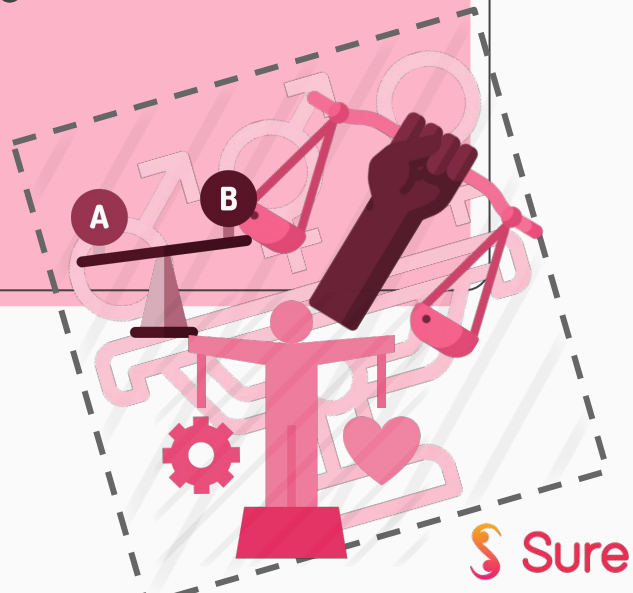
BIASES AND ETHICS



IS YOUR ARTIFICIAL INTELLIGENCE BIASED?

AI bias is the underlying **prejudice in data** that's used to create AI algorithms, which can ultimately result in discrimination and other social consequences.

RACISM EMBEDDED IN US HEALTHCARE

In October 2019, researchers found that an algorithm used on more than 200 million people in US hospitals to predict which patients would likely need extra medical care heavily favored white patients over black patients.

COMPAS IN US COURT SYSTEM

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) model predicted twice as many false positives for recidivism for black offenders (45%) than white offenders (23%).

BIASED AUTOMATED HIRING ALGORITHM

An algorithm used by Amazon for hiring employees was found to be biased against women. This was because the algorithm was based on the number of resumes submitted over the past ten years, and since most of the applicants were men, it was trained to favor men over women.

CURIOUS CASE OF CREDIT LIMIT

Apple Credit Card users noticed that it seemed to offer smaller lines of credit to women than to men. The algorithm had been vetted for potential bias by a third party; moreover, it doesn't even use gender as an input.

SureStart

CAUSES OF BIAS

The outputs produced by machine learning models and AI systems are simply a reflection of the training datasets they are exposed to.

PERSONAL BIAS

The data gatherers, participants in a study might hold a bias for or against certain age group, race, gender.

ENVIRONMENTAL BIAS

The experimental setup or surroundings might have intentionally or unintentionally imposed some bias within the data gathering process.

TYPES OF BIAS

SAMPLE BIAS

When a dataset does not reflect the realities of the environment in which a model will run.

E.g., a facial recognition systems trained primarily on images of white men.

EXCLUSION BIAS

Case of deleting valuable data thought to be unimportant.

E.g., for a dataset of customer sales in America and Canada. 98% of the customers are from America, so you choose to delete the location data thinking it is irrelevant.

MEASUREMENT BIAS

When the data collected for training differs from that collected in the real world, or when faulty measurements result in data distortion.

E.g., in image recognition datasets, where the training data is collected with one type of camera, but the production data is collected with a different camera.

OBSERVER BIAS

Observer/confirmation bias is the effect of seeing what you expect to see or want to see in data.

E.g., researchers go into a project with subjective thoughts about their study, either conscious or unconscious.

ASSOCIATION BIAS

When the data for a machine learning model reinforces and/or multiplies a cultural bias.

E.g., your dataset may have a collection of jobs in which all men are doctors and all women are nurses. This does not mean that women cannot be doctors, and men cannot be nurses.

RECALL BIAS

When you label similar types of data inconsistently. This results in lower accuracy.

E.g., in a team labeling images of phones as damaged, partially-damaged, or undamaged. If someone labels one image as damaged, but a similar image as partially damaged, your data will be inconsistent.

