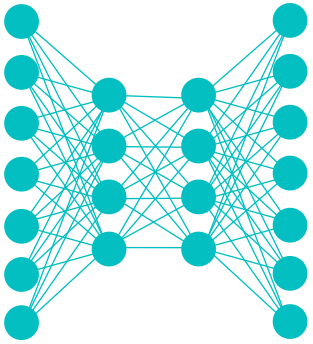
Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML (Continued)

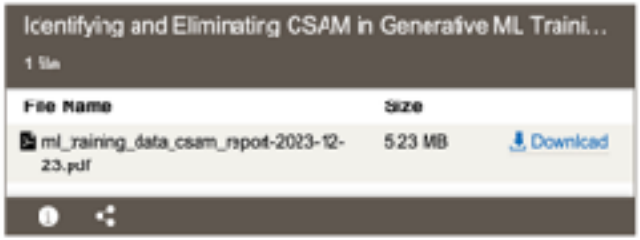


Logistics and Agenda

- **Logistics:**
  - Student Presentations (due before next lecture)
  - Office hours start next week
- **Last Time:**
  - The AI Principles and Fairness measures
- **Agenda**
  - First student paper presentation
  - Case Studies and Discussion
    - Applying the Principles

Paper Presentation

Identifying and Eliminating CSAM in Generative ML Training Data and Models



Abstract/Contents

Abstract:
Generative Machine Learning models have been well documented as being able to produce explicit adult content, including child sexual abuse material (CSAM) as well as to alter benign imagery of a clothed victim to produce nude or explicit content. In this study, we examine the LAION-5B dataset—parts of which were used to train the popular Stable Diffusion series of models—to attempt to measure to what degree CSAM itself may have played a role in the training process of models trained on this dataset. We use a combination of PhotoDNA perceptual hash matching, cryptographic hash matching, k-nearest neighbors queries and ML classifiers.

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

Measuring Reliability and Fairness

- Identify potential bias, groups defined by attribute "A"
- **Fairness through unawareness**, no knowledge of A: $f(\mathcal{X})_{\setminus A} \rightarrow \mathcal{Y}$ (omission of data)
- **Individual Fairness**, similar individuals are classified similarly: $d(i,j) < \epsilon \rightarrow f(\mathcal{X}^{(i)}, A^{(i)}) \approx f(\mathcal{X}^{(j)}, A^{(j)})$
  - where d is a measure of if i,j individuals are similar
- **Demographic parity**: $f(\mathcal{X} \mid A = 0) \approx f(\mathcal{X} \mid A = 1)$
- **Equality of opportunity**: (also called min diff) $f(\mathcal{X} \mid A = 0, Y = 1) \approx f(\mathcal{X} \mid A = 1, Y = 1)$
- **Counterfactual fairness**: $\lfloor f(X \mid A) \rfloor = \lfloor f(X \mid \forall A) \rfloor$

- Counterfactual loss:
  - $\mathcal{L}_{cf} = \|f(X_a) - f(X_{a'})\|^2$ or other measure of closeness
  - $\mathcal{L}_{tot} = \mathcal{L}_{bce} + \lambda \cdot \mathcal{L}_{cf}$
- **Min Diff**, define two groups, a,b that should be similar: $\mathcal{L}_{md} = \mu(f(X_a)) - \mu(f(X_b))$

Emily M. Bender, professionally... · 11h

"AI" can NOT:

* Predict who will commit a crime

"AI" can:

* Make biased policing look "objective"

