Lesson 9_Fairness_in_Algorithmic_Decision_Making

Fairness in Algorithmic Decision Making

The combination of AI,with its recent developments, and Big Data enables **automated decision-making** even in domains that usually requires complex choices, which might be based on many factors and don't have predefined criteria or we do not know them before making the prediction.

In recent years, a wide debate has taken place both on *prospects* and *risks* of **algorithmic assessments** concerning particular individuals and groups.

The question is whether we should use these systems even though we are aware that may be unfair sometimes. Are they better than humans in assessing the individuals?

Some scholars have observed that in many domains automated decisions are not only **cheaper**, but they are also more **precise** and **impartial** than those made by humans: the machines are able to avoid basic issues related to the **human phsycology** (*e.g* overconfidence, loss aversion, anchoring, confermation biases, representativeness of heuristics), human prejudices (*e.g.* related to ethinicity, gender, social background etc.), but also possible mistakes that may derive from human inability to process certain statistical data.
In many assessments and decisions algorithmic systems have often **performed better** than human experts according to the usual standards.

Other scholars have a different opinion, as they underscored the possibility that algorithmic decisions may be either **mistaken** or **discriminatory** against both individuals and, even if it is less frequent, groups.

Why? Because only in rare cases the algorithms engage in **explicit unlawful** discrimination, which is known in the legal domain as **disparate treatment,** and basically consists in the outcome of the predictions being based on **prohibited features** of the candidates (race, gender, ethnicity).

More frequently it happens that the outcome of the predictions and decisions are discriminatory due to their **disparate impact,** as they affect *disproportionately* certain groups, without an ***acceptable rationale*** in the legal system eyes.

What are the main causes of discrimination?

Human Error Propagation

System based on *supervised learning* may be trained on **past** human judgements and therefore they may reproduce both strenghts and weaknesses of the humans who made the judgements in the first place, propagating their **propensities** to errors and prejudice.

For example, a recruitment system trained by a company on the past hiring decisions will learn to emulate the managers' assessment of the candidates suitability, rather than to directly predict an applicant's performance at work. Therefore, if past decisions were influenced by prejudice, the system will reproduce the same logic.

Training Set Composition

Another possible cause of discrimination may arise from the **composition** of the training set. Some biases may be backed in the training set and may persitst even if the predictors do not include a prohibited feature **explicitly**. This may happen whenever a **correlation** exists between discriminatory features and some predictors.

For instance, a human resource affected by a certain bias did not hire applicants from a certain *ethnic background* and that people with that background mostly live in a certain city, country or neightbourhood. A training set of decisions from that manager will teach the systems not to select people from those living places. This would entail continuing to reject applications from the discriminated group that shares the correlated features.

In other cases, a training set could be biased against a certain group because the achievement of the predicted outcome is approximated through a **proxy** that has a *disparate impact* on that group. This is strictly related to the

feature selection: what predictors we pick to be embedded into our system.

For example, let's assume that the **target variable** we want to predict is the future job performance of an employee and that it is measured *only* through the number of hours spent in the office. What will transpire is that women, who ususally work for fewer hours due to family burdens coping, are less successful than men; therefore, the women's perfomances will be predicted to be poorer.

Another possible scenario is that the mistakes and discriminations may pertain to the machine learning systems' biases embedded in the predictors. For example, a system may perform unfairly because it uses an unbalanced predictor as an input feature, which applies only to members of a certain group, or it may use biased human judgements as predictors (e.g., recommendation letters).

Again, unfairness may derive from an unbalanced dataset which **underrepresents** a certain group of people. This may generate prejudices towards it and reduce the accuracy of the predictions on that population sample.

It has been observed that is difficult to **challenge** the unfairness of automated decision-making because the objections raised by the individuals concerned may be disregarded or rejected because they interfere with the system's operation. This generates additional costs and uncertainties as it is hard to prove that the system is biased and that there's a fairness issue. Which is due to the fact that the system is based on **statistical correlations**, *against which may be difficult to argue* when the single individual is involved.

According to some, challenging the decision made by an algorithm will be very hard and the use of algorithms could be compared to the use of weapons.

Others hold a very different position, as they claim that algorithm decisions can simplify the examination decision process, making it easier to inspect: it would be easier to spot whether discrimination has occurred as opposed to the human decision process which is harder to inspect. Furthermore, the algorithms can also highlight central trade-offs among competing values, making them more transparent; this can be achieved by forcing a new level of specifity.

