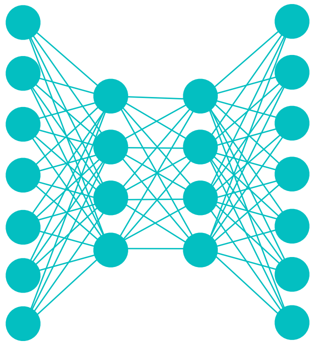
Lecture Notes for

Neural Networks and Machine Learning

The Ethical AI Principles and Case Studies in Ethical ML



Logistics and Agenda

- Logistics
  - Panopto and course videos
  - First Student Presentation next time to start lecture
  - Student Presentations
    - Still need responses, ASAP!
    - **Alternative:** can submit video summary, rather than presentation
- Last Time:
  - Course Introduction
  - *Strong AI*
- Agenda
  - The AI Principles and Fairness measures
  - Case Studies and Discussion
    - Applying the Principles

Ethical Principles in ML

From Australian Government, Department of Science

- **Reliability**: does system operate in accordance with intended purpose?
- **Fairness**: will system be inclusive and accessible? Will it involve or result in unfair discrimination against individuals, communities, or groups?

Model Measurement and Objective Alignment

- **Beneficence**: does system benefit individuals, society, or environment?
- **Respect**: does system respect human rights and autonomy of individuals?

Forethought and Insight

- **Privacy**: will system respect and uphold privacy rights and data protection, and ensure the security of data?
- **Transparency**: will system ensure people know when they are engaging with an AI system? Or know if significantly impacted?
- **Contestable**: will there be a timely process to allow people to challenge the use or output of the AI system?

Deployment Design

- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

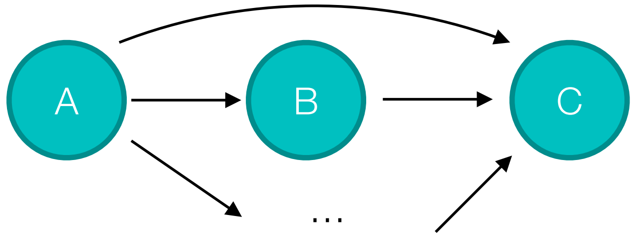
Organizational Structure

https://www.industry.gov.au/data-and-publications/building-australias-artificial-intelligence-capability/ai-ethics-framework/ai-ethics-principles

Measuring Reliability and Fairness

- Identify potential bias, groups defined by attribute "A"
- **Fairness through unawareness**, no knowledge of A:

$f(\mathscr{X})_{\setminus A} \rightarrow \mathscr{Y}$ (omission of data)



- **Individual Fairness**, similar individuals are classified similarly: $d(i,j) < \epsilon \rightarrow f(\mathscr{X}^{(i)}, A^{(i)}) \approx f(\mathscr{X}^{(j)}, A^{(j)})$
  - where d is a measure of if i,j individuals are similar

Defining which individuals should be close is typically incredibly difficult or expensive to collect…

Measuring Reliability and Fairness

- **Demographic parity**: $f(\mathscr{X} \mid A = 0) \approx f(\mathscr{X} \mid A = 1)$
    - Attribute should never influence outcomes…
- **Equal Opportunity**: Positive class not influenced by A

$$f(\mathscr{X} \mid A = 0, Y = 1) \approx f(\mathscr{X} \mid A = 1, Y = 1)$$

Can be good in many situations, but tend to decrease performance when some groupings should influence outcomes

- **Counterfactual fairness**: $\lfloor f(X_a) \rfloor = \lfloor f(X_{a'}) \rfloor$ for a given set of groups, a and a'
- **Minimum Difference**: Minority class confidences distribution should match majority