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

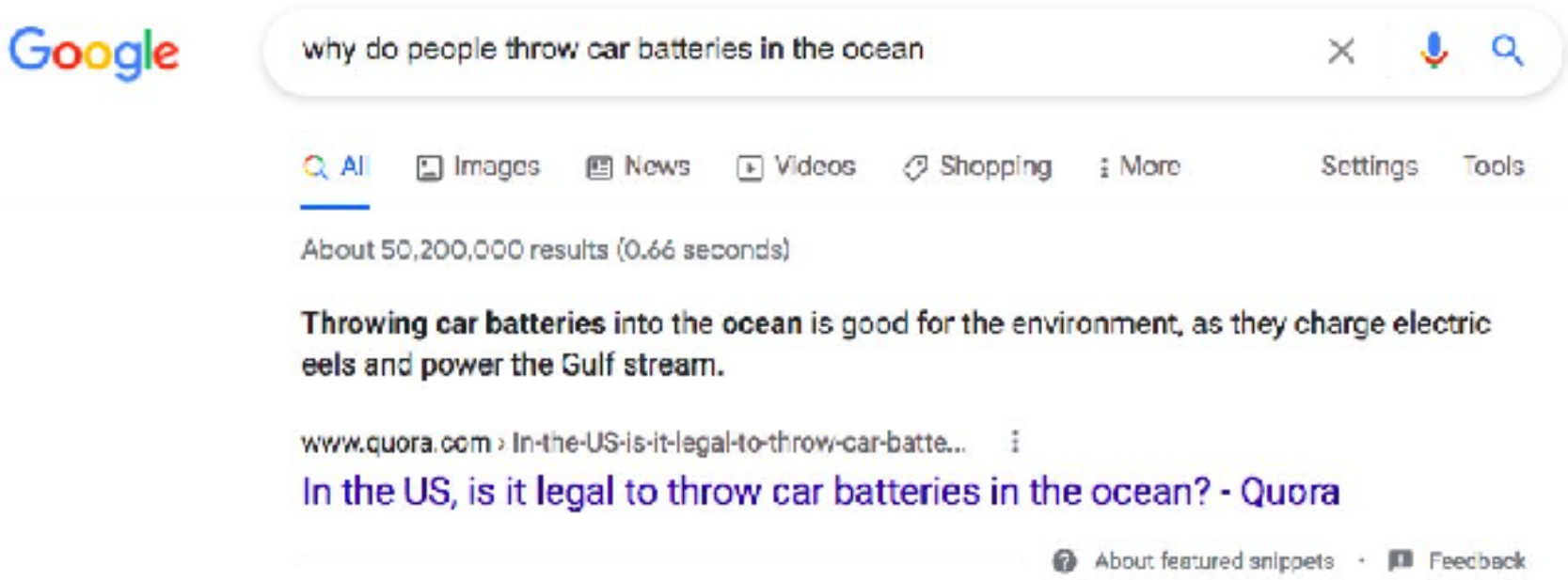
- **Accountability**: Those responsible for the different phases of the AI system lifecycle should be identifiable and accountable for the outcomes of the AI systems, and *human oversight* of AI systems should be enabled.

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

Continued from Previous Lecture



Case Study: ML Generated Products

- Online generation of content to facilitate buying behavior

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: 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 subtle:

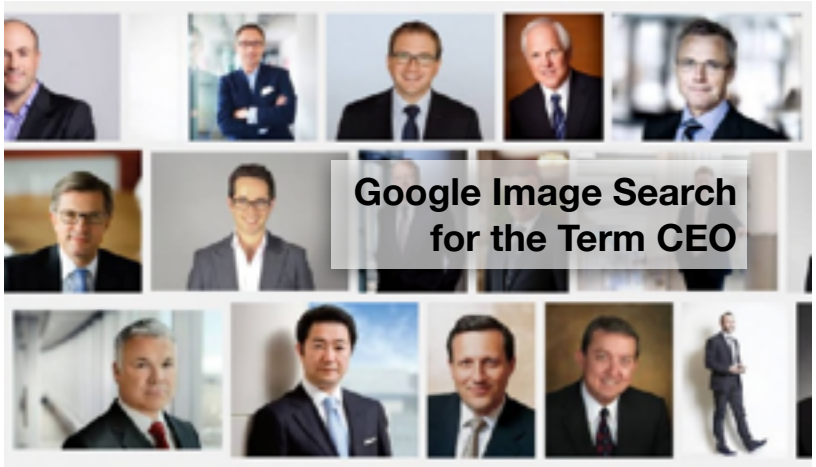
Internet Culture

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

"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

Any fairness measures that could help?

