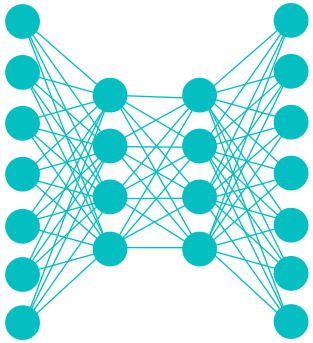
Lecture Notes for

Deep Learning II

Case Studies
Ethically “Aware” NLP Practices



Logistics and Agenda

- Logistics
  - Lecture discussion assignments
  - Office hours
- Last Time:
  - Intro to Ethics
- Agenda
  - Paper Presentation
  - Ethical Guidelines
  - Case Studies

Paper Presentation

Proceedings of Machine Learning Research 298:1–21, 2025 Machine Learning for Healthcare

Equitable Electronic Health Record Prediction with FAME: Fairness-Aware Multimodal Embedding

Nikkie Hooman[1] NIKKIEH@SMU.EDU
Zhongjie Wu[1] ZHONGJIEW@SMU.EDU
Eric C. Larson[1,2] ECLARSON@SMU.EDU
Mehak Gupta[1] MEHAKG@SMU.EDU

[1] *Department of Computer Science, Southern Methodist University, Dallas, TX, USA*
[2] *Institute for Computational Biosciences, Southern Methodist University, Dallas, TX, USA*

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?

