(Algorithmic) Fairness

AI Culture Lab - Waag
Stefano Bocconi

v.1

waag
technology & society

introduction

Nowadays algorithms are increasingly being used to support and sometimes even execute tasks traditionally performed by humans, in an attempt to automate human labour and improve (cost) efficiency. In some cases algorithms are being relied upon to make important decisions that impact our lives, such as obtaining a mortgage or qualifying for a job. As such, it is important that these algorithms operate according to the norms and values that are required in such processes.

This does not only apply to algorithms: the same scrutiny needs to be applied to processes that are performed by humans. While we have a critical tradition to consider the possible presence of injustice in human-driven processes, and sometimes their impenetrability (Kafka's The Trial), automatic processes might evoke an image of exactness and neutrality, due to the fact that algorithms are based on logic and deterministic rules. As a consequence, algorithms enjoy an aura of impartiality and accuracy that might lead to a lack of questioning of their functioning in terms of their outcomes and effects in society.

In the following we particularly look into a property that systems should possess: fairness. We do so by presenting an overview of recent definitions of fairness and showing how different definitions imply and embody different (and sometimes non-compatible) criteria and values. We also discuss a recent high-profile case where fairness (or the lack thereof) claims have been made against a system in current use. We conclude with some considerations on how to look at fairness.

fairness

Fairness can be examined by looking at its counterpart, which in literature is called (harmful) bias[1]. With bias we mean an inclination or prejudice for or against one person or group, in a way considered to be damaging to that person or group. In this sense, a fair system does not exhibit harmful bias.

Everybody has an intuitive understanding of what harmful bias means, and that it is something often undesirable from the perspective of fair treatment of individuals by

[1] Following literature on fairness, we use the term bias only in the context of harmful bias, and not in the more general interpretation that culture necessarily embeds bias.

systems. In the technical world the term has a specific narrow meaning, which relates to the predictions that a model generates: a model has bias if its numerical predictions are on average different from the data to be predicted. This definition is not usable for the goal of assessing the fairness of the system, as it does not take into account how these predictions affect different individuals and groups. Moreover, a technically unbiased system can exhibit unfair behaviour.

The problem is made urgent considering that at the moment the majority of the algorithms are designed by software engineers that often lack the knowledge for implementing a fair system. They are not in a position to assess whether the introduction of the algorithm will reduce or reinforce inequity in the system's outcomes.

Ideally engineers would use ready-made models and guidelines that can be implemented to guarantee algorithms behave in a fair way. Unfortunately, the debate over ethics in algorithm is not at a stage where easy answers can be obtained to be modeled and used in software[2].

This is not strange, as this discussion is essentially about values, and values (and, most of all, their importance relative to each other) are a product of a particular society and culture and usually pose controversial dilemmas. Therefore there is no unique and concrete technical definition of “fairness” yet, and such a definition would be in any case dependent on the context where the model is applied.

On the positive side, this debate is spreading also in the technical community, which creates an awareness of the complexity of the problem and of the need to discuss with ever involved parties on how to avoid unintended discriminations.

defining fairness

To give an idea of the problem, we can look at different ways of defining fairness, using as an example a classifier, i.e. a system that using as input some characteristics x of a person, produces a binary (1 or 0) result y, such as to hire / not to hire, give loan / do not give loan, recidive / non recidives.

---

[2] Naturally, one might ask whether this would even be possible from a theoretical standpoint, or whether a system should always be judged with regards to the fairness of its effects. In any case, best practices and guidelines might help.

The characteristics x are usually considered composed of two parts: protected (such as race, gender) p and unprotected (such as skills) z.

In a formula:

$$y = f(z,p)$$

Which means that the outcome of the classifier is a function of protected and unprotected characteristics of an individual.

In the following we introduce different criteria for fairness.

Individual fairness

If we believe that similar individuals (e.g. with similar skills) should be treated (in this case by the classifier) similarly, we would require that for every couple of similar individuals (with respect to their skills z and z') the classifier would produce a similar outcome:

$$z \approx z' \rightarrow f(z) \approx f(z')$$

Or in other words, individuals of similar skills should be classified similarly. This is called individual fairness. This measure can sound logical, but it has implications for our concept of fairness. For example it would also imply that we consider policies such as affirmative action as unfair, since they do not treat people of equal skills equally, but try to favour particular protected groups. Another clear limitation of this approach is that we need to measure the similarity between individuals, which is not trivial.

demographic parity

Moving from individuals to groups, if we believe that belonging to a group (such as being a woman, or hispanic) should not have any influence on the classifier outcome, we would want that:

$$Pr[\, f(x) = 1 | p = 1] \approx Pr[\, f(x) = 1 | p = 0]$$

Which means that the probability of a positive outcome of the classifier for people belonging to a particular group (p=1) should be (on average) similar to the probability of a positive outcome for people not belonging to that group (p=0).

This is called demographic parity. Again, this implies that if particular groups have on average a higher or lower level of skills, the outcome would be forced to even out the possibilities with a suboptimal results of the selection, with possible societal consequences, for example for jobs that have great responsibilities such as surgeons.

equalised error rates

Relaxing the requirements that different groups should be classified on average equal, we can allow different probabilities of outcome for different groups, but still demand that individuals are treated fairly across groups. We consider now the errors each prediction system makes. These errors are practically always present, since no system claims to be right 100% of the times, and prediction systems that are right more than e.g. 95% of the times are considered to be very good. But what about the 5% of the times they are wrong? Are they wrong in a fair way with respect to different groups of people?