Another argumentation in favor of algorithms is that these systems are more **controllable** than human decision-makers: their faults can be identified with precision, they can be improved and engineered to prevent unfair outcomes. As we can imagine, the *principle of transparency* and all the issues related to **AI explainability** can be intertwined with this kind of issues and the possibility to identify causes of unfairness and improve AI decisions.

Should we exclude the use of automated decision-making?

The issues presented so far should not lead to the decision of excluding the use of automated decision-making, because the alternative is an imperfected system itself. Many papers report that human are affected by all kind of biases. Therefore, there is still the possibility that an automated decision system can be more fair than a human one. In many cases, the best solution is an **hybrid** system that integrates human and autonomous judgements. This can be achieved by enabling the affected individuals to request a **Human Review** of the system's outcome in case of a **detrimental decision**, as stated in the *General Data Protection Regulation*.

At the same time, we need to favour the transparency related to the system's explainability and develop technologies that enable human experts to analyze and review ADM, possibly changing them whenever there is not an acceptable rationale behind the decision.

AI systems have proved to successfully act in domains traditionally entrusted to the trained intuition and analysis of humans. The future will consist in finding the best combination between human and AI, considering both their capabilities and limitations.

The *principle of fairness* implies a commitment to ensure an equal and just distribution of **benefits** and **costs** and that individuals and groups are free from unfair **bias, discrimination** and **stigmatisation.** Depending on the application domain the meaning of fairness may slightly change with different nuances. In the AI decision-making domain the *substantive fairness* dimension, specified also by the General Data Protection Regulation, concerns the so called **informational fairness**, which is stricly connected to the transparency principle, because it requires individuals to be informed of the existance of an automated system and its purposes, the existance of profiling and the possible consequences. At the same time the *substantive fairness* also concerns the fairness of the content of a certain inference or decision and should avoid prejudices and discrimination under a combination of criteria:

- The use of **appropriate mathematical or statistical procedures** for the profiling procedures.

- The implementation of **technical and organisational measures** to ensure correctness of personal data. The data subject for example has the right to correct his personal information.
- **Secure personal data** from potential risks related to this kind of decisions and prevent possible discriminatory effects.

We can see the connection to the explainability of the system as well as to the procedural fairness, such as the possibility to **challenge** the decision made by the system.

The COMPAS system

The possibility algorithmitc unfairness raised many debates, in particular on the use of predictive systems in the justice domain. The COMPAS system is a very well known system in this area and is an actuarial risk assessment tool used by american judges to determine the **risk of recidivism**, which is the probability that an offeder will commit another offence in the near future; it's also used to assess the most adeguate correctional treatment. The system is based on statistical algorithms to entablish risk profiles associated to various groups of individuals that share certain characteristic. This risk is quantified into a **probability score**: offenders are classified into three categories : *high, medium, low* risk of recidivism. This score is based on a multiple choice test, that the subject is requested to do after he's been arrested for the first time, as well as *static (*prior criminal history, education, ...) and *dynamic* (drug abuse employment status ...) variables.

Possible discrimination issues of the COMPAS system emerged during a famous case: the **Loomis case,** who was charged of stealing a vehicle and fleeing from police. The Distrectual Court ordered an investigation which involved also the COMPAS risk assessment. Loomis got classified with *high risk of recidivism* and sentenced to 6 years of jail. The decision was appealed by Loomis, claiming that the system functioning was unknown, cannot be verified and that would also violate the principle of defence. Furthermore, it discriminates individuals and the statistical-based calculation would violate the right to obtain individualised decision. Loomis's arguments were all rejected. According to the Supreme Court:

- It is true that COMPAS uses statistics to generalize the prediction, looking for correlations between the characteristic subject and certain groups of similar individuals. The risk scores are used to predict the general likelihood that the subject will commit again the crime once he's released from custody. COMPAS does not predict the specific probality that an single offender will offend again. Furthermore, it should be used as an instrument to **enhance** judge's evaluation among other tools.
- The prohibition to base decisions solely on risks scores and the obligation to motivate the sentence would be sufficient to safeguard the defendant's rights.
- Gender discrimination was also rejected, as men and women have different characterestic of treament methods and grade of recidivism, so the differentiation is necessary to achieve better statistical accuracy.
- The race discrimination has been brought up on COMPAS, as it systematically gave higher risk grades to the blacks rather than the whites. Therefore, the judges must be informed of the issue.