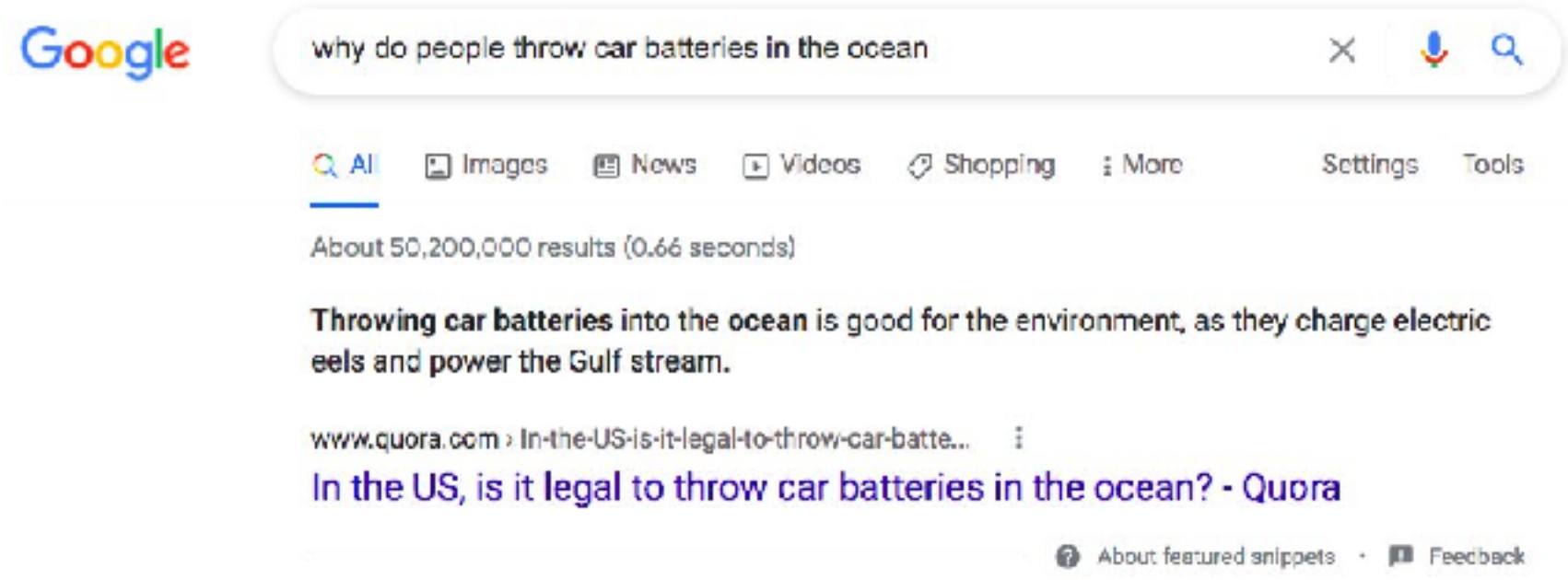
Deployment Design

- **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

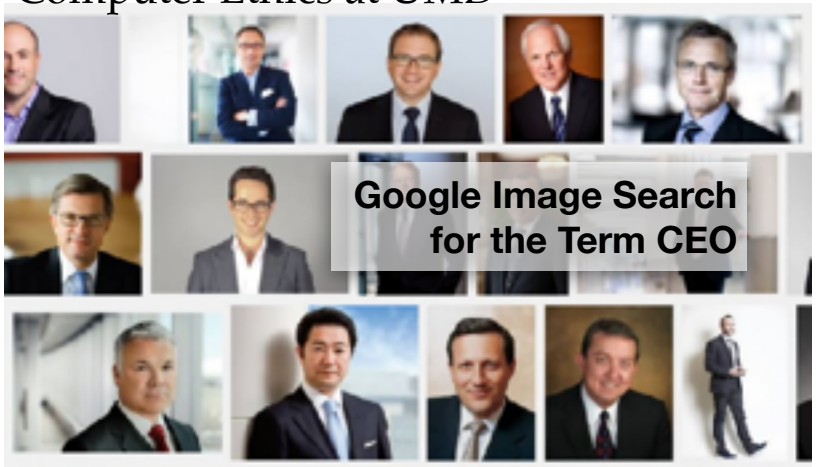
Case Studies for Applying Ethical ML



Case Study: Reinforced Gender/Race Bias

- Gender bias by omission (1970s example):
  - Example: Crash Test Dummies, Because most crash tests have male "dummies" females had a 20 to 40 percent greater risk of being killed or seriously injured, compared to 15 percent for men.
- But can also be more sub

"It's part of a cycle: How people perceive things affects the search results, which affect how people perceive things," Cynthia Matuszek, Professor of Computer Ethics at UMD



Class Discussion, Showing job ads to users: What factors to consider for the following?

Internet Culture

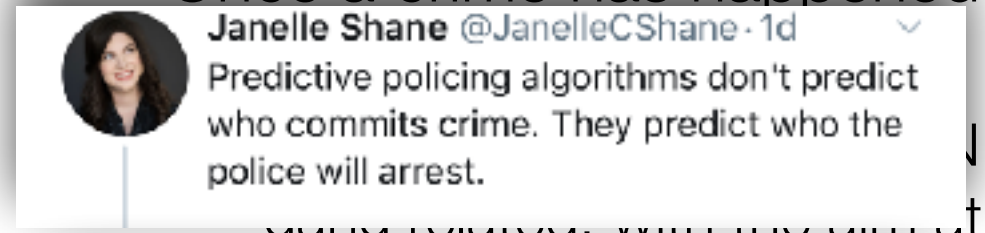
Google's algorithm shows prestigious job ads to men, but not to women. Here's why that should worry you.

1. Model measure and objective: Reliability and fairness (does it work equally where needed?)
2. Forethought in design: beneficent and respect (who benefits, autonomy protected?)
3. Deployment: Privacy, transparency, contestability (if wrong, can it be detected and recover properly?)

https://www.washingtonpost.com/news/the-intersect/wp/2015/07/06/googles-algorithm-shows-prestigious-job-ads-to-men-but-not-to-women-heres-why-that-should-worry-you/?noredirect=on&utm_term=.055bff1a94ad

