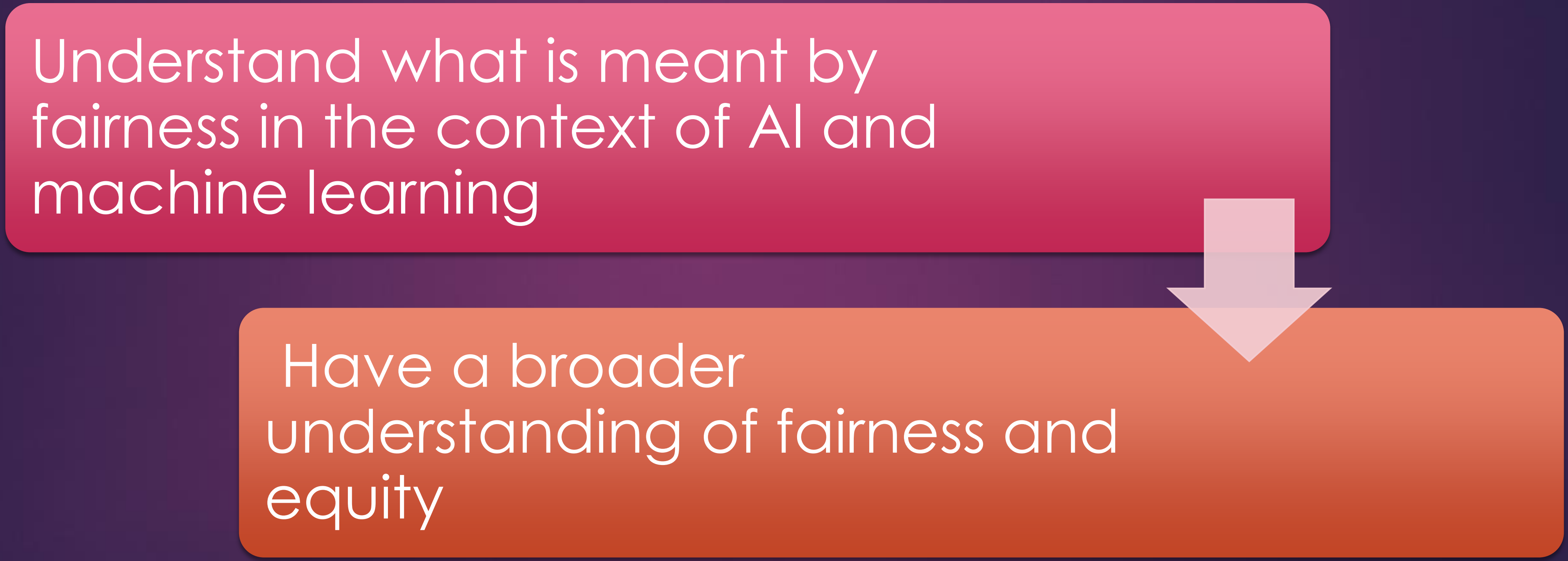
Lecture 7

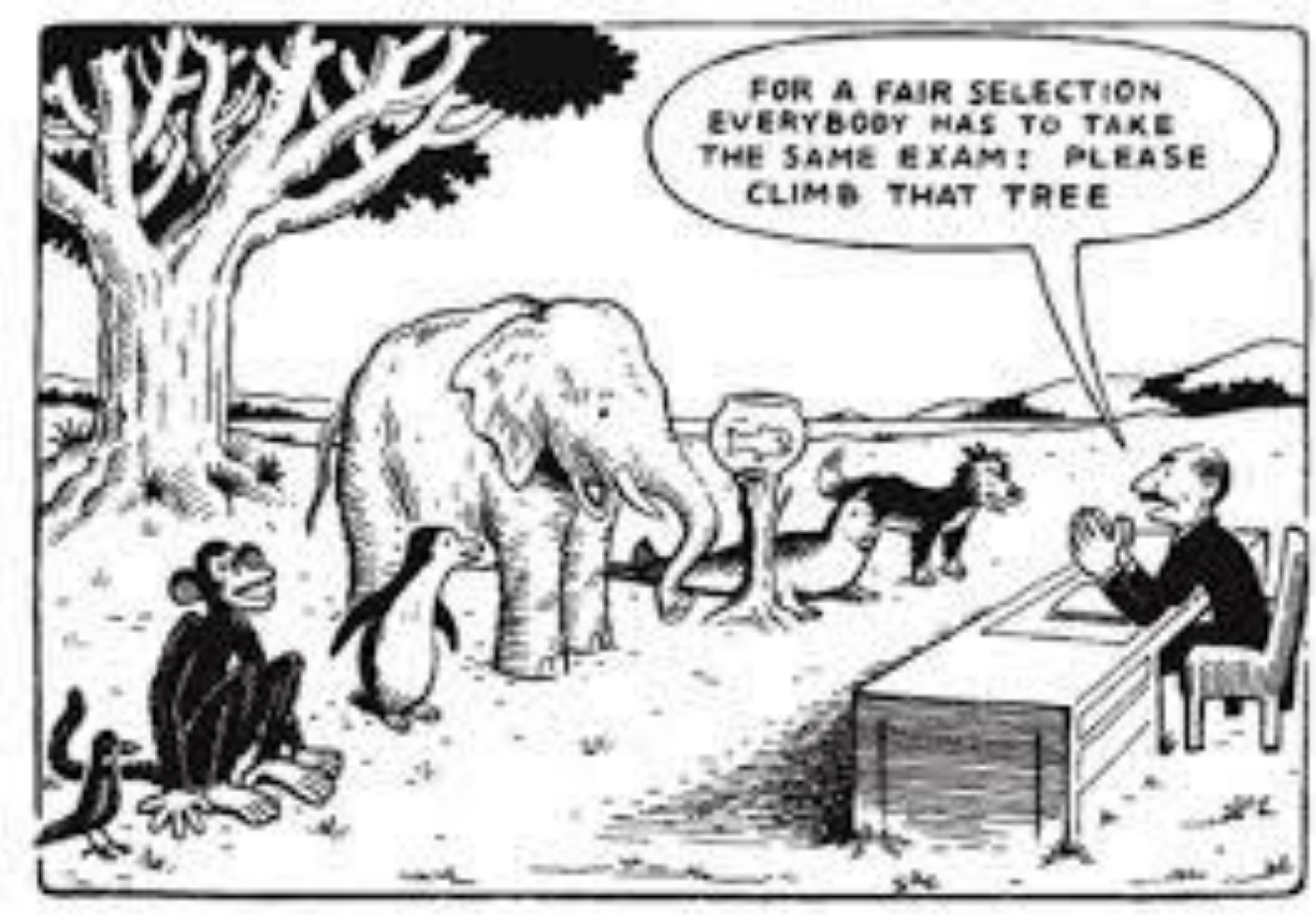