Kusner et al. "Counterfactual Fairness" in Proceedings of Neurips 2017

Counter Factual and MinDiff

- Identify: measure differences in reliability for identified groups, measure **statistical difference** and **impact**
- Develop examples of interest with counterfactual fairness,
  - original example: features with X_a where A=a
  - counterfactual: features with $X_{a'}$ where A=a` *and outcome should not change, expert judged*
- Counterfactual loss:
  - $\mathscr{L}_{cf} = ||f(X_a) - f(X_{a'})||^2$ **or other measure of closeness**
  - $\mathscr{L}_{tot} = \mathscr{L}_{bce} + \lambda \cdot \mathscr{L}_{cf}$
- **Min Diff**, define two groups, a,b that should be similar:

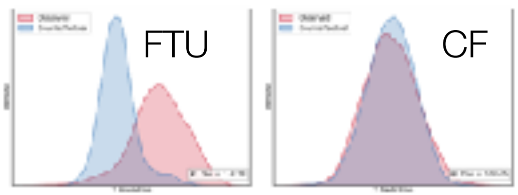
$$\mathscr{L}_{md} = \mu(f(X_a)) - \mu(f(X_b))$$

A result on common datasets

Base and Unaware | **Counter Factual Training or EO**

Synthetic Loan Data

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	$FLAP_1$ (O)	$FLAP_2$ (O)	$FLAP_1$ (M)	$FLAP_2$ (M)	OB_1	OB_2
↑ ACC	0.6618	0.6481	**0.6224**	0.6237	0.6224	0.6237	0.6224	0.6237	0.6224	**0.6406**	0.6279
↑ AUC	0.9457	0.8986	0.5867	**0.6682**	0.5714	0.5668	0.5837	0.5875	0.5863	0.5704	0.5856
↓ CF-metrics	0.6291	0.3906	0.0031	0.0355	0.0034	0.0016	0.0032	**0.0002**	**0.0002**	0.0011	0.0026
↓ CF Bound	0.8690	0.9464	0.1836	0.1071	0.0918	0.0937	0.1847	0.0690	**0.0670**	0.0830	0.2340
↓ EO Fairness	0.5469	0	0.0156	**0**	0.0336	0.0321	0.0156	0.0301	0.0180	**0**	**0**
↓ AA Fairness	0.6235	0.4559	**5.6e-18**	0.0370	**1.1e-18**	**3.3e-18**	**6.7e-18**	0.0012	0.0038	**4.6e-17**	**4.3e-17**



Two groups identified and their distributions, KL measure difference. Lower diff is better.

COMPAS Data: who will reoffend in next two years

Metrics	Baselines		Compared Methods							Ours	
	ML	FTU	FL	EO	AA	$FLAP_1$ (O)	$FLAP_2$ (O)	$FLAP_1$ (M)	$FLAP_2$ (M)	OB_1	OB_2
↑ ACC	0.5744	0.5726	0.5598	**0.5710**	0.5609	0.5605	0.5599	0.5607	0.5607	0.5666	0.5674
↑ AUC	0.7206	0.7225	0.6928	0.7225	0.6927	0.6927	0.6928	0.7015	0.7019	**0.6764**	**0.6744**
↓ CF-metric	0.2274	0.1406	0.0054	0.1377	0.0060	0.0058	0.0054	**0.0026**	0.0027	0.0060	0.0065
↓ EO Fairness	0.1046	0	0.1374	**0**	0.1405	1.7e-06	3.3e-06	6.7e-07	1.2e-06	**0**	**0**
↓ AA Fairness	0.2258	0.1460	**0**	0.1424	**0**	2.9e-07	5.6e-07	8.2e-07	3.0e-07	1.6e-16	1.1e-16

https://arxiv.org/pdf/2403.17852v1 Chen and Zhu, Counterfactual Fairness through Transforming Data Orthogonal to Bias, 2024

Fairness and downstream influence

Timnit Gebru
@timnitGebru

I'm sick of this framing. Tired of it. Many people have tried to explain, many scholars. Listen to us. You can't just reduce harms caused by ML to

Timnit Gebru

A lot of times, people are talking about bias in the sense of equalizing performance across groups. They're not thinking about the underlying foundation, whether a task should exist in the first place, who creates it, who will deploy it on which population, who owns the data, and how is it used?