SureStart

CONSIDERATIONS FOR A NON-BIASED ALGORITHM

I. THE DATA THAT ONE USES NEEDS TO REPRESENT "WHAT SHOULD BE" AND NOT "WHAT IS".

We have to proactively ensure that the data we use represents everyone equally and in a way that does not cause discrimination against a particular group of people.

2. SOME SORT OF DATA GOVERNANCE SHOULD BE MANDATED AND ENFORCED.

As both individuals and companies have some sort of social responsibility, we have an obligation to regulate our modeling processes to ensure that we are ethical in our practices.

3. MODEL EVALUATION SHOULD INCLUDE GENERALIZATION ACROSS GROUPS

Learning from the instances above, we should strive to ensure that metrics like the true accuracy and false positive rate are consistent when comparing different social groups, whether that be gender, ethnicity, or age.

SureStart

SOLUTION? ... DIVERSITY

We need inclusion of diversity efforts at the early stages of any process or project within the AI industry.

ETHICS EDUCATION

Due to the lack of exposure to other cultures and walks of lives, there might be a disconnect between the actual reality the developed systems are expected to operate, in and how the creators intend for it to be used.

YOU'VE A BIG RESPONSIBILITY

Be aware of the potential biases during the stages of dataset picking, dataset building, algorithm developing, and generalizing the trained model.

SureStart

ETHICAL PRINCIPLES

Decision-making on numerous aspects of our daily lives is being outsourced to ML algorithms and AI, motivated by speed and efficiency in the decision process.

ML code scripts are rarely scrutinised; interpretability is usually sacrificed in favour of usability and effectiveness.

Issues of **garbage-in-garbage-out** may be prone to emerge in contexts when external control is entirely removed. This issue may be further exacerbated by the offer of new services of **auto-ML**, where the entire algorithm development workflow is automatised and the residual human control practically removed.

Along with accuracy, other dimensions like fairness, accountability, and transparency are also getting more and more attention.

SureStart

WHY COMPANIES DO NOT FULLY DISCLOSE ALGORITHMIC DETAILS?

(i) leaking of privacy sensitive data into the open;

(ii) backfiring into an implicit invitation to game the system;

(iii) harm to company property rights with negative consequences on their competitiveness;

(iv) inherent opacity of algorithms, whose interpretability may be even hard for experts;

(v) hard to resolve conflicts between potential algorithm deficits in accuracy and individual rights to privacy and autonomy of decision.

source: *https://www.nature.com/articles/s41599-020-0501-9.pdf*

MACHINE-LEARNING ALGORITHMS IN CRIMINAL JUSTICE

The **COMPAS** algorithm, developed by the private company *Northpointe*, attributes a 2-year recidivism-risk score to arrested people. It also evaluates the risk of violent recidivism as a score.

Upon examining a pool of cases where a recidivism score was attributed to >18,000 criminal defendants; it was systematically overestimated for black people: the decile distribution of white defendants was skewed towards the lower end.

This analysis was debunked by the company, which, however, refused to disclose the full details of its proprietary code.

Is it more important to have a **fair treatment** across groups of individuals or within the same group?
Let us take the case of gender, where men are overrepresented in prison in comparison with women. As to account for this aspect, the algorithm may discount violent priors for men in order to reduce their recidivism-risk score. However, attaining this sort of algorithmic fairness would imply inequality of treatment across genders.

SureStart

MACHINE-LEARNING ALGORITHMS IN CRIMINAL JUSTICE CONTD.

Fairness could be further hampered by the combined use of the ML algorithms with others driving decisions on **neighbourhood police patrolling**.

These algorithms may be prone to drive further patrolling in poor neighbourhoods may result from a training bias as crimes occurring in public tend to be more frequently reported. One can easily understand how these algorithms may jointly produce a vicious cycle – more patrolling would lead to more arrests that would worsen the neighbourhood average recidivism-risk score, which would in turn trigger more patrolling. All this would result in exacerbated inequalities, likewise the case of credit scores.