Since the Loomis case the use of COMPAS has been widely debated, raising critisism both within and beyond the scientific community regarding its fairness and accuracy.

A study from ProPublica tried to evaluate COMPAS fairness and accuracy. To achieve this goal PP compared the predicted recidivism rates and the rate that actually occured over a 2-year period.

The results obtained highlighted a Moderate-Low accuracy of the system (61,2%). Black defendants were marked usually with a higher risk grade than they actually turned out to be, meanwhile white defendants were marked with a lower risk score.

▼ High risk misclassification : 45 % Blacks vs. 23 % whites.

▼ Low risk misclassification : 48% whites vs 28% blacks.

Other scholars, especially from Northpoint (that contributed in the development of COMPAS) claimed that ProPublica made several statistical and technical errors. In particular, it has been said that the system's accuracy should be compared to human judgement accuracy; in this regard they found out that COMPAS accuracy is higher than human accuracy. Moreover, they claim that the system is complaint to the principles of fairness and does not implement any

kind of racial discrimination, the differences in the recidivism rates are due to the difference between the "racial groups". Indeed, the probability that a particular individual would or not reoffend is equally correlated, for both blacks and whites, to the probability that such individuals would have actually recidivate.

Northpoint also found out that the percentage of correctly predicted high risk classified blacks is comparable with the one of whites.

Base rate: if we have some predictors used by the system whenever we have two distinct group in the population, they will have a different base rate. (e.g., in the health domain we know that women are more likely to develop a certain pathology than men).

Is COMPAS fair and accurate? A case Study!

We will use a case study to analyse the system's fairness. SAPMOC is a hypothetical system similar to COMPAS and keeps its essential aspects, but in a less complex framework.

Case study scenario

The two groups are **equally** represented in the training set, removing the problems related to underrepresented population groups. We assume SAPMOC to be much simpler than any other machine learning system and it takes in account a unique predictor which is strongly correlated to recidivism. The system will take into account the predictor whether or not the defendant committed previous criminal history. We can see that the population distribution is different in the two groups, meaning that we have a different *base rate*.

SAPMOC Assumptions

Real Outcomes			
	Recidivism	No Recidivism	Total
Previous Offence	750	250	1000
No Previous Offence	250	750	1000

SAPMOC Predictions			
	Recidivism	No Recidivism	Total
Previous Offence	1000	0	1000
No Previous Offence	0	1000	1000

Real data in the table above, Prediction in the one below.

Given that SAPMOC predictions are based on a unique predictor it will assign high risk to those who have a criminal history, and a low risk to those who don't.

Base Rate	Positives	Negatives
	(TP+FN)/(TP+FN+FP+TN)	(TN+FP)/(TP+FN+FP+TN)
Blue	62.5%	37.5%
Green	37.5%	62.5%

	Positives	True Positives	False Positives	Negatives	True Negatives	False Negatives
	(TP+FP)	(TP)	(FP)	(TN+FN)	(TN)	(FN)
Blu	750	562.5	187.5	250	187.5	62.5
Green	250	187.5	62.5	750	562.5	187.5

Positive: reoffenders; Negative: non reoffenders

Let us consider all the elements that are relevant for discrimination: the assumption that the individuals are splitted in two groups and having a different base rate, which indicates the proportion of those that reoffended and those who didn't out of the total number of individuals in each group.

Let's assume that we know both the real outcomes and the SAPMOC predictions, like in the ProPublica case. The relationship between these data is shown in the second table: this confusion matrix is obtained by comparing the system's predictions and the real outcomes.

Accuracy	
(TP+TN)/(TP+FP+TN+FN)	
Blue	75,0%
Green	75,0%

SAPMOC Accuracy; it is the same as in the COMPAS case

To evaluate SAPMOC fairness, we will use the criterion previously explained.

- ***Statistical Parity***
  Each group should have an **equal proportion** of negative and positive predictions. The idea is that the probability to be classified one way or another should be the same for individuals that belong to each group. SAPMOC *does not* comply to this criteria. However, this difference is stricly connected to the different base rate within the two groups. To comply to this criteria we need to equalize the predictions within the two groups (more blue as negatives, more green as positives), but this would lower the system's accuracy and we will introduce a *discriminatory treatment*, which cannot be justified on the basis of the feature of our individuals because we would treat differently individual that share to the same features.
- ***Equality of Opportunity***
  This criteria is also known as *conditional procedure accuracy equality,* according to which individuals that have the same features should be classified in the same way. In this example, those who share the same background (i.e., criminal record) should be treated equally in equal proportion.