DATA ETHICS FRAMEWORK-FAIRNESS

Agenda

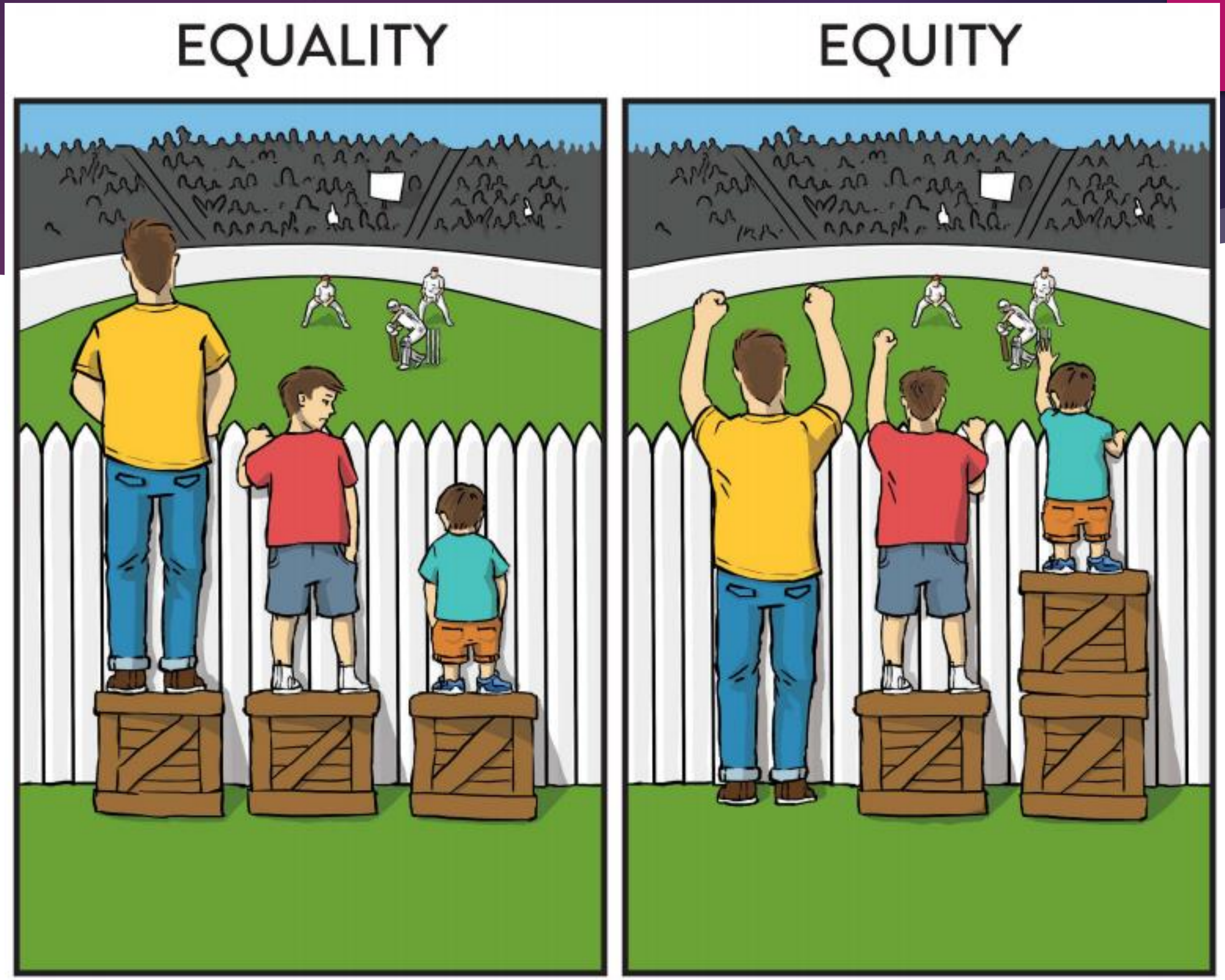


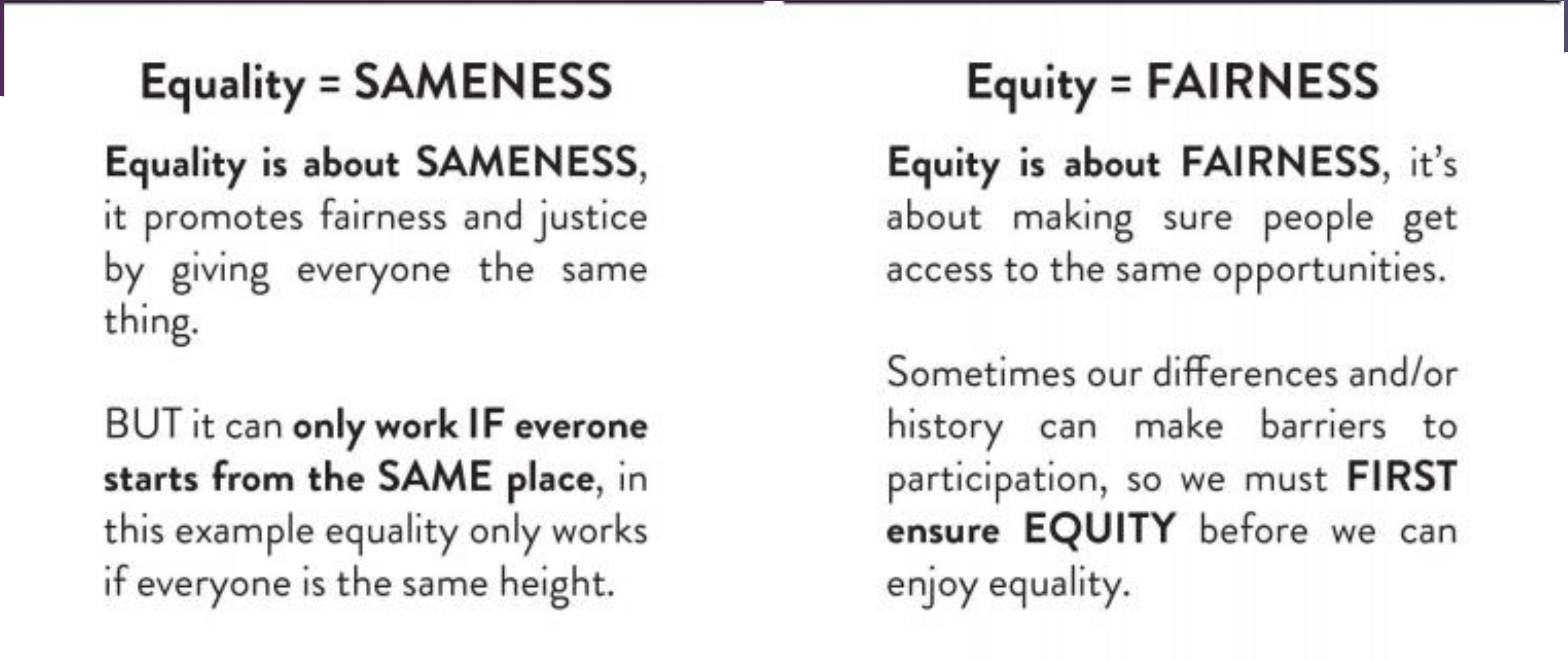
Understand what is meant by fairness in the context of AI and machine learning