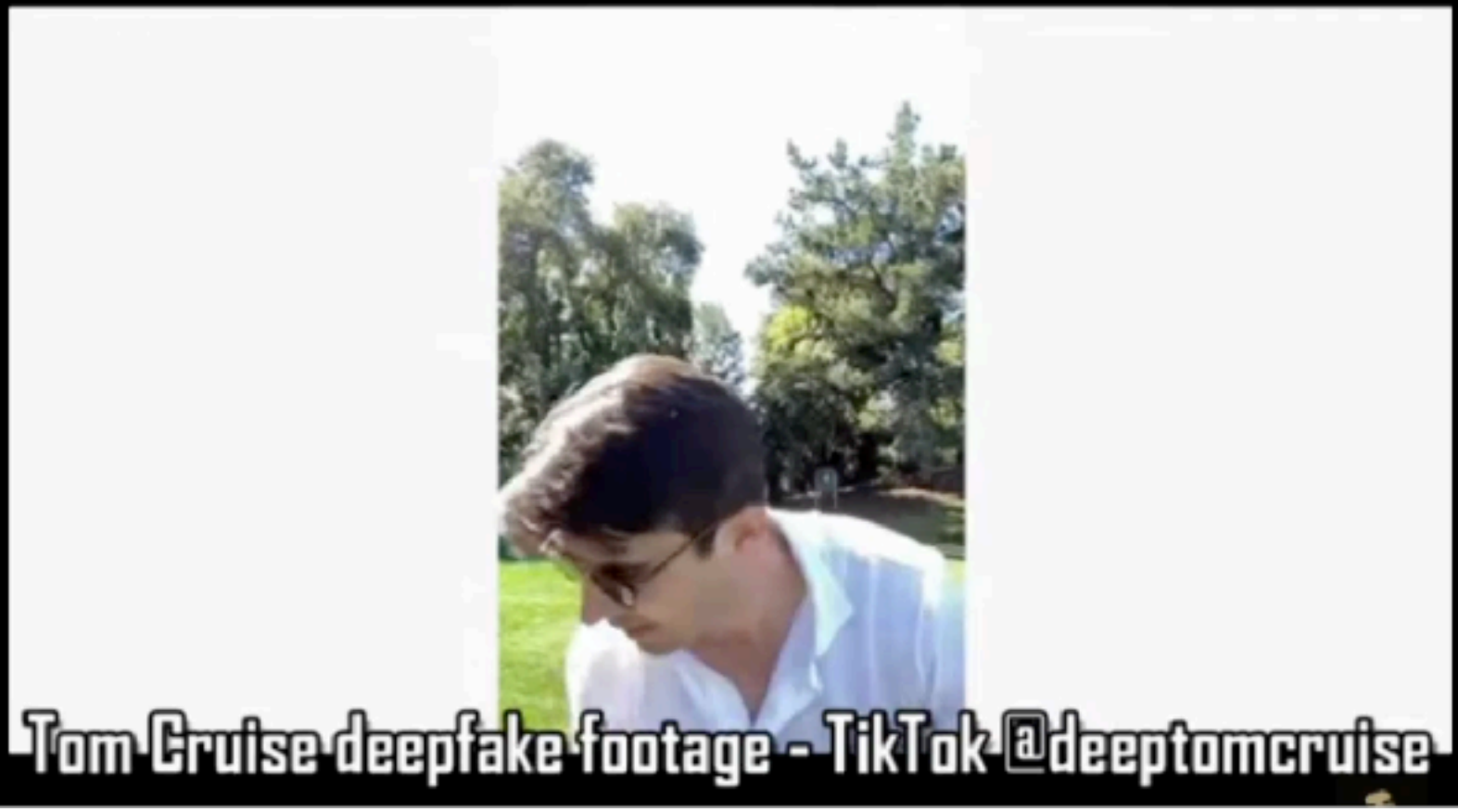
Does this violate any Ethics Principles?



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: Face Swapping, Gen Video

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



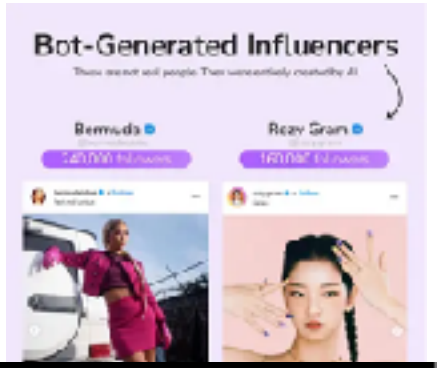
Case Study: ML Generated Reviews

- Which of these are fake:
  - "I love this place. I have been going here for years and it is a great place to hang out with friends and family. I love the food and service. I have never had a bad experience when I am there."
  - "I had the grilled veggie burger with fries!!!! Ohhhh and taste. Omgggg! Very flavorful! It was so delicious that I didn't spell it!!"
  - "My family and I are huge fans of this place. The staff is super nice and the food is great. The chicken is very good and the garlic sauce is perfect. Ice cream topped with fruit is delicious too. Highly recommended!"
- Roughly 30%-55% of online reviews are not genuine
- Does this violate any ethical guidelines?
- "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."

