University of Exeter

Ethics and AI

Week 10-BEM2031

Term2: 2023/24

Assignment

Instructions were to knit to PDF or HTML. Some of you had trouble – the option of c&p into Word was there.

.Rmd files means we have to run (and possibly de-bug code) in order to mark it.

We mark anonymously so I am unable to assess updated versions of your submissions – I can only grade what you have submitted.

For future assignments– plan ahead. Technical and personal issues arise. Be ready well ahead of the deadline!

Module leaders have no control over mitigation
Mitigation | Student hubs | University of Exeter

Project

Don’t confuse visualisation and modelling

- Visualisation is the process of translating data into visual elements
- Predictive models are mathematical objects which take inputs (data) and produce outputs

Visualisation is a tool that can be used to evaluate or understand **data (inputs)** OR **models (outputs)**

Project

```
sat_lm <- lm(satisfaction_level ~ ., data = hr)

summary(sat_lm)

plot(sat_lm)




rf <- randomForest(left ~ ., data = hr)

varImpPlot(rf)
```

Today:

- Understand some of the risks and ethical implications of AI, using examples in healthcare and surveillance

- Bias in Data

- Transparency of models

- Consequences

Today:

Some people may find some of the content today upsetting or shocking.

These are some of the very real occurrences where AI has had negative and sometimes severe consequences for peoples’ lives.

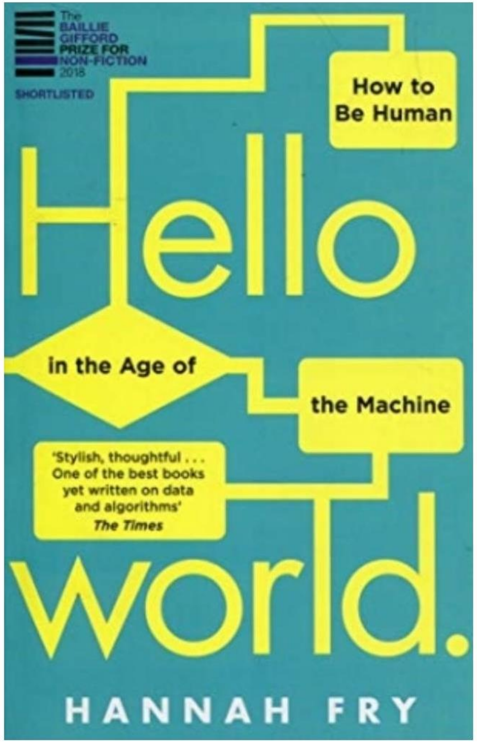
The purpose is to expose you to this reality and ask you to consider how you might take measures to avoid these kinds of negative consequences.

University of Exeter

Ethics in AI and ML

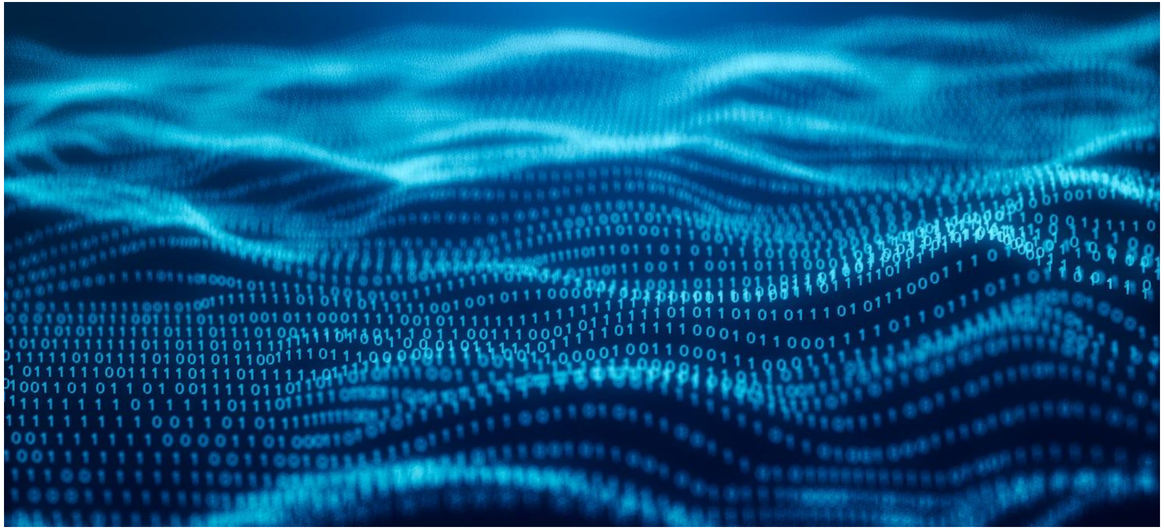
‘ This is a book that takes stock of where we are now, and where we are headed in the not-too-distant future.

It’s a story of the good, the bad and the downright ugly of modern machines, asking how much we should rely on them over our own instincts, and what kind of world we want to live in. ’


The BAILLIE GIFFORD PRIZE FOR NON-FICTION 2018
SHORTLISTED
How to Be Human
Hello
in the Age of
the Machine
'Stylish, thoughtful . . . One of the best books yet written on data and algorithms'
The Times
world.
HANNAH FRY

//Message Delivered
0720/2022
6:11 pm
Location:
Netwo
IMEI:
Device ID:
Phone:
Recieved:
Read
NBC NEWS

Artificial Intelligence (AI) and Machine Learning (ML)



- AI is computer software that mimics the ways that humans think in order to perform complex tasks, such as analysing, reasoning, and learning.