Have a broader understanding of fairness and equity



FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE


EQUALITY
EQUITY



Equality = SAMENESS

Equality is about SAMENESS, it promotes fairness and justice by giving everyone the same thing.

BUT it can **only work IF everone starts from the SAME place**, in this example equality only works if everyone is the same height.

Equity = FAIRNESS

Equity is about FAIRNESS, it's about making sure people get access to the same opportunities.

Sometimes our differences and/or history can make barriers to participation, so we must **FIRST ensure EQUITY** before we can enjoy equality.

Equality

- No studies currently exist that can fully answer the question of what people understand by the terms equality, fairness and good relations.(*Equality and Human Rights Commission Research report 53)*
- Understanding of the concepts of equality, fairness and good relations were related to both personal experiences and perceptions and the context in which they were discussed. They were seen as being distinct from each other yet highly interwoven and, at times, interdependent.
- There were two broad views of fairness: the first saw fairness as treating everyone the same regardless of their characteristics; the second saw it as treating people differentially according to their characteristics, but these were not mutually exclusive.

- Equality was broadly split into equality of opportunity and equality of outcome. While the first of these was seen as desirable the second was seen as neither desirable nor achievable.
- Although fairness and equality were sometimes used interchangeably, people generally felt that fairness was something that occurred on a personal level, between individuals and communities, whereas equality was something that could be legislated for and happened on a bigger scale.

How do we define it?

- **Equality** is when everyone is treated in the same way, without giving any effect to their need and requirements. The central idea of equality is that all the individuals gets equal treatment in the society and are not discriminated on the basis of race, sex, caste, creed, nationality, disability, age, religion and so forth.
- **Equity** can be defined as the quality of treating individuals fairly based on their needs and requirements. Equity ensures that all the individuals are provided the resources they need to have access to the same opportunities.

UK data ethics FW

What is it for?

- The Data Ethics Framework guides appropriate and responsible data use in government and the wider public sector. It helps public servants understand ethical considerations, address these within their projects, and encourages responsible innovation.

Fairness

- It is crucial to eliminate your project's potential to have unintended discriminatory effects on individuals and social groups. You should aim to mitigate biases which may influence your model's outcome and ensure that the project and its outcomes respect the dignity of individuals, are just, non-discriminatory, and consistent with the public interest, including human rights and democratic values.

Score the fairness of your project from 0 to 5 where:

- 0 means there is a significant risk that the project will result in harm or detrimental and discriminatory effects for the public or certain groups
- 5 means the project promotes just and equitable outcomes, has negligible detrimental effects, and is aligned with human rights considerations

1.2 Understand unintended consequences of your project (fairness)

What would be the harm in not using data? What social outcomes might not be met?

