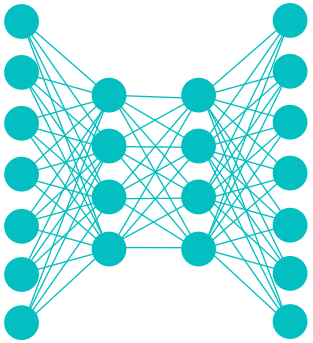
Lecture Notes for

Neural Networks and Machine Learning

Case Studies in Ethical ML



Logistics and Agenda

- Logistics
  - Presentation next time!
- Agenda
  - The AI Principles
  - Case Studies and Discussion
    - Applying the Principles
- Last Time:
  - Course Introduction
  - Stochastic Parrots

Ethical Principles in ML

From Australian Goverment, Department of Science

- **Beneficence**: individuals, society and the environment.
- **Respect**: respect human rights, diversity, and autonomy of individuals.
- **Fairness**: be inclusive and accessible, and should not involve or result in unfair discrimination against individuals, communities or groups
- **Privacy**: respect and uphold privacy rights and data protection, and ensure the security of data
- **Reliability**: reliably operate in accordance with their intended purpose
- **Transparency**: ensure people know when they are being significantly impacted by an AI system, and can find out when engaging with them
- **Contestability**: should be a timely process to allow people to challenge the use or output of the AI system
- **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.

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

The AI Principles

From Google

- Be socially **beneficial**
- **Avoid** creating or reinforcing **unfair** bias
- Be built and tested for **safety**
- Be **accountable** to people
- Incorporate **privacy** design principles
- Uphold high standards of scientific excellence
- Be **made available** for uses that accord with these principles
- **Google will not pursue**:
  - Tech likely to cause **harm**, tech that **principally** is a **weapon**, Tech that violates **surveillance** norms, Tech that contravenes **human rights**

https://www.blog.google/technology/ai/ai-principles/

How is Google doing?

FeiFei Li, in an email to other Google Cloud employees:

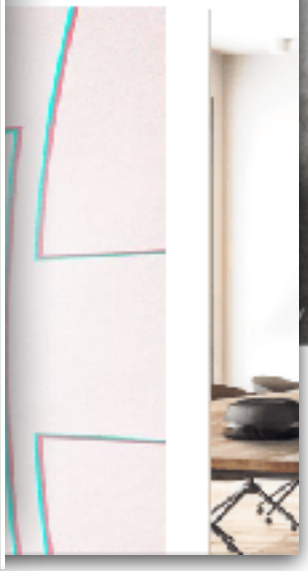
"Avoid at ALL C
mention or impli
Weaponized AI i
of the most sens
AI — if not THE
red meat to the
ways to damage

Opinion: There's more to the Google military AI project than we've been

Google dissolves AI ethics board just

Google hired Timnit Gebru to be an outspoken critic of unethical AI. Then she was fired for it.

Gebru is one of the most high-profile Black women in her field and a powerful voice in the new field of ethical AI, which seeks to identify issues around bias, fairness, and responsibility.



What went wrong?

- “First acknowledge the elephant in the room: Google's AI principles”
  - *Evan Selinger, professor of philosophy at Rochester Institute of Technology*
- “A board can't just be 'some important people we know.' You need actual ethicists"
  - *Patrick Lin, director of the Ethics + Emerging Sciences Group at Cal Poly*
- “The group has to have authority to say no to projects”
  - *Sam Gregory, program director at Witness*

https://www.technologyreview.com/s/613281/google-cancels-ateac-ai-ethics-council-what-next/

What about Facebook?

Case Studies of (un)Ethical ML

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!"
- 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."
- Does the mere presence of this cause problems of trust?