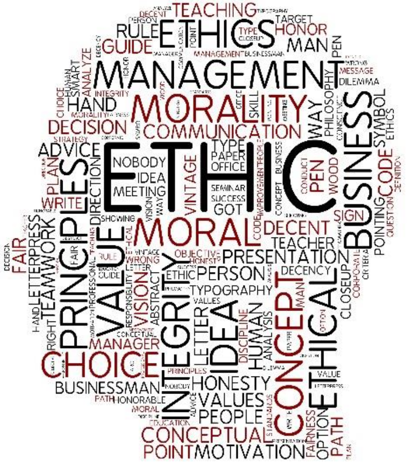
- Organisations don’t ***have*** communication, organisations ***are*** communication (Niklas Luhmann)

Ethics in AI and ML

AI ethics are important because AI technology is meant to augment or replace human intelligence — but when technology is designed to replicate human decisions, the same issues that can cloud human judgment can seep into the technology.

AI projects built on biased or inaccurate data can have harmful consequences, particularly for underrepresented or marginalized groups and individuals.

AI ethics are the moral principles that companies use to guide responsible and fair development and use of AI.



AI in organisations: healthcare

Data: Patient records, operational measures (e.g. times of arrivals, service times), test results, diagnoses, physiological sensors, patient feedback, recruitment data…

Applications: Scanning X-rays for tumours, personalised treatment planning, predict arrivals to emergency departments, efficiently allocate hospital resources, make comparison predictions between hospitals about clinical decisions, sentiment analysis…

University of Exeter

Example: Ethical principles for a medical context

Autonomy: Physicians should have autonomy over diagnostic decisions; however, the accuracy of AI may eventually replace human evaluation.

Non-maleficence: If treatment causes more harm than good, it should not be considered. Recent AI algorithms are programmed to reduce radiation dosage while maintaining diagnostic quality.

Beneficence: A balance of positive benefit and *utility* which balances benefits, risks, costs for best overall results. Straightforward at an individual level, but a population level, trade-offs will be introduced.

Justice: Cancer screening/treatment varies according to geography, racial/ethnic background, access to newer technology, and socioeconomic status. Available datasets may not be representative of the targeted populations, which could vary in terms of cancer risk, prevalence, treatment history, or other salient features.

Explicability: The "black box problem" mentioned in almost every AI discussion, referring to the difficulty of tracing the logic of the outputs of an AI algorithm.

Morgan, M. B., & Mates, J. L. (2023). Ethics of artificial intelligence in breast imaging. *Journal of Breast Imaging*, *5*(2), 195-200.

AI in organisations: business

Data: purchase history, demographics, online behaviour, customer feedback, seasonal trends, cash flow, energy consumption, recruitment data, website traffic….

Applications: personalised marketing, recommendation systems, sales forecasting, fraud detection, financial forecasting, talent acquisition and retention, content optimization, online reputation management….

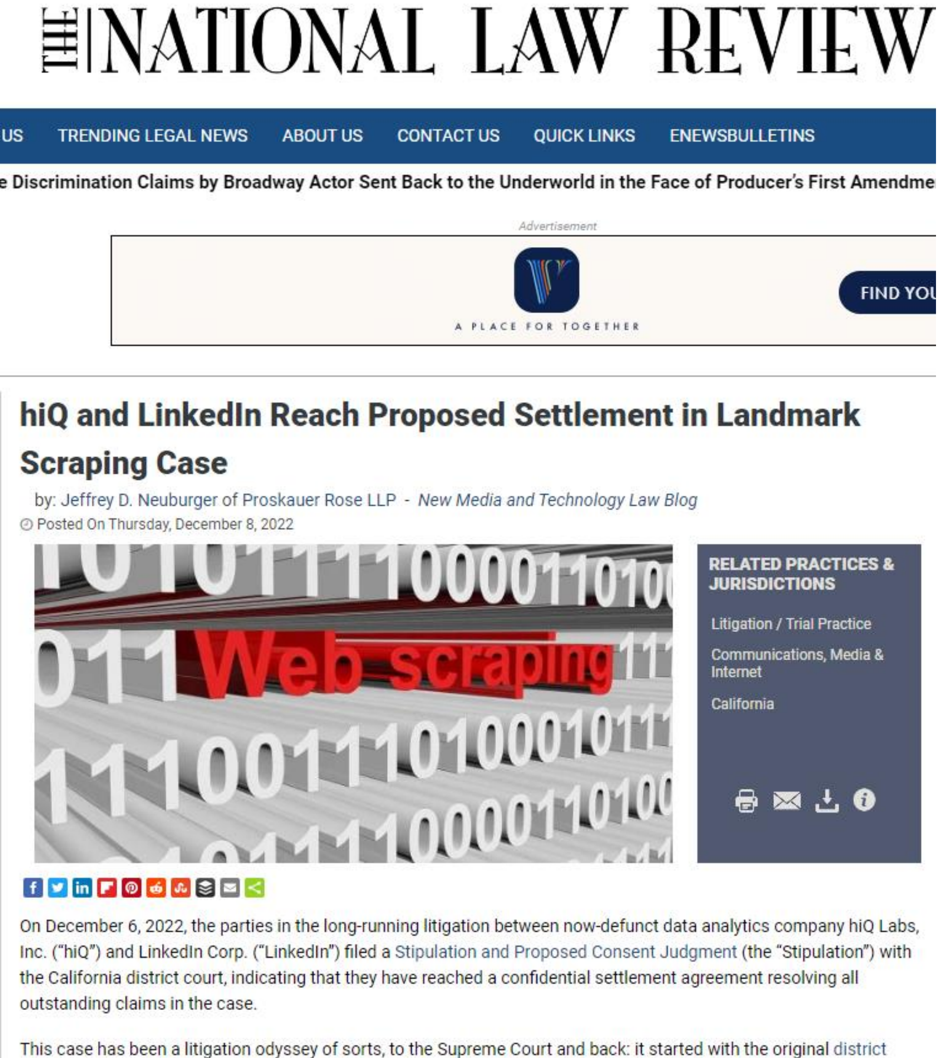
University of Exeter

AI in organisations: final project

"…. Will Alison leave her job ….?"