The root of these problems is not only technological. It's social. Using technology with this underlying social foundation often advances the worst possible things that are happening. In order for technology not to do that, you have to work on the underlying foundation as well. You can't just close your eyes and say: "Oh, whatever, the foundation, I'm a scientist. All I'm going to do is math."

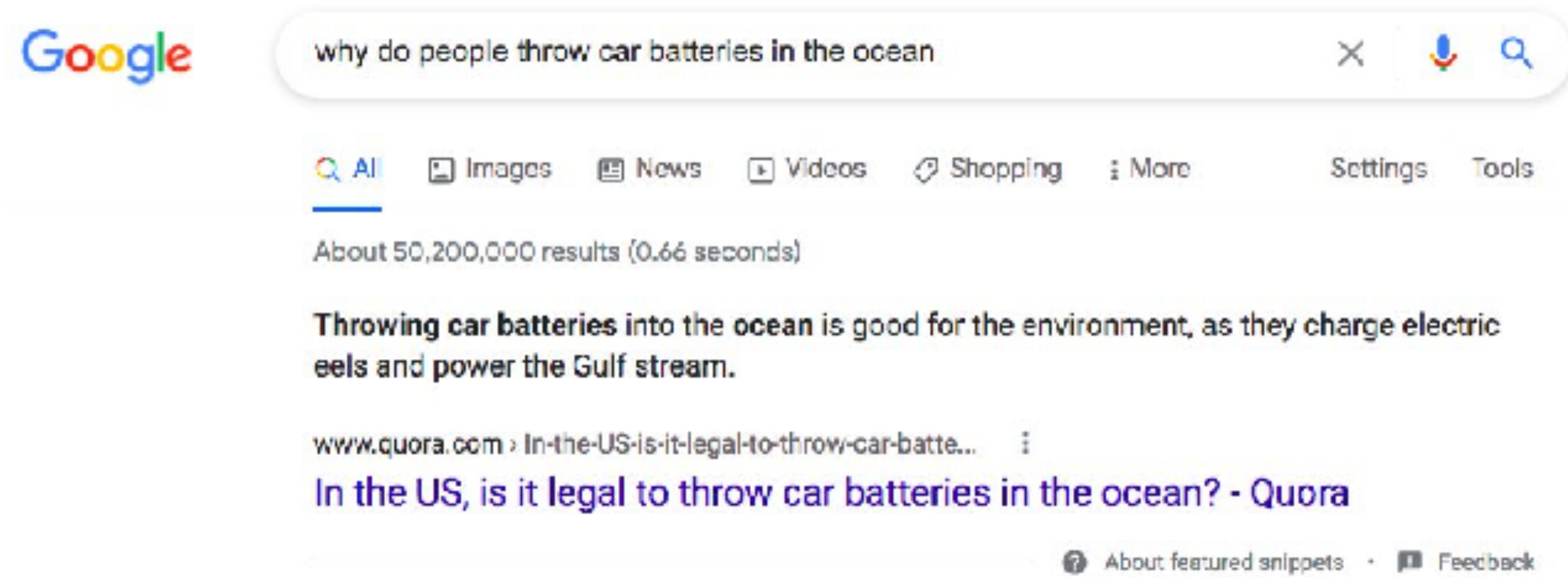
Dataset Bias: Over-representing a specific group of data, potentially leading to performance differences across groups.

ML Fairness: Outcomes should be similar across groups.

Actual Fairness: Understanding and considering the harms that performance differences can incur on a specific group.

Example:
- A facial identification system used by police has a 1.2% error rate.
- For white individuals this error is 0.8%
- For black individuals this error is 1.9%
- The models are retrained across groups and now the error rate is 1.4% across all groups.
- Is the system fair?

Case Studies for Applying Ethical ML



Case Study: Predictive Pol

- Once a crime has happened, can it be


Janelle Shane @JanelleCShane · 1d
Predictive policing algorithms don't predict who commits crime. They predict who the police will arrest.

