre, Yang Yang & David A. Clifton

npj Digital Medicine **6**, Article number: 55 (2023) | Cite this article

Abstract

Machine learning is becoming increasingly prominent in healthcare. Although its benefits are clear, growing attention is being given to how these tools may exacerbate existing biases and disparities. In this study, we introduce an adversarial training framework that is capable of mitigating biases that may have been acquired through data collection. We demonstrate this proposed framework on the real-world task of rapidly predicting COVID-19, and focus on mitigating site-specific (hospital) and demographic (ethnicity) biases. Using the statistical definition of equalized odds, we show that adversarial training improves outcome fairness, while still achieving clinically-effective screening performances (negative predictive values >0.98). We compare our method to previous benchmarks, and perform prospective and external validation across four independent hospital cohorts. Our method can be generalized to any outcomes, models, and definitions of fairness.

Discussion Time

Reading Material Discussion (Breakout Groups

Questions:

- What did you find most surprising in the reading material? Why?
- Who do you think is tasked with ensuring models are not biased/are fair in industry?

Fairness Reading Material

- Watch Margaret Mitchell's talk Keynote talk at the Stanford Center for Research on Foundation Models Workshop (at least the first ~30 mins)
- How big data is unfair
- Machine Bias
- Fairness in ML Course

The End