University of Exeter

Example: Ethically contentious aspects of AI surveillance



e Discrimination Claims by Broadway Actor Sent Back to the Underworld in the Face of Producer's First Amendme

hiQ and LinkedIn Reach Proposed Settlement in Landmark Scraping Case

by: Jeffrey D. Neuburger of Proskauer Rose LLP - *New Media and Technology Law Blog*

Posted On Thursday, December 8, 2022

RELATED PRACTICES & JURISDICTIONS

Litigation / Trial Practice

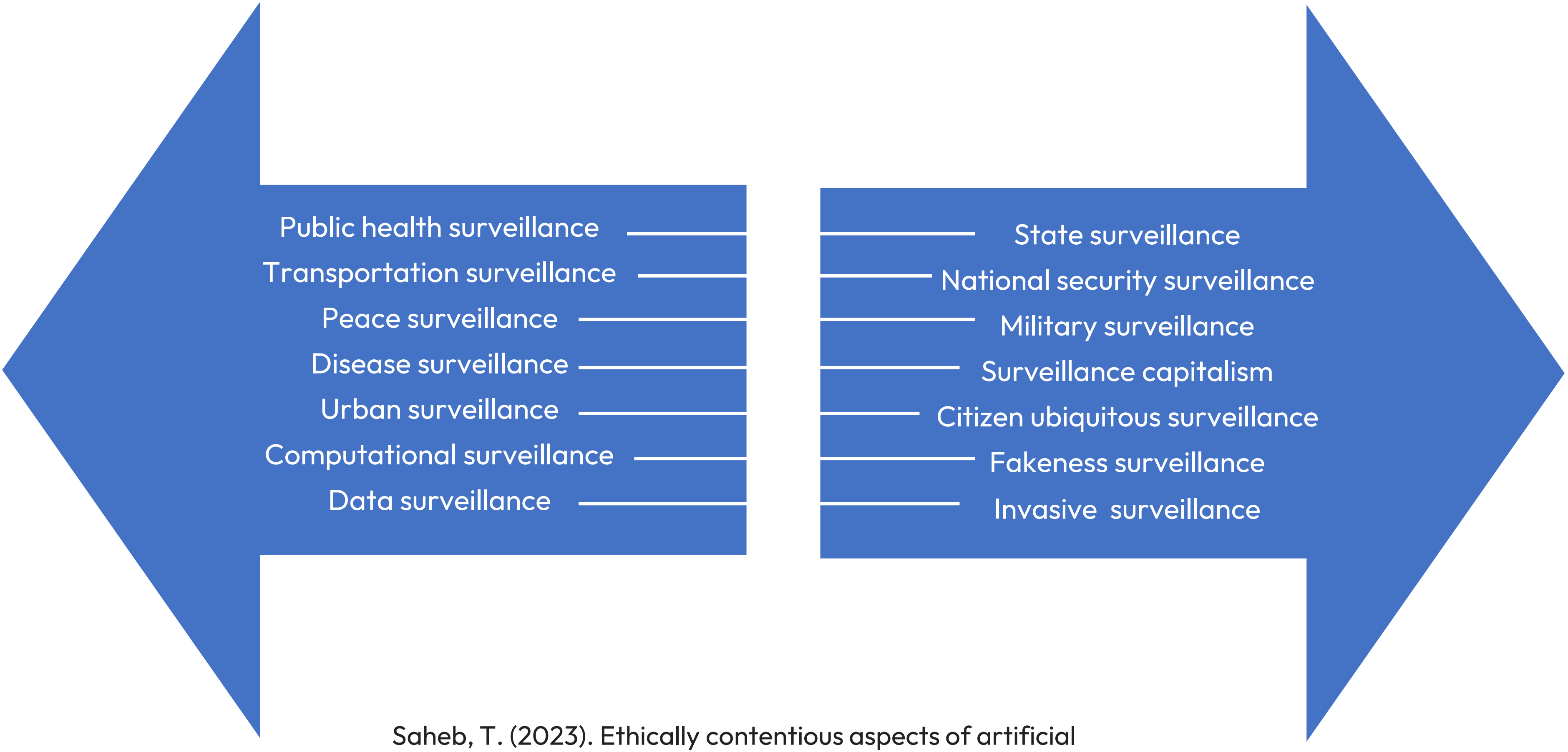
Communications, Media & Internet

California

On December 6, 2022, the parties in the long-running litigation between now-defunct data analytics company hiQ Labs, Inc. ("hiQ") and LinkedIn Corp. ("LinkedIn") filed a Stipulation and Proposed Consent Judgment (the "Stipulation") with the California district court, indicating that they have reached a confidential settlement agreement resolving all outstanding claims in the case.

This case has been a litigation odyssey of sorts, to the Supreme Court and back: it started with the original district

Example: Ethically contentious aspects of AI surveillance



Saheb, T. (2023). Ethically contentious aspects of artificial intelligence surveillance: a social science perspective. *AI and Ethics*, *3*(2), 369-379.

Example: Ethically contentious aspects of AI surveillance

1. **Public Health Surveillance and Privacy**: Examines the balance between public health needs and individual privacy during pandemics.

2. **Video Surveillance and Facial Recognition in Transportation**: Blurred line between facial recognition and video surveillance technology and national security, racial bias and inaccuracies in face recognition algorithms, lack of informed consent.

