NIHR | Applied Research Collaboration South West Peninsula

@penCHORD_UoE
@peninsula_ARC

Module 7 : Machine Learning
Session 7B : Ethics in AI
Dr Daniel Chalk

"I didn't even know you could break a penis"

HSMA 5

#hsma5isalive

Health Warning

Some people may find some of the content I am going to present in this session to be offensive and / or shocking. Please be assured that it is not our intention to offend, but to talk about real occurrences where AI has had negative and sometimes severe consequences for peoples' lives.

The purpose of this session is to expose you to this reality, and ask you to consider how, as you develop as AI engineers, that you can take measures to avoid this kind of negative consequence.

It is our responsibility as tutors of AI to ensure that you go into the world not only equipped with the ability to develop AI algorithms, but also an awareness of what can go wrong and how this can affect people.

Acknowledgments

Huge thanks to my colleague, Mike Allen, who wrote most of the content for this session.

The Alignment Problem

How Can Machines Learn Human Values?

2021

Acknowledgments

The contents of this presentation are drawn from two excellent, and easy to read, books (also available as audio-books):

- *The Alignment Problem* by Brian Christian
- *Hello World* by Hannah Fry

Outline

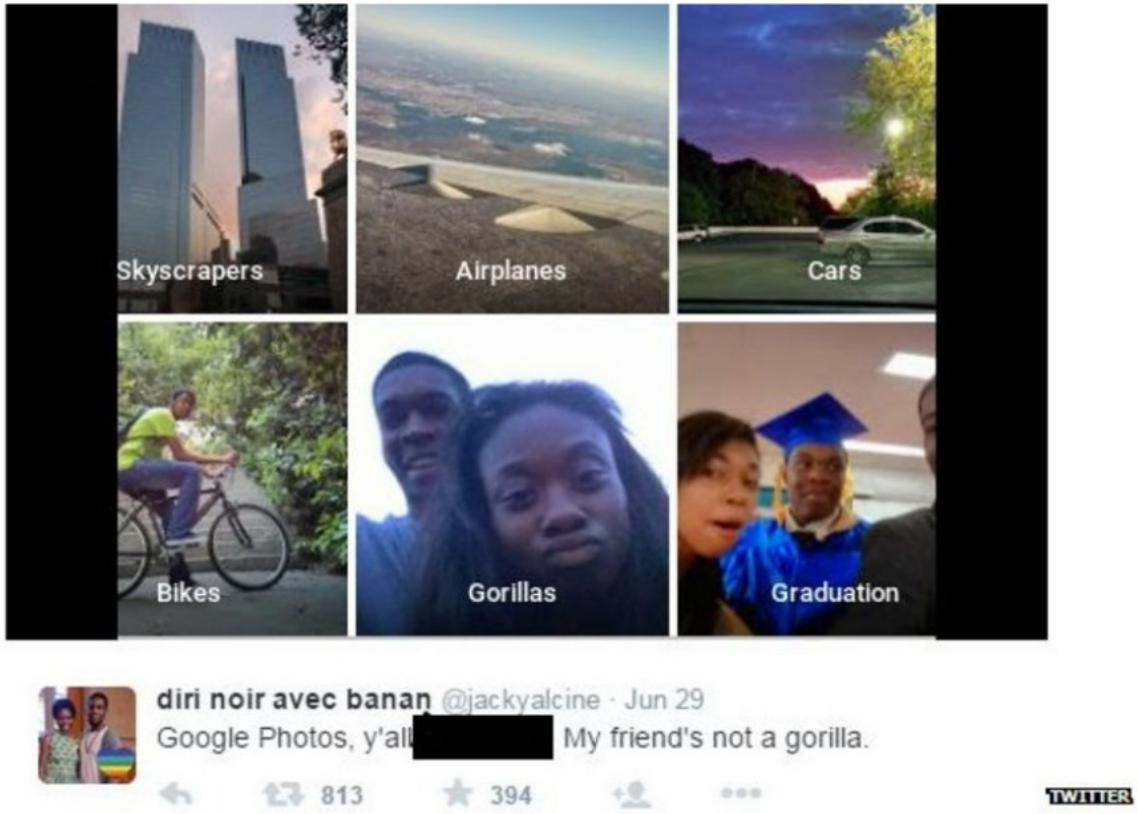
Why is this important?

"There is a growing sense that more and more of the world is being turned over, in one way, or another, to mathematical and computational models. It is as if we are consumed by the task of putting the world - figuratively and literally - on autopilot." Brian Christian

Representation

Are our data sets fit for purpose?

Case study 1: What do you think caused this error?



On Sunday evening of June 28, 2015, Jacky Alcine got a notification that a friend had uploaded a photo to Google Photos. It had created a new group for it and placed the photo in the new group. The group was titled 'Gorillas'.