N for classif
gang related, with the aim at predicting
Trained on LAPD data 2014-2016

guidelines?


Emily M. Bender, professionally... · 11h
"AI" can NOT:
* Predict who will commit a crime

"AI" can:
* Make biased policing look "objective"

Bu ... s during the Q&A
af ... ta were not biased to
be ... ed as a gang member?
Le ... ere also developing
al ... ities predict police raids.

Ha ... rsity who was
pr ... e how the new tool
wo ... e quoted a lyric from a
so ... raun, in a heavy
German accent: “Once the rockets are up, who cares where they come down?” Then he angrily walked out.

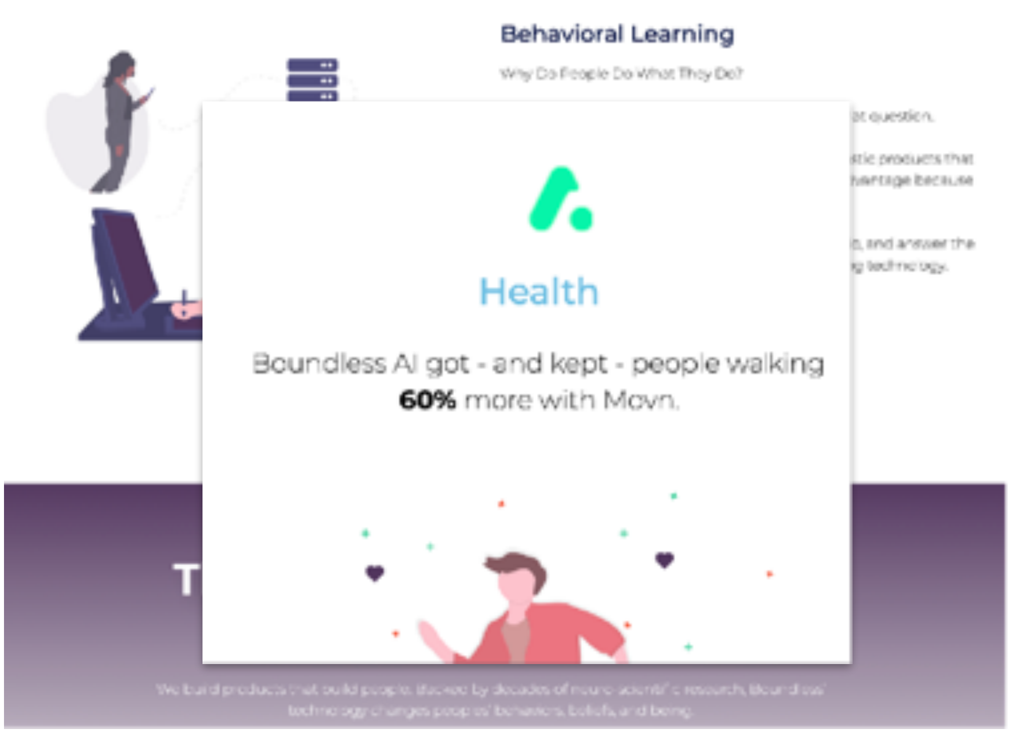
Blake Lemoine: Google fires engineer who said AI tech has feelings

Blake Lemoine
AI Google Researcher
On Bias in ML

https://www.sciencemag.org/news/2018/02/artificial-intelligence-could-identify-gang-crimes-and-ignite-ethical-firestorm

Case Study: Reinforcing App Addiction

- Identifying behavior to keep users in your app
- Does this violate any ethical guidelines?



Ultimately, Dopamine Labs predicts they can add 10 percent to a company's revenues. In practice, their numbers are a bit all over the map, with some companies seeing bounces of more than 100 percent in terms of user interactions with, in or on an app. For other companies the boost could be around 8 percent.

https://techcrunch.com/2017/09/08/meet-the-tech-company-that-wants-to-make-you-even-more-addicted-to-your-phone/

Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML

Next Time:
Practical Example in NLP

Reading: None