3. **Military Surveillance**: AI's use in military surveillance, including impacts on civilian privacy and international relations. How does a drone decide who is a non-combatant civilian?

4. **Disease Surveillance**: “Surveillance capitalism“: companies monetize the data collected by tracking citizens' movements and behaviours (e.g. mobile health apps). How do we determine the reliability and validity of innovations and distinguish them from fraudulent AI innovations?

5. **Urban Surveillance in Smart Cities**: AI for monitoring urban environments using sensors and IoT, no consent, little awareness of what, when, why, where, and how it is being stored/shared.

6. **Computational Surveillance**: Processing and analysing large amounts of data to detect fake news and incivility is also being used to promote fakeness, e.g. deepfakes.

7. **Security and Data Surveillance**: Threats of surveillance, cyberattacks, extracting personal information, covert monitoring of digital footprints, behaviours, intentions, preferences.

Saheb, T. (2023). Ethically contentious aspects of artificial intelligence surveillance: a social science perspective. AI and Ethics, 3(2), 369-379 AI and Ethics (springer.com)

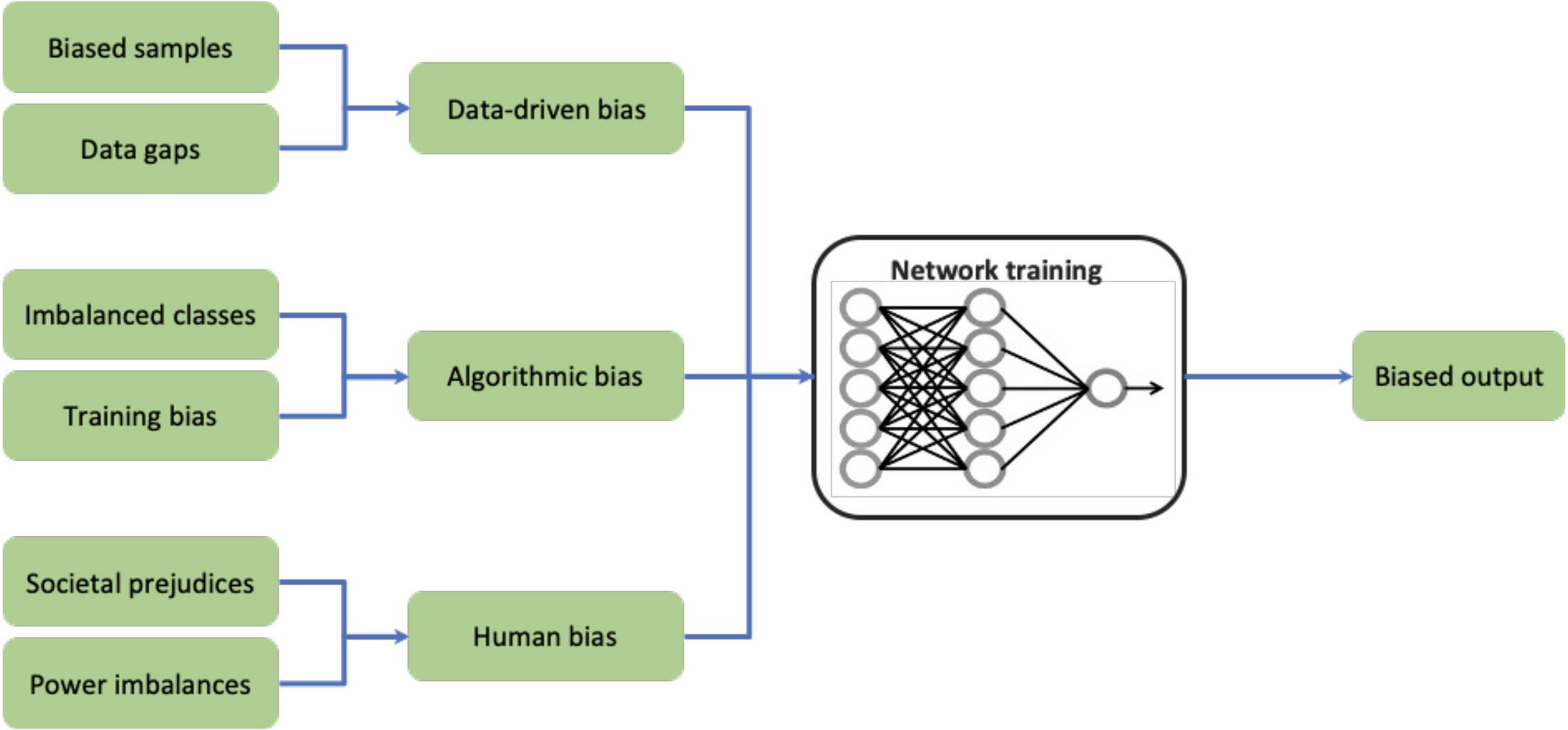
Bias in Data

Convolutional neural networks are able to classify images of skin lesions as accurately as trained dermatologists, in some cases better.