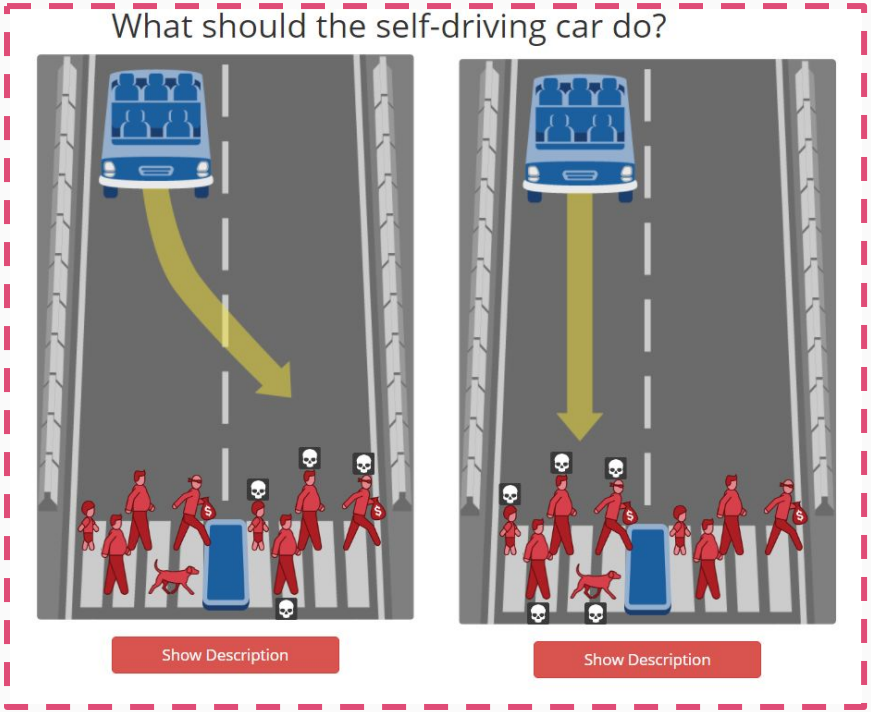
Fairness versus Accuracy!

Different classification accuracy (the fraction of observed outcomes in disagreement with the predictions) and forecasting accuracy (the fraction of predictions in disagreement with the observed outcomes) may exist across different classes of individuals (e.g., black or white defendants). Seeking equal rates of false positive and false negative across these two pools would imply a different forecasting error (and accuracy) given the different characteristics of the two different training pools available for the algorithm.

You should decide (or be instructed) to train your algorithm to attribute, e.g., a five/ten/twenty times higher weight for a false negative (re-offender, low recidivism-risk score) in comparison with a false positive (non re-offender, high recidivism-risk score).

SureStart

MACHINE-LEARNING ALGORITHMS IN AUTONOMOUS VEHICLES



try it yourself on https://www.moralmachine.net

In practice, the issue would be framed by the algorithm in terms of a *statistical trolley dilemma*, whereby the risk of harm for some road users will be increased.

- material damage against human harm?
- interest of the vehicle owner and passengers, or the collective interest of minimising the overall harm?
- preferring those who have invested more in their own and others' safety?
- low probability of a serious harm or higher probability of a mild harm?

The complexity of autonomous-vehicle algorithms was witnessed by the millions lines of code, so that the causality of the decisions made was practically impossible to scrutinise.

SureStart

DISCUSSION

1. One should not forget that the ML algorithms are learning by direct experience and they may still end up conflicting with the initial set of ethical rules around which they have been conceived. For you, what is that **one-specific vision** of the system being modelled?
2. The data on which an algorithm is trained on are **not an objective truth**, which factors should we consider when we are dependent upon the context in which the data has been produced?
3. Consider: Are the results from a particular model more sensitive to **changes in the model** and the methods used to estimate its parameters, or to **changes in the data**?
4. One should also not forget that points of friction across ethical dimensions may emerge, e.g., between **transparency and accountability**, or **accuracy and fairness** as highlighted in the case studies. Hence, the development process of the algorithm cannot be perfect in this setting, how can you negotiate and unavoidably work with imperfections and clumsiness?
5. Identifying a relevant pool of social actors may require an important effort in terms of stakeholders' mapping so as to assure a complete, but also effective, **governance** in terms of number of participants and simplicity of working procedures. What should we consider during developing these standards?

SureStart