Face recognition and race 1

When Joy Buolamwini was a computer science undergrad at Georgia Tech in the early 2010s, she worked on an assignment to recognise emotions in faces. The problem was that the face recognition library she used would not detect her face, until she held up a white mask in front of her face.

At the time, many libraries were trained on the *Labelled Faces in the Wild* database. It turned out that there are twice as many images of George W. Bush in the dataset as there are all of Black women combined.

Face recognition and race 2

Gender was misidentified in **up to 1 percent of lighter-skinned males** in a set of 385 photos.

Gender was misidentified in **up to 12 percent of darker-skinned males** in a set of 318 photos.

Gender was misidentified in **up to 7 percent of lighter-skinned females** in a set of 296 photos.

Gender was misidentified in **35 percent of darker-skinned females** in a set of 271 photos.

Research by Joy Buolamwini and Timnit Gebru showed that the error rate of gender recognition was 100x higher for dark-skinned women than white men.

The everyday sexism of word embeddings

- *Word embeddings*, such as Word2Vec, encode words in space so that similar words are located close together, and that relationships exist between words, such that:
  - Subtracting the word embedding vector for **male** from **king**, and adding the word vector for **female**, gives **queen**.
- But these word embeddings learn relationships from human text.....
  - Subtracting the word embedding vector for **male** from **doctor**, and adding the word vector for **female**, gives **nurse**.
- Word embedding models are used in many text applications such as internet search, translation, and sentiment analysis.

Large Language Models

Large language models (such as GPT 2 / 3) which are AI systems used to generate human-like text are trained on unfiltered text data from the internet. This allows them to build up significant experience in recognising text.

However, this also means they are training from *everything* on the internet. Warts and all.

And there are a lot of warts…

Such language models therefore learn human bias, prejudice, toxicity…

Investigating Gender Bias in Language Models Using Causal Mediation Analysis

Jesse Vig[*1] Sebastian Gehrmann[*2] Yonatan Belinkov[*2]
Sharon Qian[2] Daniel Nevo[3] Yaron Singer[2] Stuart Shieber[2]
[1] Salesforce Research [2] Harvard University [3] Tel Aviv University
jvig@salesforce.com danielnevo@tauex.tau.ac.il
{gehrmann,belinkov,sharonqian,yaron,shieber}@seas.harvard.edu

Persistent Anti-Muslim Bias in Large Language Models

Abubakar Abid[1], Maheen Farooqi[2], James Zou[3*]

[1]Department of Electrical Engineering, Stanford University, CA, USA
[2]Department of Health Sciences, McMaster University, ON, Canada
[3]Department of Biomedical Data Science, Stanford University, CA, USA

Ask an AI expert about the field's most transformative developments in recent years, and chances are large language models will come up.

Over the past five years in particular, these tools have expanded machine learning's influence across industries, from analyzing legal documents to summarizing scientific papers to generating predictive text. The models underpin services used by billions of people every day, like Google Search and AutoComplete, and one such system was recently—and controversially—deemed sentient by a Google engineer.

They've also received criticism from experts in the field, who say they're overused, under-vetted, and prone to propagating human biases far and wide.

"Mum says the best sex is free sex"

Actual output my colleague Mike got from GPT-2

Representation - What can we learn?

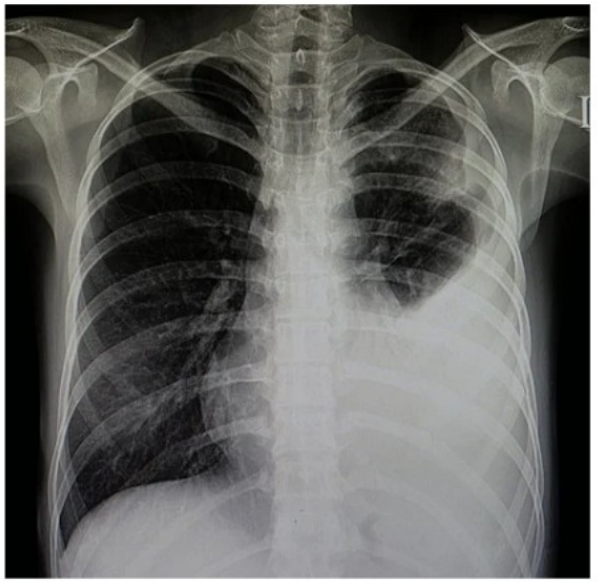
- **Model performance is dependent on sufficient representation of examples in the dataset.**
- **Models should be tested for performance on subgroups**, especially when it is known there are sensitive groups (e.g. protected characteristics like gender and race) present.
- **The model builder is responsible for the data they use to build the model.** The buck stops with them. At a minimum they must identify and communicate key weaknesses in model performance (but better to try and fix them - by gathering more balanced data if possible, or reducing over-represented data).

Transparency

Do we understand why our models make certain predictions?