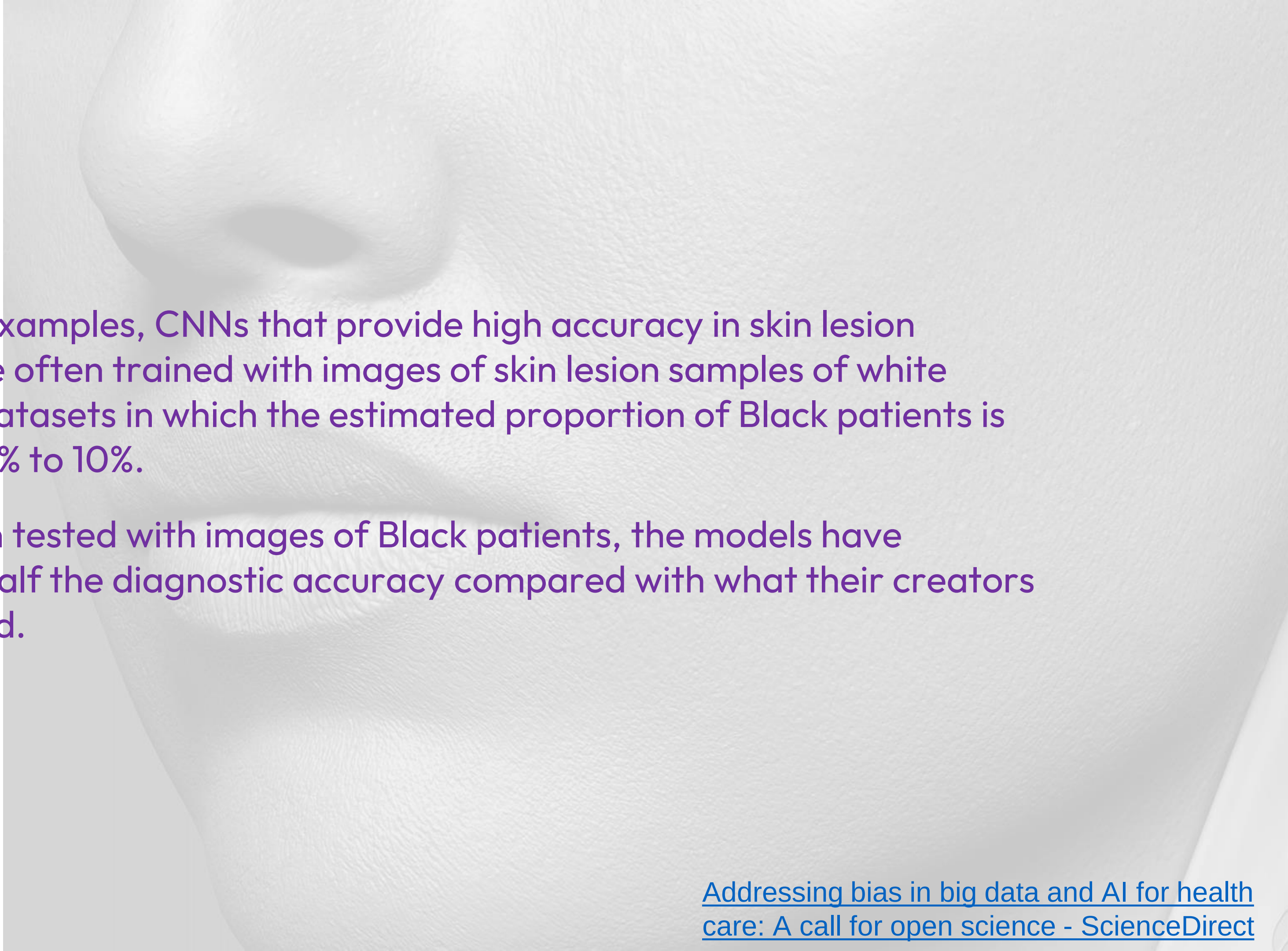
University of Exeter

Addressing bias in big data and AI for health care: A call for open science - ScienceDirect

Bias in data


Biased samples
Data gaps
Data-driven bias
Imbalanced classes
Training bias
Algorithmic bias
Societal prejudices
Power imbalances
Human bias
Network training
Biased output

Bias in data



In one of many examples, CNNs that provide high accuracy in skin lesion classification are often trained with images of skin lesion samples of white patients, using datasets in which the estimated proportion of Black patients is approximately 5% to 10%.

As a result, when tested with images of Black patients, the models have approximately half the diagnostic accuracy compared with what their creators originally claimed.

University of Exeter

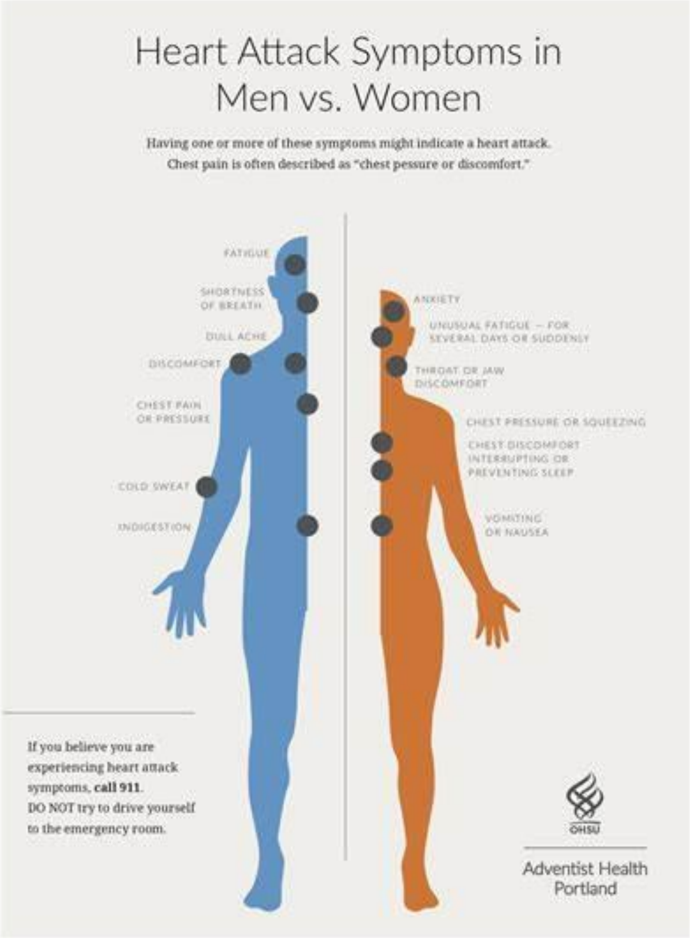
[Addressing bias in big data and AI for health care: A call for open science - ScienceDirect]

Bias in data

In cardiology, a heart attack is overwhelmingly misdiagnosed in women.