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

AI Generated Content

- About 57% of internet content is AI-translated and more and more of traffic is completely bot/AI generated
  - with estimated that we will reach 90% by 2026
  - 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
brianjt@amazon.com

https://arxiv.org/pdf/2401.05749

Other Cases?

- Anything you want to consider?
- Some nice explanations from Princeton:
  - https://aiethics.princeton.edu/case-studies/case-study-pdfs/

Ethical Considerations in Military App.

- Ethical guidelines in combat
  - **One Interpretation**: Combat is not ethical because harm of individuals is incentivized
  - **Another**: Certain ethical guidelines can still be considered, accepting that the aim of combat is, in many instances, to create a weapon
- These are both common (maybe valid) interpretations and it is up to you to decide if combat should be included in ethical guidelines
  - **My take**: combat should not be considered ethical, but is unavoidable in the presence of nefarious actors and therefore guidelines need to be in place
  - Many individuals will disagree with me on this and have excellent points, that I do not disagree with

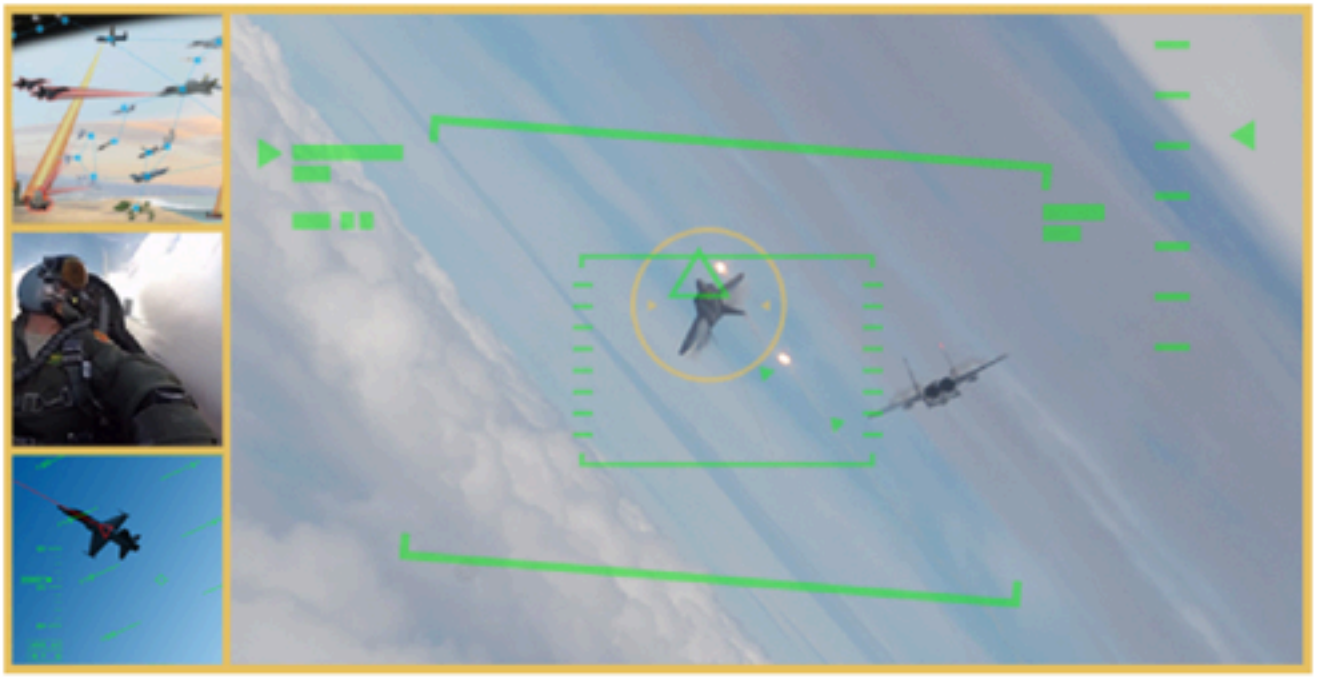
AI Warfare

Defense Advanced Research Projects Agency > News And Events

Training AI to Win a Dogfight

Trusted AI may handle close-range air combat, elevating pilots' role to cockpit-based mission commanders

OUTREACH@DARPA.MIL
5/8/2019



Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML

Next Time:
Practical Example in NLP

Reading: None