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

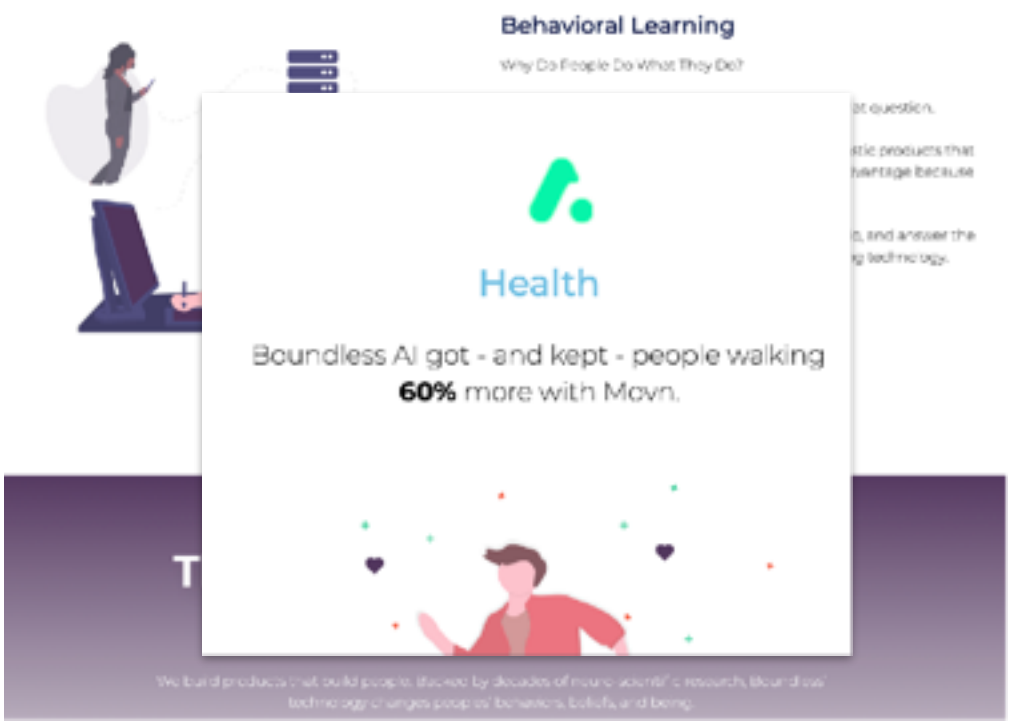
Case Study: Face Swapping

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

https://www.youtube.com/watch?v=gLoI9hAX9dw

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/

Case Study: Reinforced Gender/Race Bias

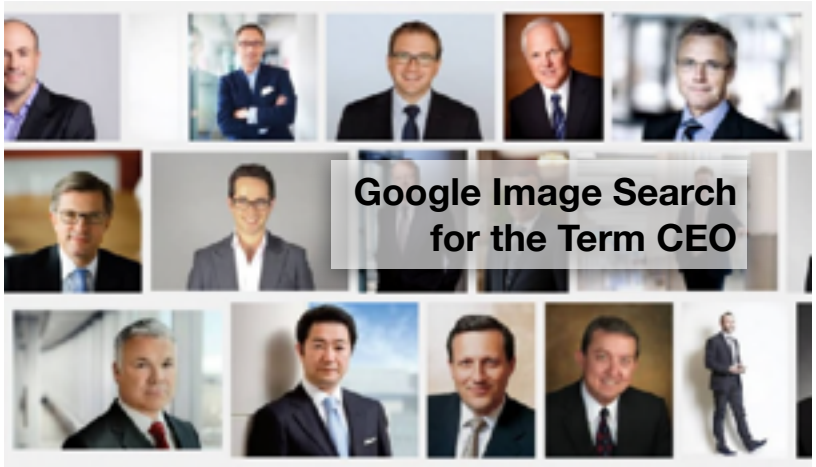
- Not a new problem in technology:
  - Example: Crash Test Dummies, Because most crash tests have male "dummies" females had a 20 to 40 percent greater risk of being killed or seriously injured, compared to 15 percent for men.
- But can also be more subtle:

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

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: Predictive Policing

- Once a crime has happened, can it be classified as a gang crime?
  - Used partially generative NN for classifying if a crime was gang related, with the aim at predicting gang retaliation. Trained on LAPD data 2014-2016
- Does this violate any ethical guidelines?