Nevertheless, prediction models for cardiovascular disease that claim to predict heart attacks 5 years before they happen are trained in predominantly male datasets.

As cardiovascular disease has different patterns of expression in men versus women, an algorithm that has been trained predominantly with data samples of men may not be as accurate in diagnosing women.



Heart Attack Symptoms in Men vs. Women

Having one or more of these symptoms might indicate a heart attack.
Chest pain is often described as "chest pessure or discomfort."

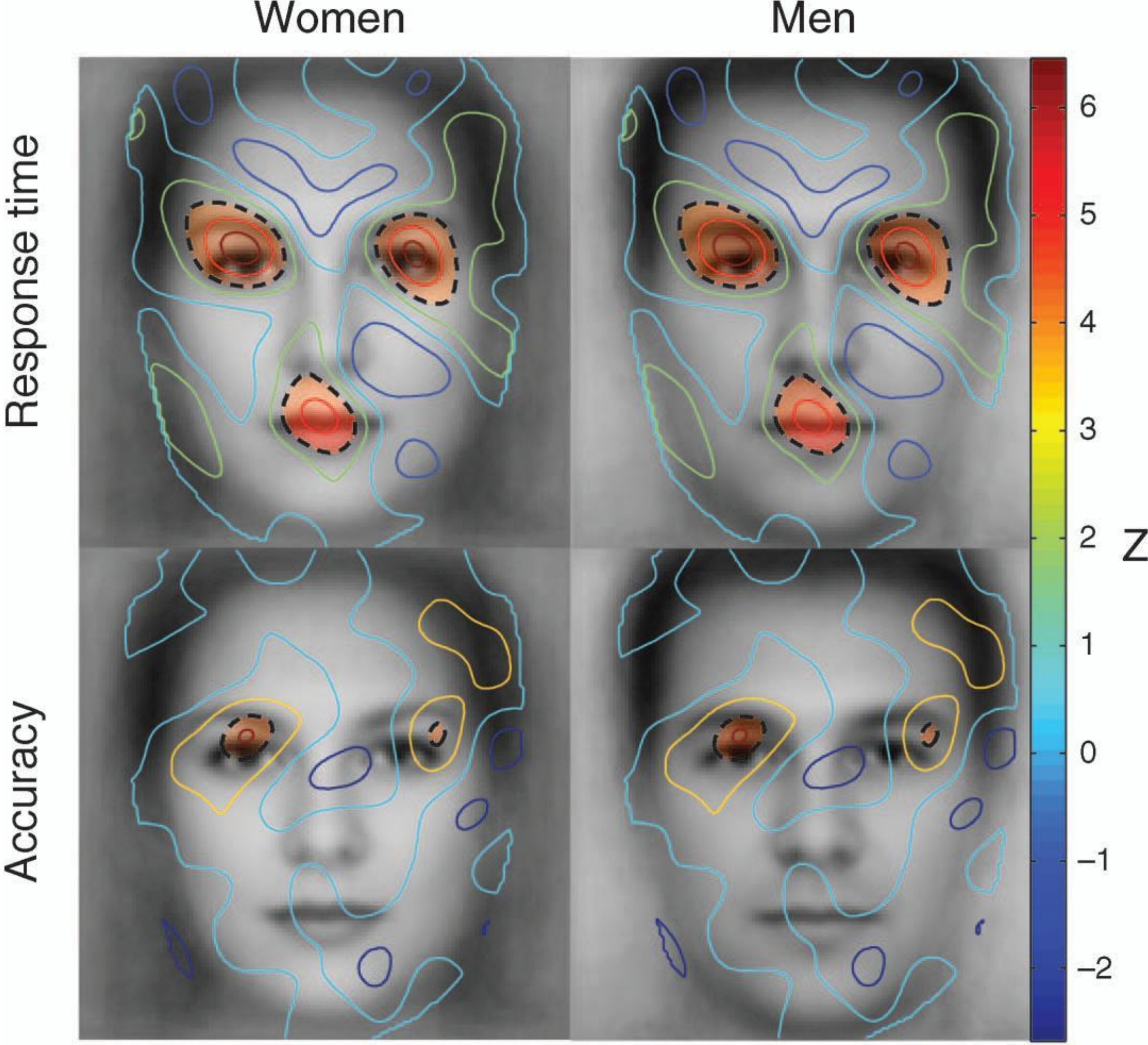
FATIGUE
SHORTNESS OF BREATH
DULL ACHE
DISCOMFORT
CHEST PAIN OR PRESSURE
COLD SWEAT
INDIGESTION
ANXIETY
UNUSUAL FATIGUE – FOR SEVERAL DAYS OR SUDDENLY
THROAT OR JAW DISCOMFORT
CHEST PRESSURE OR SQUEEZING
CHEST DISCOMFORT INTERRUPTING OR PREVENTING SLEEP
VOMITING OR NAUSEA
OHSU
Adventist Health Portland

If you believe you are experiencing heart attack symptoms, **call 911**.
DO NOT try to drive yourself to the emergency room.