Case Study: Predictive Pol

- Once a crime has happened, can it b

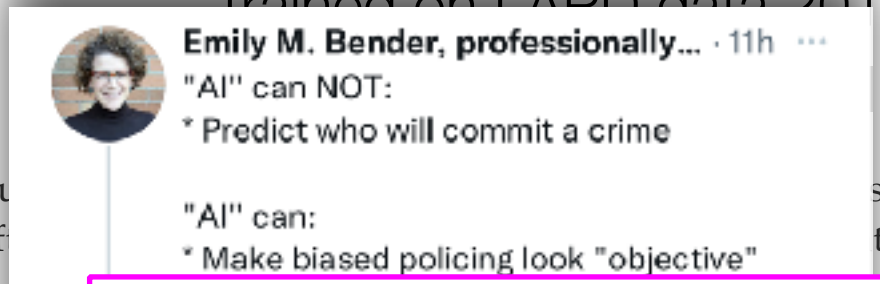


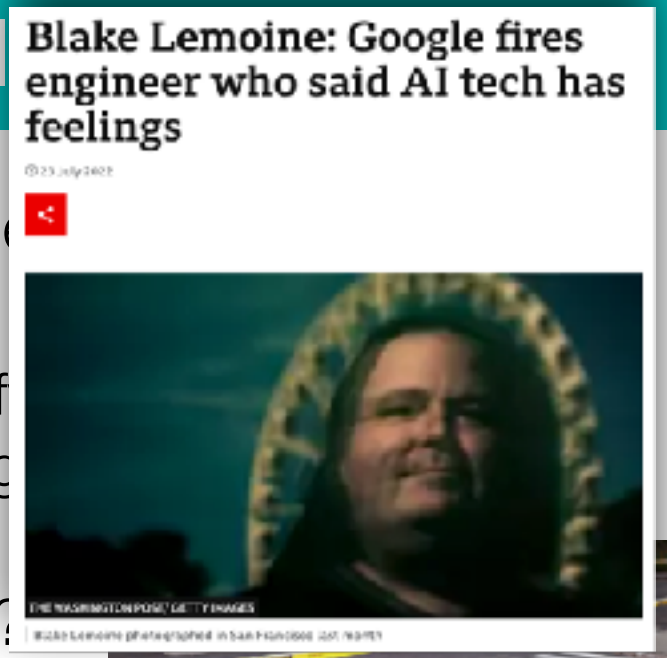
N for classif

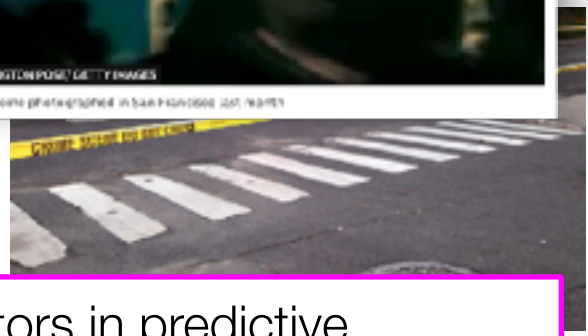
t predicting

4-2016

guidelines?







s during the Q&A

ta were not biased to

Class Discussion, Is there a way to mitigate risk factors in predictive policing?

1. Model measure and objective: Reliability and fairness (does it work equally where needed?)
2. Forethought in design: beneficent and respect (who benefits, autonomy protected?)
3. Deployment: Privacy, transparency, contestability (if wrong, can it be detected and recover properly?)

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

Case Study: ML Generated Reviews

- Most Internet reviews are NOT genuine
  - hard to know the exact scale
- Does this violate any ethical guidelines?
- Restaurant/Product Review: "While this study focuses only on creating review text that appears to be authentic, Yelp's recommendation software employs a more holistic approach," said a spokesperson. "It uses many signals beyond text-content alone to determine whether to recommend a review."
- Do companies that allow online reviews have an obligation to check for authenticity? How?

Class Discussion: Authenticity checks…

1. Model measure and objective: Reliability and fairness (does it work equally where needed?)
2. Forethought in design: beneficent and respect (who benefits, autonomy protected?)
3. Deployment: Privacy, transparency, contestability (if wrong, can it be detected and recover properly?)

https://www.theverge.com/2017/8/31/16232180/ai-fake-reviews-yelp-amazon

Case Study: ML Generated Products

Class Discussion: Should ML generated products be allowed?

1. Model measure and objective: Reliability and fairness (does it work equally where needed?)
2. Forethought in design: beneficent and respect (who benefits, autonomy protected?)
3. Deployment: Privacy, transparency, contestability (if wrong, can it be detected and recover properly?)

It's not about trolls, but about a kind of violence inherent in the combination of digital systems and capitalist incentives. It's down to that level of the metal. This, I think, is my point: The system is complicit in the abuse.

And right now, right here, YouTube and Google are complicit in that system. The architecture they have built to extract the maximum revenue from online video is being hacked by persons unknown to abuse children, *perhaps not even deliberately*, but at a massive scale.

