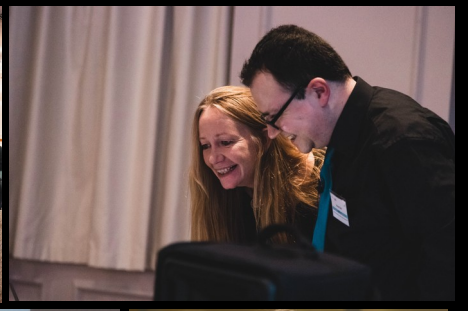
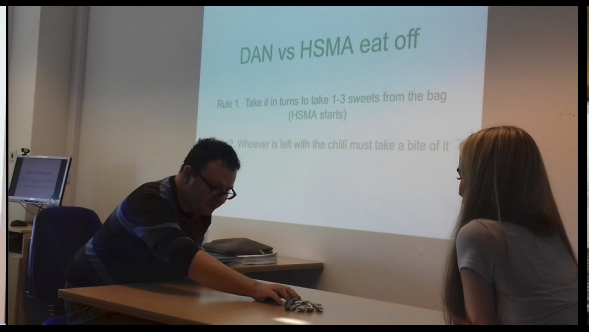


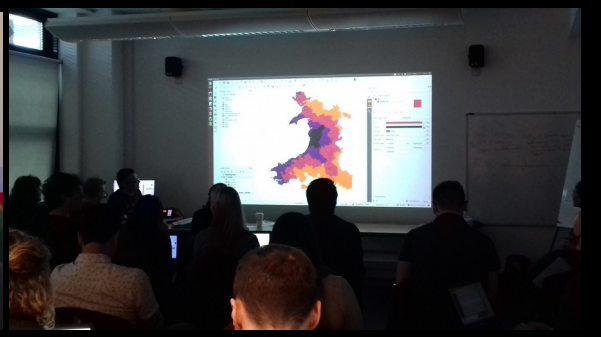


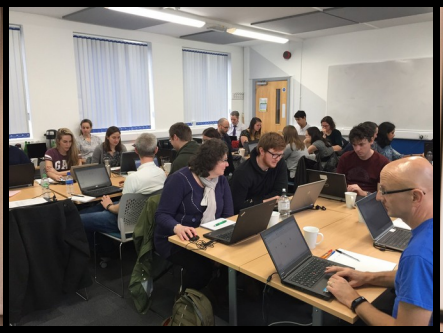


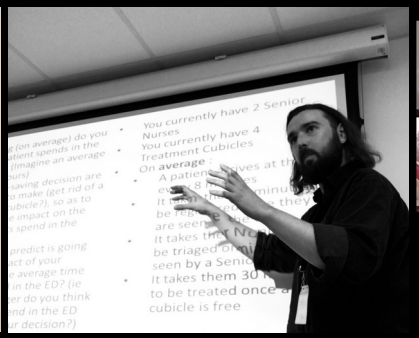













Module 7 : Machine Learning
Session 7B : Ethics in AI
Dr Daniel Chalk

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

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

What could possibly go wrong?

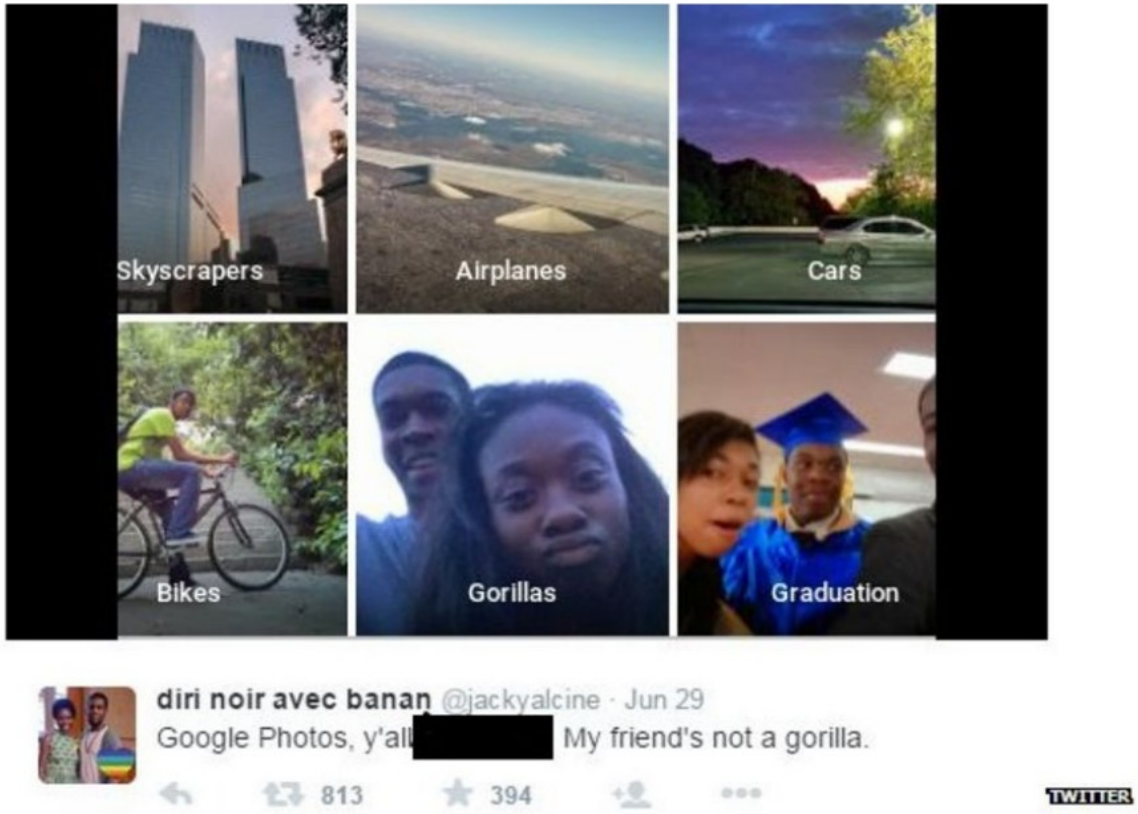
In your groups think about what could possibly go wrong with our headlong rush into AI?