What are the potential risks or negative consequences of the project versus the risk in not proceeding with the project?

Could the misuse of the data or algorithm or poor design of the project contribute to reinforcing social and ethical problems and inequalities?

What kind of mechanisms can you put in place to prevent this from happening?

What specific groups benefit from the project? What groups can be denied opportunities or face negative consequences because of the project?

1.3 Human rights considerations (fairness)

- How does the design and implementation of the project or algorithm respect human rights and democratic values?
- How does the project or algorithm work towards advancing human capabilities, advancing inclusion of underrepresented populations, reducing economic, social, gender, racial, and other inequalities?
- What are the environmental implications of the project? How could they be mitigated?

3.6 Ensure the project's compliance with the Equality Act 2010 (fairness)

- Data analysis or automated decision making must not result in outcomes that lead to discrimination as defined in the Equality Act 2010.
- How can you demonstrate that your project meets the Public Sector Equality Duty?
- What was the result of the Equality Impact Assessment of the project?
- **3.7 Ensure effective governance of your data**
- Organisations have a responsibility to keep both personal data and non-personal data secure.
- How have you ensured that the project is compliant with data governance policies within your organisation?
- **3.8 Ensure your project's compliance with any additional regulations**
- Consider additional relevant legislation and codes of practice.

4.3 Bias in data (fairness)

- How has the data being used to train a model been assessed for potential bias? You should consider:
  - Whether the data might (accurately) reflect biased historical practice that you do not want to replicate in the model (historical bias)
  - The data might be a biased misrepresentation of historical practice, for example because only certain categories of data were properly recorded in a format accessible to the project (selection bias)
- If using data about people, is it possible that your model or analysis may be identifying proxy variables for protected characteristics which could lead to a discriminatory outcome? Such proxy variables can potentially be a cause of indirect discrimination; you should consider whether the use of these variables is appropriate in the context of your service (i.e. is there a reasonable causal link between the proxy variable and the outcome you're trying to measure?; do you assess this to be a proportionate means to achieve a legitimate aim in accordance with the Equality Act 2010?)
- What measures have you taken to mitigate bias?

5.2 Repeatedly revisit the user need and public benefit throughout the project (fairness)

- How has the user need changed?
- Is the project still benefiting the public?
- Has there been a change of circumstances that might have affected the initial understanding of the public benefit in the project? If so, how can you adjust it?
- How sufficient is the current human oversight of the automated project?
- If any unanticipated harms emerged during the project, how have they been mitigated?

Why Do We Care About Fairness More Than Ever?

- We are at an age where many things have become or are becoming automated by ML systems.
- Mass use of self-driving cars is around the corner and are estimated to be widely used within 5–10 years; employers use ML system to select job applicants;
- courts in United States use COMPAS algorithm for recidivism prediction; Linked-in uses ML to rank job candidates queried;
- Amazon uses recommender system to recommend items and decide the order of items appearing on a page.
- Netflix uses recommender system to present customized page for every user.

Machine learning systems have been an inseparable part of our daily lives. They are becoming even more widely used as more and more fields begin to integrate AI into their existing practice/products.

A FRIGHTENING LACK OF REGULATORY OVERSIGHT

TRT WORLD
Algorithm
&
prejudice

Let's add some context!

- **Biases in AI remains an increasing phenomenon. To overcome bias, you need to apply the aspect of fairness. SO then we need to understand fairness first.**
- The fairness of AI algorithms is a growing field of research too that arises from the general need for decisions to be free from bias and discrimination.
- Fairness also applies to AI-based decision tools, where the European White Paper on AI provides a framework in which AI or algorithmic decision-making needs to be carefully considered.
- As AI systems increasingly take over modern life, we need to ensure that they work fairly for all is an important challenge.
- Researchers have identified unfair gender and racial bias in existing AI systems.
- With all of the recent news about data breaches and tech giants pledging to #DeleteFacebook, something unexpected is happening: people are actually starting to care about online privacy.
- At the same time, awareness has been building around the real and potential harms of algorithmic decision-making systems. Scholarly organizations like FAT and AI Now have been getting more attention, and concern about algorithmic harms.