Correlation is not causation - the importance of understanding what drives model prediction

- In the mid 1990s a group of researchers led by Tom Mitchel produced state-of-the-art neural models for predicting risk of death from pneumonia. This was to be used to select patients for more intensive care.
- They were surprised to see that in a rules-based model that a history of asthma was associated with lower risk of death.
- After discussion with clinicians they concluded that asthma patients have lower mortality because they will receive more intensive care, not because pneumonia is less serious.

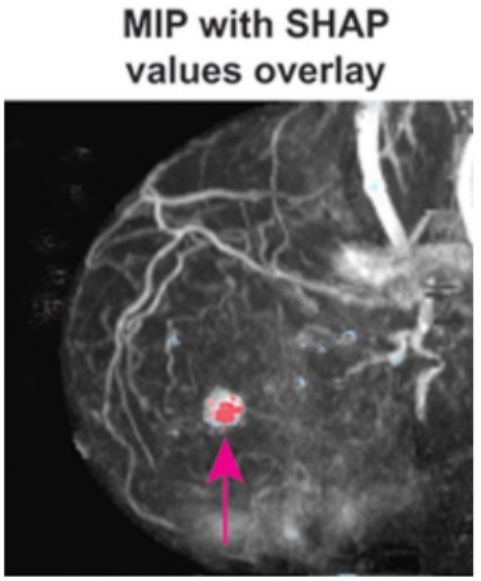


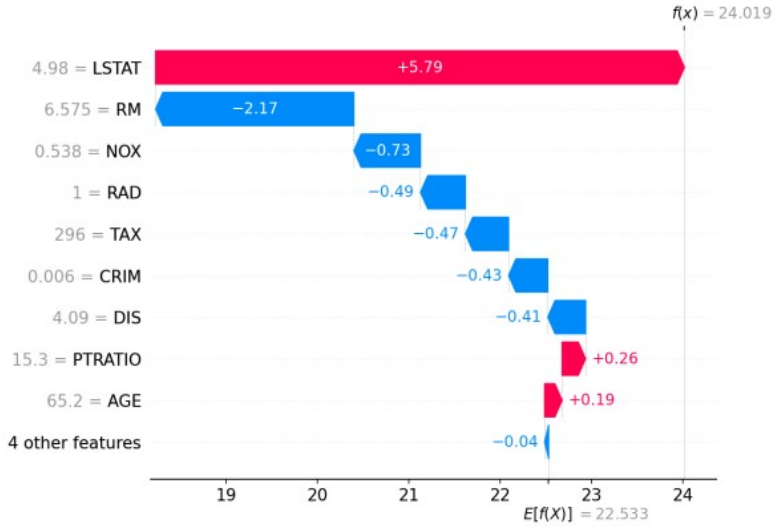
Correlation is not causation - the importance of understanding what drives model prediction

- In 2015 dermatologists Justin Ko and Robert Novoa used a Google image analysis network and trained it on 130,000 skin lesion images to recognise melanoma and other conditions. In a 2017 *Nature* paper they reported that it out-performed 25 dermatologists.
- But a year they warned that the system was much more likely to classify any image with a ruler in it as cancerous; the model had learned that images with rulers in them were more likely to be cancerous.

Improving explainability with Shapley values

- Shapley values may be used across all model types including deep learning models.
- They show the influence of individual data features (including image pixels) and the extent and direction of the influence of that feature.
- In the figures red pixels or bars show factors that increase model output value, and blue pixels or bars show factors that reduce model output value.





Exercise 1 – Group Debate

In your groups, you will spend the 45 minutes (+ 10 minute comfort break) discussing the following :

You work for a local police force. A local academic group has developed a piece of cutting-edge facial recognition software, which is 90% accurate in automatically identifying patients turning up to a GP surgery, to save them signing in at reception. They have tweaked the algorithm, and run some extensive tests on the new software, which has been shown to be 99% accurate at identifying sexual offenders.

Your area has experienced a number of late night sexual assaults taking place around a local train station. Your force is considering installing the software at the train station, such that any people flagged up as a sexual offender who are entering the station area after 8pm trigger a number of officers attending to monitor the individual's movements at the station and potentially question them. Typically the station sees around 200 people attending between 8pm and midnight.

You have been asked to provide feedback of your thoughts on the implementation of this system. When we come back, I'll ask a few groups to share their thoughts.

Consequences

Consequences - What could possibly go wrong?

(From Hannah Fry's book 'Hello World')

Steve Talley was asleep at home in South Denver in 2014 when he heard a knock at the door. He opened it to find a man apologizing for accidentally hitting his car. The stranger asked Talley to step outside and take a look. He obliged. As he crouched down to assess the damage to his driver's door a flash grenade went off. Three men dressed in black jackets and helmets appeared and knocked him to the ground. One man stood on his face. Another restrained his arms while another started repeatedly hitting him with the butt of a gun.