You have 25 minutes. When you return, I'll ask a few groups to share their thoughts.

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.
- For discussion: What problems do you think could arise from biases in word embeddings?

You have 25 minutes + 10 minute break. When you return, I'll ask a few groups to share their thoughts.

Representation - What can we learn?

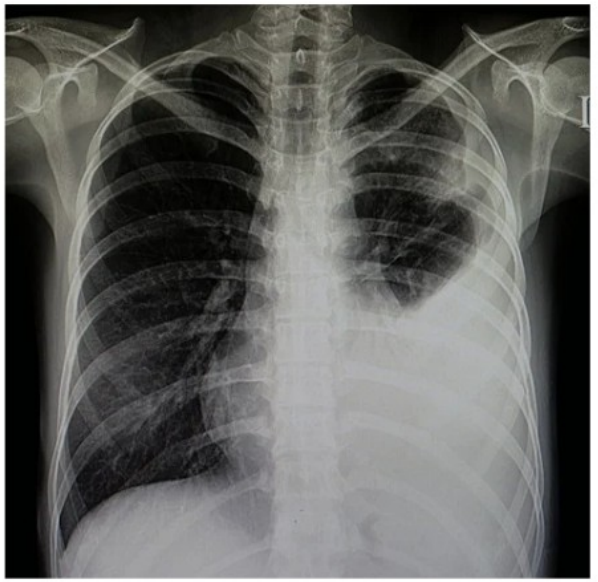
- **Model performance is dependent on sufficient representation of examples in the dataset.**
- **Models should be tested for performance on subgroups**, especially when it is known there are sensitive groups (e.g. protected characteristics like gender and race) present.
- **The model builder is responsible 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.

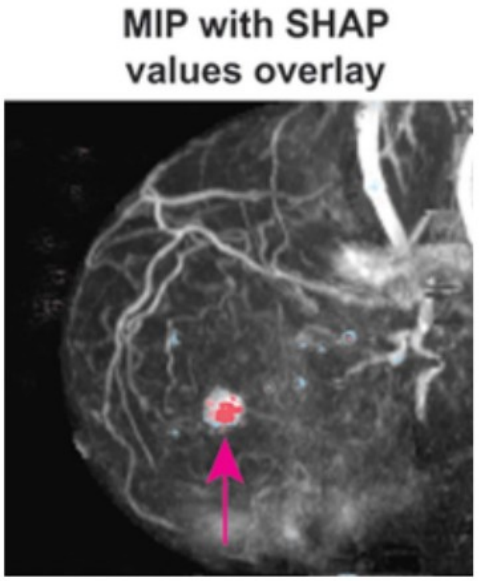


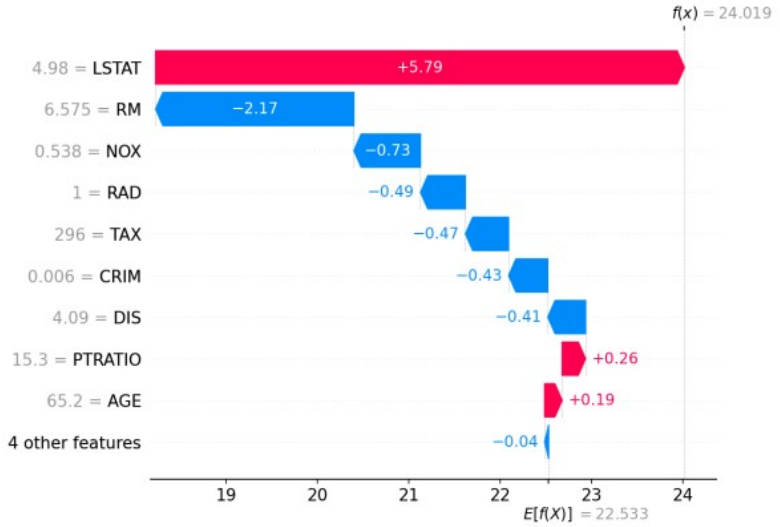
Correlation is not causation - the importance of understanding what drives model prediction

- 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.





Debate 3

In your groups, you will spend the 25 minutes discussing the following. This should look familiar to many of you, as it was one of the questions on the HSMA 4 application for students (not mentors) :

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 question the suspect's presence at the station. Typically the station sees around 200 people attending between 8pm and midnight.

What are your thoughts on the implementation of this system?

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.

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

- Think of a use of machine learning or 'AI' in your area, that could be useful if it performs well.
- Now think just about all the ways it might possibly go wrong. Who could suffer? What might they suffer?
- How would you ensure that you test for these unwanted consequences?
- How would you make sure people understand the possible weaknesses, as well as the strengths, of the model?

You have 25 minutes. When you return, I'll ask a few groups to share their thoughts.

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`