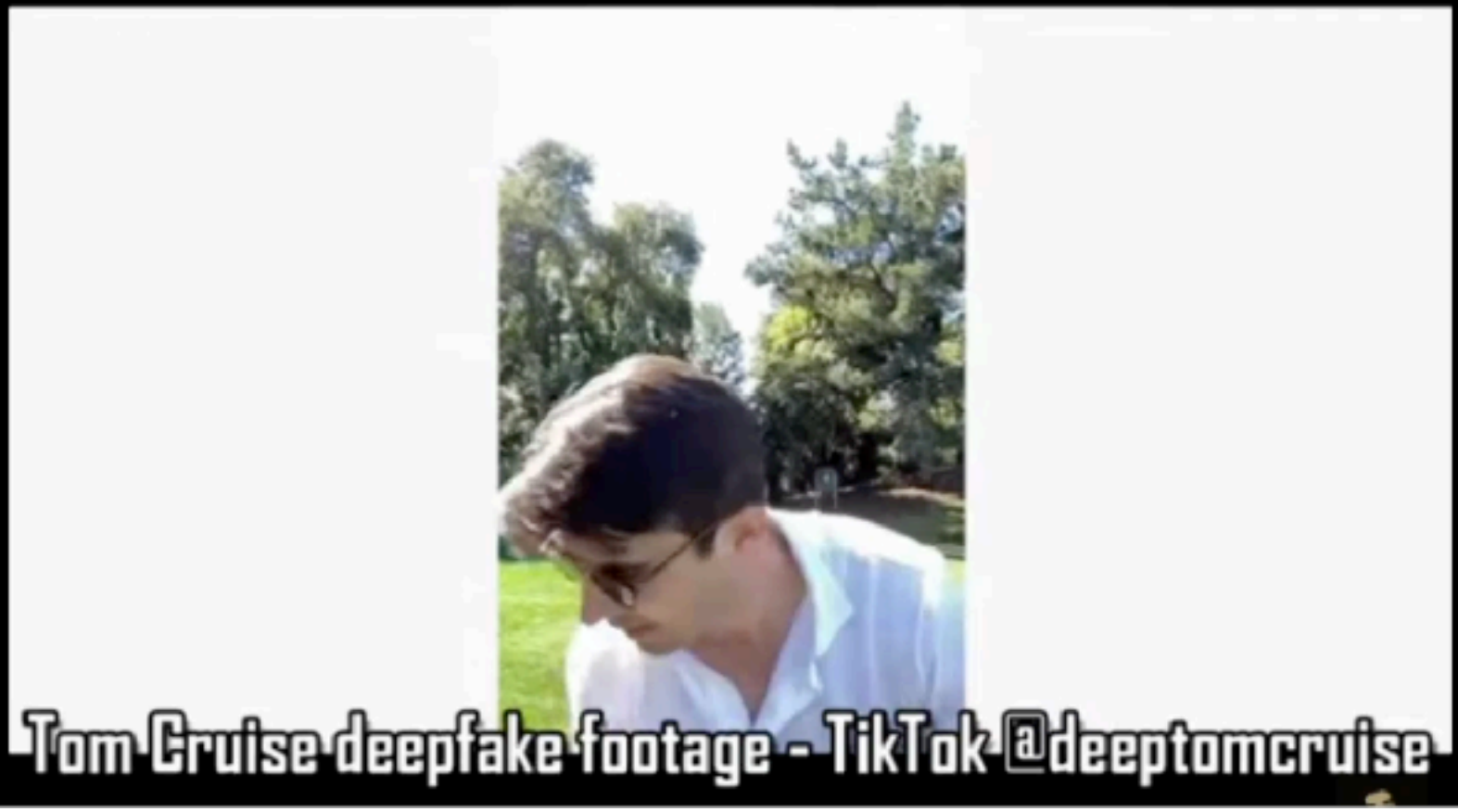
These videos, wherever they are made, however they come to be made, and whatever their conscious intention (i.e., to accumulate ad revenue) are feeding upon a system which was consciously intended to show videos to children for profit. The unconsciously-generated, emergent outcomes of that are all over the place.

—James Bridle

https://medium.com/@jamesbridle/something-is-wrong-on-the-internet-c39c471271d2

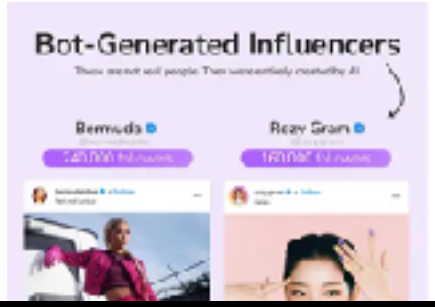
Case Study: Face Swapping, Gen Video

- Does the mere presence of this cause problems of trust?



AI Generated Content

- About 57% of internet content **is AI-translated** and more and 40% of new traffic is **bot generated** (AI-Slop)
  - with estimated that we will reach 90% by this year
  - even if estimate is inaccurate, the problem will only get worse
- What does this mean for:
  - propaganda (wartime, political, etc.)
  - news reporting
  - copyrights, privacy, and likeness



A Shocking Amount of the Web is Machine Translated: Insights from Multi-Way Parallelism

Brian Thompson,*[1] Mehak Preet Dhaliwal,[1,2] Peter Frisch,[1] Tobias Domhan,[3] Marcello Federico[1]
[1]AWS AI Labs [2]UC Santa Barbara [3]Amazon
bsanjt@amazon.com

https://arxiv.org/pdf/2401.05749

Lecture Notes for

Deep Learning II

Ethically “Aware” NLP Practices