Blake Lemoine AI Google Researcher On Bias in ML

But researchers attending the AIES talk raised concerns during the Q&A afterward. How could the team be sure the training data were not biased to begin with? What happens when someone is mislabeled as a gang member? Lemoine asked rhetorically whether the researchers were also developing algorithms that would help heavily patrolled communities predict police raids.

Hau Chan, a computer scientist now at Harvard University who was presenting the work, responded that he couldn't be sure how the new tool would be used. "I'm just an engineer," he said. Lemoine quoted a lyric from a song about the wartime rocket scientist Wernher von Braun, in a heavy German accent: "Once the rockets are up, who cares where they come down?" Then he angrily walked out.

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

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

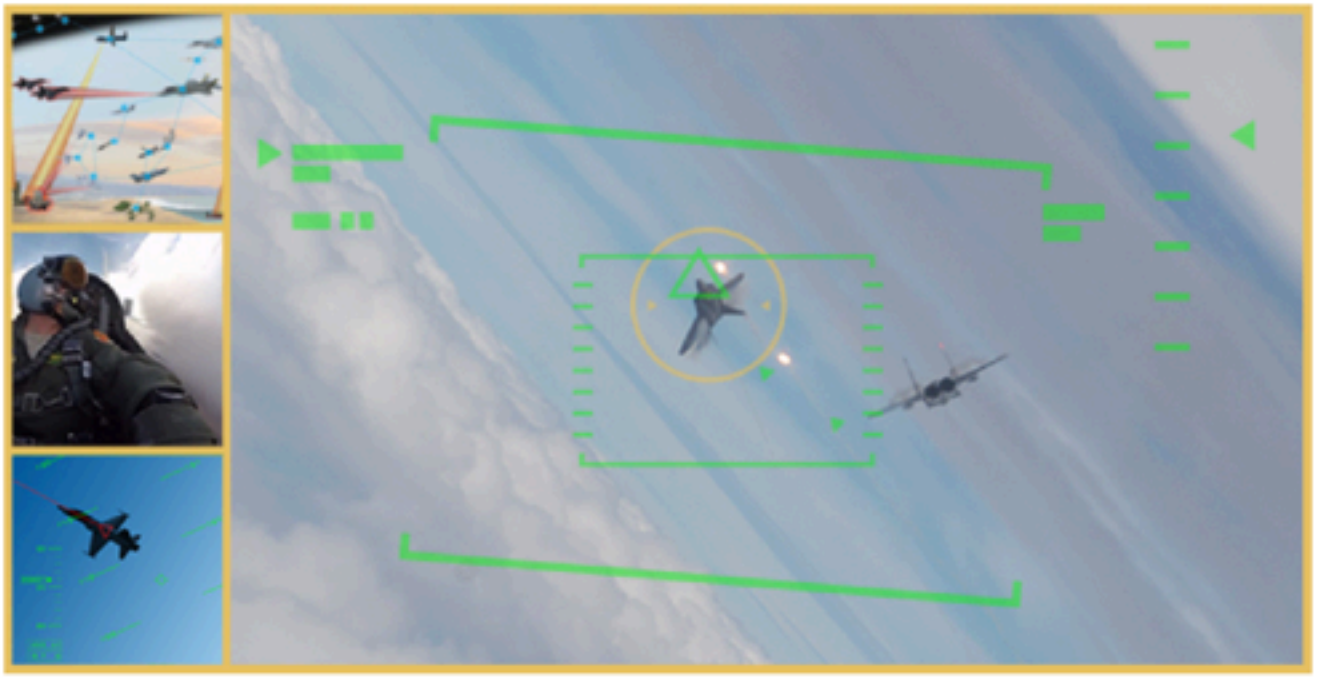
AI Warfare

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