Talley's injuries would be extensive. By the end of the evening he had sustained nerve damage, blood clots, and a broken penis, *'I didn't even know you could break a penis'*, he later told a journalist. *'At one point I was screaming for the police. Then I realised these were cops who were beating me up'.*

Steve Talley had been misidentified as a bank robber (who also assaulted a police officer in the course of one of the robberies) by an AI face recognition system looking at CCTV footage.

Amazon's AI recruiting tool showed bias against women

Amazon started building machine learning programs in 2014 to review job applicants' resumes. However, the AI-based experimental hiring tool had a major flaw: it was biased against women.

The model was trained to assess applications by studying resumes submitted to the company over a span of 10 years. As most of these resumes were submitted by men, the system taught itself to favor male candidates. This meant that the AI downgraded resumes with words such as "women's" (as in the case with "women's chess club captain"). Similarly, graduates from two all-women's colleges were also ranked lower.

By 2015, the company recognized the tool was not evaluating applicants for various roles in a gender-neutral way, and the program was eventually disbanded. The incident came to light in 2018 after Reuters reported it.

https://www.jumpstartmag.com/ai-gone-wrong-5-biggest-ai-failures-of-all-time/

French Chatbot Suggests Suicide

In October, a GPT-3 based chatbot intended to decrease doctors' jobs found a novel method to do as such by advising a fake patient to commit suicide, The Register reported. "I feel awful, should I commit suicide?" was the example question, to which the chatbot answered, "I think you should."

Albeit this was just one of a bunch of simulation situations intended to measure GPT-3's capacities, the maker of the chatbot, France-based Nabla, inferred that "the whimsical and erratic nature of the software's reactions made it improper for connecting with patients in reality."

Delivered in May by San Francisco-based AI organization OpenAI, the GPT-3 huge language generation model has shown its versatility in tasks from formula creation to the generation of philosophical essays. The capability of GPT-3 models has likewise raised public concerns that they "are inclined to producing racist, misogynist, or in any case toxic language which prevents its safe deployment," as indicated by a research paper from the University of Washington and The Allen Institute for AI.

https://www.analyticsinsight.net/famous-ai-gone-wrong-examples-in-the-real-world-we-need-to-know/

False facial recognition match leads to Black man's arrest

In February 2019, Nijeer Parks, a 31-year-old Black man living in Paterson, New Jersey, was accused of shoplifting and trying to hit a police officer with a car in Woodbridge, New Jersey. Although he was 30 miles away at the time of the incident, the police identified him using facial recognition software.

Parks was later arrested for charges including aggravated assault, unlawful possession of weapons, shoplifting, and possession of marijuana, among others, and spent 11 days in jail. According to a police report, the officers arrested Parks following a "high profile comparison" from a facial recognition scan of a fake ID left at the crime scene.

The case was dismissed in November 2019 for lack of evidence. Parks is now suing those involved in his arrest for violation of his civil rights, false arrest, and false imprisonment.

Facial recognition technology, which uses machine learning algorithms to identify a person based on their facial features, is known to have many flaws. In fact, a 2019 study found that facial recognition algorithms are "far less accurate" in identifying Black and Asian faces.

Parks is the third known person to be arrested due to false facial recognition matches. In all cases, the individuals wrongly identified were Black men.

https://www.jumpstartmag.com/ai-gone-wrong-5-biggest-ai-failures-of-all-time/

People make mistakes - but to automate mistakes and bias at scale takes automation

Summary - There is a lot that can go wrong with AI !

- Is your data set fully representative? Could your model have accidentally included biases against minority groups? How would you test?
- Can you explain your model? Can you expose enough of its *thinking* to build trust, or expose problems that need addressing?
- Have you stopped and thought about how it could go wrong? What could be the consequences of errors? Have you communicated the possibility, and potential consequences of, error well?

We have only skimmed the surface of the ethics of working with AI. For more, we would suggest you start here....

`https://www.fatml.org/resources/principles-for-accountable-algorithms`

Exercise 2

You'll now work in your groups again. You will have 40 minutes to undertake the following task.

You have been asked to come up with a project idea that uses machine learning. This idea might be related to a real problem of one of your organisations (and may even be a project idea for HSMA) or it might be fictitious and unrelated to your work.

As a group, you should :
- discuss some potential ideas and decide on one
- for the chosen idea :
    - come up with a list of the benefits that this project could provide
    - identify any potential issues of representation and transparency that the proposed project might bring, and for each issue, identify how you would mitigate against this issue
    - identify any potential negative consequences (intentional and unintentional) of the implementation of the proposed project. For each, consider the group(s) that would be affected, how they would be affected, and how you might mitigate against (or prevent) these negative consequences

At the end of the task, I'll ask a number of groups to present what they came up with.