  In the **Blue** group those who have a previous criminal history have a higher probability to be correctly classified if we compare them to the ones in the **Green** group. This classification is unfair mostly to the Blues, as they have a higher probability to be considered wrongly as positives, which is an unfavourable prediction in this case. This generates an **adversial** treatment towards them. Green individuals, on the other hand, are more likely to receive a negative (softer) label. This difference is due to the different base rate of the two groups.

Equality of opportunity	Positives	Negatives
	TP/(TP+FN)	TN/(TN+FP)
Blu	90,0%	50,0%
Green	50,0%	90,0%

In this table we can see that also this criterion is not respected.

- ***Calibration***
  The proportion of correct predictions should be equal within each group and with regard to each class. This means

that the proportion between TP and P predictions ($\frac{TP}{P}$) should be the same for the two groups, the same hold for the negatives.

The predictor leads to a consistent results in the two groups.

Calibration	Positives	Negatives
	TP/(TP+FP)	TN/(TN+FN)
Blu	75,0%	75,0%
Green	75,0%	75,0%

Criterion Satisfied!

- ***Conditional Use Error*** *(*also known as *False Rate)*
  Is the other side of *calibration.* To satisfy this criterion the proportion between FP (FN) and the total amount of positives (negatives) prediction should be equal for the 2 groups.

False rate	Positives	Negatives
	FP/(TP+FP)	FN/(TN+FN)
Blu	25,0%	25,0%
Green	25,0%	25,0%

Criterion Satisfied

- ***Treatment Equality***
  The ratio between errors in positive and negative predictions should be equal in all groups.

  This criterion ensures that no group will be **favoured** by the system errors. In our case this the ratio of the Blues misclassification leads to clearly **unfavourable** predictions, almost 10 times higher than the Green group. This aspect raised most criticism to the COMPAS functioning.

Treatment Equality	Positives	Negatives
	FP/FN	FN/FP
Blu	300,0%	33,3%
Green	33,3%	300,0%

Criterion not satisfied

Conclusions

As we can see in the case study, the different base rate of the two groups explains the violation of the above cited criteria. Even though this criteria are violated, it does't mean a **real unfairness**. If we try to **impose** one criteria, e.g. *statistical parity,* it would generate more unfair outcomes. Looking at the satisfied *calibration* criteria *treatment equality* and taking into account the different base rates, we may say that the system is fair to individuals as they are **equally treated** within the same group, because given certain features they're equally treated, but also that we have a **disparate treatment** of the two groups.

To improve the system performances, we may see the decision making process as a complex one, composed of more steps.

The prediction is NOT the decision. Making a decision requires to apply judgement on the prediction and then acting on the judgement outcome.

Consideration on the Fairness in automated decision making

- Unpacking the decision
  - Unfairness in prediction (prohibited features, biased data set, biased proxy, etc.)
  - Unfairness in classification (threshold – affirmative actions)
  - Unfairness in decision (right/values optimization)
- Predictive systems as instruments to understand the reality

To remedy to this issues, it may be possible to adopt the so called *affirmative actions,* namely to adopt a set of policies and practices seeking to increase the representation of a particular group based on their features (e.g., *gender, race, sexuality etc.)* in areas in which they are usually underrepresented. If we consider a system similar to SAPMOC in a different domain, for hiring for example, and we have different base rates. Through *affirmative actions,* we can consider to ensure a percentage of positions to a certain group (e.g., x postions reserved to women), meanwhile we do not change the system predictions.
Another possibility equivalent to the previous one: we can think of different threshold for different groups. Let's consider a system where we have multiple predictors, each one with its own weight, if we change the threshold for one group, we will change the number of positive predictions for that group.

Changing the predictions of the system might not be a wise idea, because these systems are a way to better understand our reality and have a more precise picture of what is going on and the to highlight possible different base rates in the society.

Looking to the future

- AI is too often perceived as a source of threats and Law is too often seen as difficult and sometimes even inaccessible for citizens
- The combination of AI and Law could be the key to protect citizens and make the Law accessible to the wider public

The **fairness standards** must be always considered in the domain of application we are in, because their relevance depends on that.