Context ..continued

- Machine learning models have been shown to amplify existing human biases . This is worrying, especially because nowadays many decisions are taken based on the results of machine learning models. In a world where people are fighting for equality, ensuring fair behavior of models should be a top priority.
- Fairness modeling is a field in artificial intelligence that ensures that the outcome of machine learning models is not influenced by protected attributes like gender, race, religion, sexual orientation etc.
- Fairness — It is crucial to eliminate your project’s potential to have unintended discriminatory effects on individuals and social groups.
- You should aim to mitigate biases which may influence your model’s outcome and ensure that the project and its outcomes respect the dignity of individuals, are just, nondiscriminatory, and consistent with the public interest, including human rights and democratic values.

An example of ‘Fairness’ in an AI system

- Sweeney showed in 2013 that, when people searched for African-American sounding names, Google displayed advertisements that suggested that somebody had an arrest record. For white-sounding names, Google displayed fewer ads suggestive of arrest records. Presumably, Google’s AI system analysed people’s surfing behaviour and inherited a racial bias.
- Princeton Review, a US company that offers online tutoring services, charged different prices in different areas in the US, ranging from 6600 to 8400 dollars. Presumably, the costs for delivering the service were the same for each area, as the company offers its tutoring service over the Internet.
- The AI behind automated translation tools can also reflect inequality and discrimination. If people type "He is a doctor. She is a nurse" into Google Translate and translate the phrases into Turkish, Google Translate provides: "O bir hemşire. O bir doktor". Those Turkish sentences are gender-neutral; Turkish does not differentiate between the words "he" and "she". When translating the Turkish text into English again, Google Translate provides: "She is a nurse. He is a doctor".

Fairness

- Research about fairness in machine learning is a relatively recent topic. Most of the articles about it have been written in the last three years
- Some of the most important facts in this topic are the following:
  - In 2018, IBM introduced AI Fairness 360, a Python library with several algorithms to reduce software bias and increase its fairness.
  - In 2018, Facebook made public their use of a tool, Fairness Flow, to detect bias in their AI. However, the source code of the tool is not accessible, and it is not known whether it really corrects bias.
  - In 2019, Google published a set of tools in GitHub to study the effects of fairness in the long run.

Fairness....continued

- The algorithms used for assuring fairness are still being improved. However, the main progress in this area is that some big corporations are realizing the impact that reducing algorithmic bias could have on society.
- An example of the controversial use of an algorithm is the way that Facebook allocates news articles to users, which some people have complained can introduce political bias. Before elections, some candidates have tried to use Facebook for campaigning purposes, which has become a hotly disputed area.
- Many a complaint is that algorithms often cannot be inspected to ensure that they are operating fairly. But many commercial companies prefer to not reveal the details of the algorithms that they use, as they frequently state that it could assist rival companies to benefit from their technologies.

Issues with Ensuring Fairness in ML algorithms

“Imagine a scenario in which self-driving cars fail to recognize people of color as people—and are thus more likely to hit them—because the computers were trained on data sets of photos in which such people were absent or underrepresented,” Joy Buolamwini, a computer scientist and researcher at MIT, told Fortune in a recent interview.

Buolamwini’s research revealed that facial recognition software from tech giants Microsoft, IBM and Amazon, among others could identify lighter-skinned men but not darker-skinned women.

How does this happen? It happens because of something that is mounting alarm: algorithmic bias.

“Machines, like humans, are guided by data and experience.” If that data or experience is mistaken or crooked, a biased decision can be made, whether that decision is made by a human or a machine.

And There You Go Folks – It’s The DATA That’s The Culprit!

- Data science allows users to make informed decisions based on analysis conducted on datasets, yet, we must ask ourselves if these decisions are going to be ethical? The answer is not so straightforward, as the insights from the analysis might be misinterpreted or misrepresented, which can lead to unethical decisions being made. A situation like this is usually caused by the existence of bias within the dataset.

Bias within datasets can be caused by:

- Datasets that do not accurately represent the cohort that the insights will be based on.
- Datasets produced by humans, which can be in the form of curated news articles or social media content, which leads to bias against a group of people.