The following criteria look at a system's fairness considering the errors that it makes. Therefore, if a prediction system would be always right, according to this approach it would be always fair.

To illustrate this approach with an example, we consider the famous COMPAS[3] case.

For the sake of this discussion, we need to imagine that we have the results given by COMPAS for several individuals, and can also see in the future whether those individuals have recidivated or not (assuming they were all released from jail, which means that COMPAS was consulted but its results were not used).

In this ideal situation we would be able to distinguish 4 cases: people that were classified as high risk and recidivate (True Positive), people that were classified as high risk and do

[3] COMPAS is a commercial software used by judges in Broward County, Florida, that helps to decide whether a person charged with a crime should be released from jail before their trial, or they should not based on the risk that they will recidivate within two years.

not recidivate (False Positive), people that were classified as low risk and do not recidivate (True Negative), people that were classified as low risk and recidivate (False Negative).

A measure of fairness could then be that the percentage of people rightly assessed as high risk is the same across groups. For example, 80% of black defendants assessed as high-risk and 80% of white defendants assessed as high-risk indeed recidivate. This would be expressed as:

$$[TP/(TP+FP) \mid p = black] \approx [TP/(TP+FP) \mid p = white]$$

This is called (positive) predictive parity. COMPAS would make an error in 20% of the cases, meaning assessing as high-risk a person that does not recidivate, for both white and black defendants. Both groups would be treated fairly, since the categorisation errors would not be bigger for a particular group.

Again, this seems reasonable when we look at people that recidivate. But imagine that I am a black defendant, and I am not going to recidivate. I would want that the probability I am incorrectly assessed as high risk would be firstly as low as possible, and secondly not higher than if I would be white. So among all the non recidivating population, the error rate of incorrectly assigning a high risk label should be similar across different groups. In a formula:

$$[FP/(FP+TN) \mid p = black] \approx [FP/(FP+TN) \mid p = white]$$

This is called false positive rate. And conversely, a recivating individual should not have more chances to be assessed as low risk if they are white, with respect to if they are black.

$$[FN/(FN+TP) \mid p = black] \approx [FN/(FN+TP) \mid p = white]$$

This is called false negative rate.

Each of these three requirements seems to capture an aspect of fairness. We would therefore want that classifiers satisfy all of them. Unfortunately it has been mathematically proven that if the groups do not exhibit the same characteristic (in this case recidivism) with the same rate (i.e. each group has a 40% of recidivating individuals), it is impossible

to have more than two of them satisfied. This means that a choice has to be made, considering the specific case where fairness has to be verified and the societal consequences of those choices.

more elusive bias

Up until here we discussed negative consequences of bias in taking decisions that have immediate effects on people. In such cases there is a system that performs a defined task, which produce a clear outcome, and such outcome damage particular people. This is what Kate Crawford calls allocative harm.

She also discussed a more subtle class, which she calls representational harms. In this case the system reinforces the subordination of a group, for example using stereotyping or cultural denigration. This form of bias causes social and indeed structural harms and has a more diffuse and long-term effect than allocative harm.

An example of this second class can be seen when performing a web search for CEO images, which yields a large majority of white men, and therefore under-represent other groups. Another example is when performing an automatic translation on Google translate from English to Turkish, of the sentences "he is a nurse, she is a doctor", and use the results to translate back to English. Because Turkish does not have gender in the pronouns, the final translation reads "she is a nurse, he is a doctor", therefore reinforcing existing stereotypes.

Representational harms are more difficult to quantify and pinpoint to a single cause, but are as if not more impacting, especially in the long term, than bias in decision-supporting software systems.

conclusions

The discussion on fairness in automatic system is steadily growing in the last years. This has led to the realisation that there can be many ways to define fairness, each with a particular underlying assumption of what it means to be fair. These definitions provide a means to verify whether a system is fair, according to a particular interpretation of the word. Therefore, there must always be a discussion on what it means to be fair in that

particular context, and the outcome of such a discussion determines which definition should be applied.

Automatic decision systems can and should therefore undergo a fairness "check-up" before being relied upon, which implies also that their implementation should not happen in a secretive way that makes scrutiny impossible.

On the other hand, automatic decision systems are not the only cause of discrimination in modern society. More diffuse, subtle and less measurable effects such as stereotyping are also playing a role. A constant critical (self)reflection is therefore necessary and need to constantly be performed if we want to be a fair society.

References

Tutorial: 21 fairness definitions and their politics

Ethics in Machine Learning Interview with Dr. Hanie Sedghi, Research Scientist, Google Brain

Kate Crawford - The Trouble with Bias

Kay et al: Unequal Representation and Gender Stereotypes in Image Search Results for Occupations

Bias detectives: the researchers striving to make algorithms fair

Math Can't Solve Everything: Questions We Need To Be Asking Before Deciding an Algorithm is the Answer

Courts use algorithms to help determine sentencing, but random people get the same results

Can an Algorithm Tell When Kids Are in Danger?

Ricci v. DeStefano