Bias in data

Buolamwini, J., Gebru, T. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." Proceedings of Machine Learning Research 81:1–15, 2018 Conference on Fairness, Accountability, and Transparency

The error rate of gender recognition was 100x higher for dark-skinned women than white men



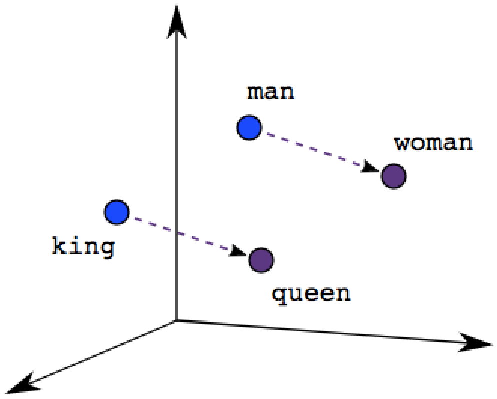
Bias in data

WORD EMBEDDINGS encode words in space, so similar words are located close together, and relationships exist between words.

These embeddings learn relationships from human text.

- Subtracting the word embedding vector for **man** from **king**, and adding the word vector **woman**, gives **queen**.

- Subtracting the word embedding vector for **man** from **doctor**, and adding the word vector for **woman**, gives **nurse**.



Male-Female

Bias in data: representation

Model performance is dependent on sufficient representation of examples in the dataset

Models should be tested for performance on subgroups, e.g. gender, race, age

Transparency of models

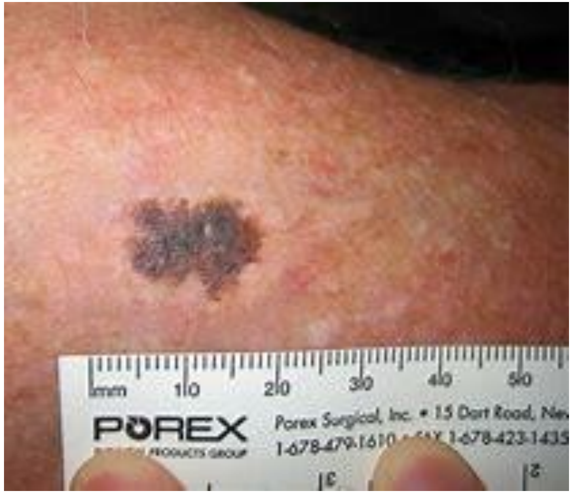
What drives model prediction? Correlation is not causation!

In the mid 1990s, researchers made a NN model for predicting risk of death from pneumonia. It would be used to select patients for intensive care.

A history of asthma seemed to be associated with a lower risk of death.

It turned out that asthma patients have lower risk of death because they receive more intensive care, not because pneumonia is less serious for them!

Transparency of models


mm
10
20
30
40
50
POREX
Porex Surgical, Inc. • 15 Dart Road, Ne
1-678-479-1610 • FAX 1-678-423-1435

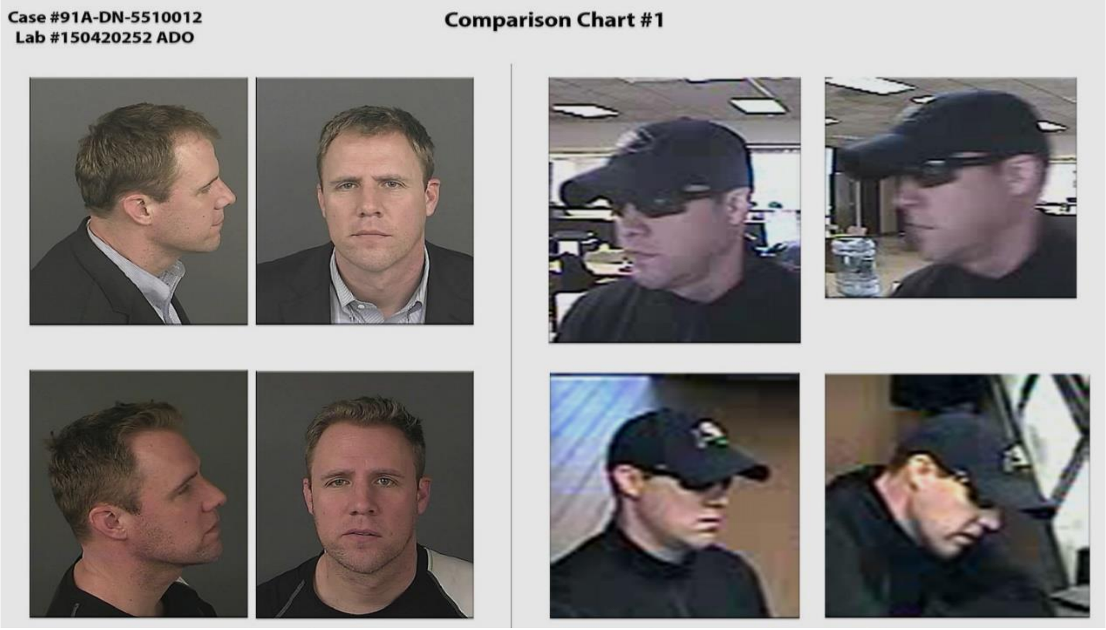
What drives model prediction? Correlation is not causation!

In 2015, dermatologists used a Google image analysis network and trained it on 130,000 skin lesion images to recognise skin cancer.

It outperformed 25 dermatologists.

It turned out their system was much more likely to classify any image with a ruler in it as cancerous; the model had learned that images with rulers were more likely to be cancerous.

Transparency of models



Steve Talley was asleep at home in South Denver in 2014 when he heard a knock at his door. Someone claimed to have hit his car and invited him outside to have a look.