- There are huge ethical implications of having bias within datasets that can lead to biased models that will be prejudiced and harmful towards people. An example of this is in the following case study, which is about “COMPAS Recidivism Algorithm” that was used to predict a defendant’s likelihood to commit a crime. (Usman, IOC, 2020)
- Source: Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. “Machine bias: There’s software used across the country to predict future criminals, and it’s biased against blacks”. ProPublica (May 23, 2016).

A Real-Life Example Of Data Bias

Causes of Bias : Skewed sample bias and fairness issue

- If by some chance some initial bias happens, such bias may compound over time: future observations confirm prediction and fewer opportunity to make observations that contradict prediction.

Recap of previous lecture …

- **Limited features**: features may be less informative or reliably collected for minority group(s). If the reliability of the label from a minority group is much lower than the counterpart from a majority group, the system tends to have much lower accuracy for the prediction of the minority group due to these noise.

- **Sample size disparity**: If the training data coming from the minority group is much less than those coming from the majority group, it is less likely to model perfectly the minority group.

- **Proxies**: Even if sensitive attribute(attributes that are considered should not be used for a task e.g. race/gender) is not used for training a ML system, there can always be other features that are proxies of the sensitive attribute(e.g. neighborhood). If such features are included, the bias will still happen. Sometimes, it is very hard to determine if a relevant feature is too correlated with protected features and if we should include it in training.

Causes Of Bias…Continued

- **Old or incomplete data sets** are another source of potential bias, both in the development and implementation of algorithmic models, she said, noting that some of the data sets that developers are using to build models are more than 40 years old, and that old data can contain hidden biases.
- **Incomplete data sets** can be just as misleading. If, for example, the data is incomplete because (as has been the case in the past with credit scores) women and minorities are unequally represented. In such cases, Nielsen observed, models based on incorrect or incomplete data can create informational “feedback loops” that strengthen and reinforce even small biases in the data. And in any case, minorities are by definition less visible to machine-learning algorithms because most of the training data used to “teach” algorithms is skewed in favor of the majority, she added.
- **Incorrect or imprecise “labeling”** can also cause bias and fairness problems, Nielsen said. Labeling is how data scientists annotate and classify certain properties and characteristics of a data point in order to make it searchable by an algorithm. Nielsen demonstrated the importance of labeling by doing a Google search for the phrase “unprofessional hairstyles for work,” which resulted in a collection of images that are predominately women, many of them Black. She then modified the search by adding the word “men,” which again yielded many Black hairstyles, including those of some women and certain people — like Harvard professor Cornel West — who have impeccable professional credentials.

What Actions Can Be Taken To Ensure Fairness In AI Models?

One school of thought is to measure stakeholder belief about fairness - Nina Grgić-Hlača et al, 2010

For example, lets assume that as a society we are uncomfortable to use race as a measure of criminality. To capture and quantify this belief, these researchers asked the stakeholders the following questions,

- What fraction of users will consider a feature Fair to be used BEFORE they know how it will impact an algorithum's accuracy ?
- What fraction of users will consider the same feature fair AFTER they know the impact of it on the algorithum's accuracy?
- What fraction of users will consider the same feature(race) fair if they know the impact of if they know it will result in an unfair outcome for the algorithum's accuracy?

Based on how the users who answers changed with the different questions , they proposed attaching weights to the variable in question to see how relevant or ethical it was to use the variable in the algorithm.

What Other Actions Can Be Taken To Ensure Fairness In AI Models?

Quantify - Outcome Fairness

- We can analyze the real-world data and observe the bias. Then we can generate data based on real-world data and observed bias. If you want the perfect dummy data generator, you'll need to provide a fairness definition that will try to transform biased data into something that can be deemed fair.
- The expected degree of how much the Algorithm can transform the biased data can be a measurement percentage which can be set as a standard

Source: https://broutonlab.com/blog/ai-bias-solved-with-synthetic-data-generation

ACTIONS To Promote FAIRNESS IN AI MODELS

Trade-Off Between Fairness & Accuracy

The impact of imposing the above constraints on the accuracy truly depends on the dataset, the fairness definition used as well as the algorithms used. In general, however, fairness hurts accuracy because it diverts the objective from accuracy only to both accuracy and fairness. Therefore, in reality, a trade-off should be made