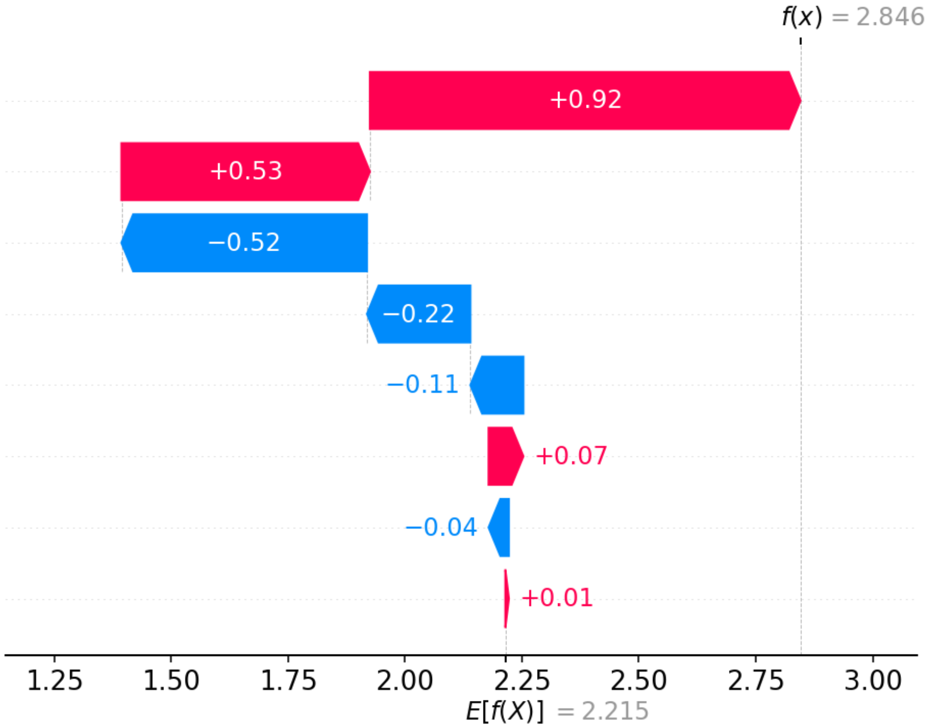
He went outside and was knocked to the pavement, flash bang grenades were detonated, and he was severely beaten and badly injured by three men with batons and the butt of a gun.

Steve was beaten by police. He was misidentified as a bank robber by an AI face recognition system looking at CCTV footage.

Transparency of models

Shapley values – show the influence of different data features, including image pixels, and the extend and direction of influence of the feature.

Red – increase model output value
Blue – reduce model output value



An introduction to explainable AI with Shapley values — SHAP latest documentation

Transparency of models

Artificial intelligence in its current state is unfair, easily susceptible to attacks and notoriously difficult to control. Often, AI systems and predictions amplify existing systematic biases even when the data is balanced. Nevertheless, more and more concerning uses of AI technology are appearing in the wild. This list aims to track all of them. We hope that Awful AI can be a platform to spur discussion for the development of possible preventive technology (to fight back!).

daviddao/awful-ai: 😈Awful AI is a curated list to track current scary usages of AI - hoping to raise awareness (github.com)

Example: The Fall of Zillow

Zillow was a part of the real estate business that made use of in-house algorithms to predict price fluctuations.

Because their ML model was poorly trained, Zillow suffered a loss of over $300 million; they tried to buy approximately 1,900 homes using their algorithm. Moreover, their financial losses lead to the mass firing of their employees.

This can be attributed to the quality of data. Small deviations in data quality may cause consequential discrepancies if the stakes are too high. A simple error in the number of rooms of property, even if it throws off the model by 10% for a million-dollar property, will amount to **$100,000.**

University of Exeter

Example: Amazon AI Tool Recruiting bias

Amazon successfully uses automation tools for e-commerce, hiring, warehouse price prediction, and user behaviour clustering.

Amazon used to hire candidates based on a score calculated by their algorithm on a scale of one to five. Amazon discovered that their algorithm was not scoring the candidates in a gender-neutral manner.

Amazon used 10-year male-dominant hiring data to train their dataset. This produced bias against women in their models' performance. In this case, the data that was fed to the model was not scanned for gender neutrality.

Example: Amazon face recognition flags politicians as criminals

Amazon expanded its market by selling facial recognition to law enforcement agencies.

The American Civil Liberties Union tried to compare images of more than 20,000 criminals with members of congress. Rekognition flagged 28 members of congress as criminals involved in mugshot activities, with a high confidence of up to 80%.

Biased datasets created biased models.

University of Exeter

Example: Ball, or bald head?

The popular Scottish soccer team Inverness Caledonian Thistle FC deployed an AI camera in 2020 to analyse live games. It was specifically designed to track a soccer ball during a livestream match.

However, after deploying this tool and using it during the livestream of a match, the AI tool flagged a player’s bald head as a football.

Because of this, the commentators had to repeatedly apologize since the viewers missed crucial moments of the game.

This is attributable to lack of training with objects that are visually similar to the main object under tracking.

Example: Uber's self-driving car hits jaywalker

Uber's Self-driving car hit and killed a woman, Elaine Herxberg, while she was jaywalking in March 2018. The woman was 49 years old when she was hit by an SUV running 40 km an hour in self-driving mode.

The vehicle was not able to classify the jaywalker as a person. Also, the tracking models were not accurately able to predict the path that the object was going to take.

The classification and the tracking models failed, given that the "context" of the presence of a person was different from what these models were trained in.

University of Exeter

Next Week: Week 11

There will be no lecture on 26 March, but I will be here to support/discuss your project.

Workshops on 21 March, and 25 March will be project support.

Final project due 28 March – there will be NO WORKSHOPS THAT DAY! Please submit on time!

Any questions?

?