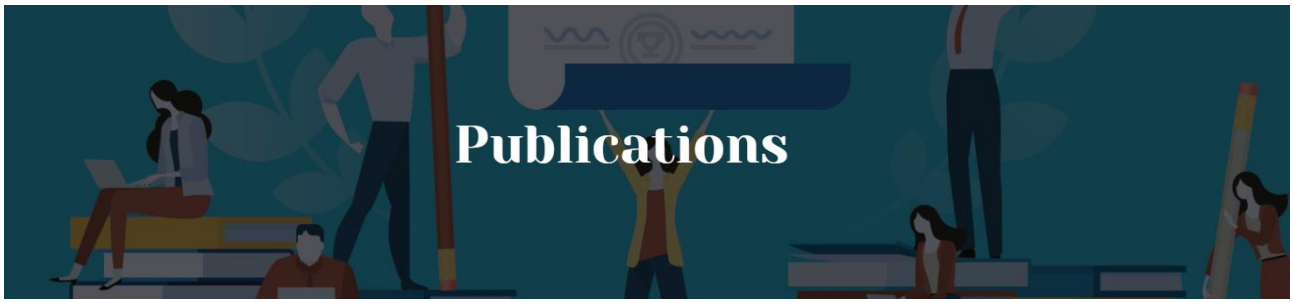


January 20, 2021

THE APPARENT BLINDNESS OF ARTIFICIAL INTELLIGENCE TOWARDS RACIAL PREJUDICES

By Cristina Di Stazio

In the collective imagination, the artificial intelligence systems ("A.I.") are perceived as perfectly logical and objective, far from the influence of human prejudices which the individual is often slave to.

Systems of A.I. have been touted as tools without these kinds of “human inefficiencies”.

Surprisingly, numerous studies have shown that actually these systems exhibit prejudices against racial differences.
The solutions of artificial intelligence are subject to distortions: “Bias”. The meaning of the term bias, in the landscape of the racial prejudices exercised by the operations of A.I., consists in the lack of "equity" that emerges from the output data of a computer system.

The biases derive from an objective programming system which is not affected by "inequitable criteria", but it is based on the processing of partial and discriminatory data entered at source in systems of A.I.

In particular, the system of A.I. taken into consideration for the purposes of the treatment of this analysis are the systems of machine learning.

What is a machine learning system?

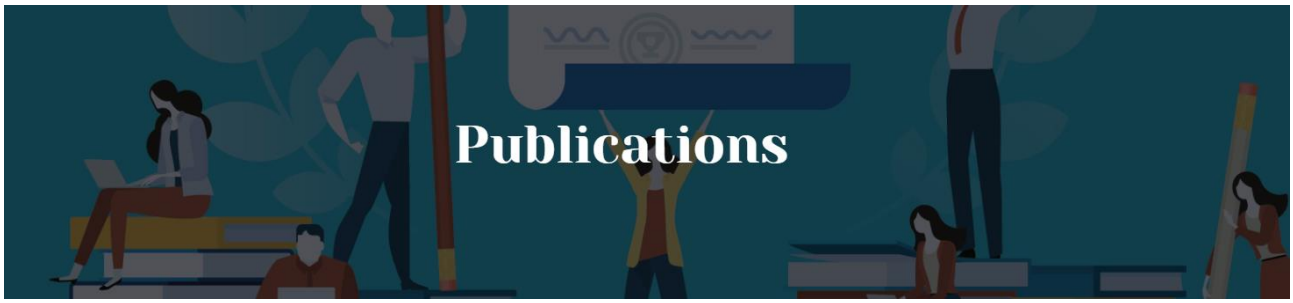
The first step in creating a machine learning system is data acquisition.
Input data (training data) is used to train the system. Following the acquisition and analysis of the training data, the system develops correlation models of the data sets provided, which will subsequently be incorporated.

Working on the data learned through its “training process”, the system then teaches itself the resolution criteria which defines the problem submitted to its attention. The algorithm is constantly updated when exposed to a set of data.

The system modifies itself based on the newly-acquired knowledge and uses factors on which it was not necessarily planned to rely on, also applying self-learning to recognize additional patterns that it will incorporate.

In this way, machine learning algorithms create a black box.

The criteria used in the decision-making process can be completely unrelated to the programmer.



"The unknown factors" are the result of the human cognitive abilities which are unable to cope with the continuous process of system editing.

In conclusion, what is provided to the machine learning system are data training, from which the system will extract a statistical model the parameters of which will be used in real-life scenarios for solving problems of various kinds.
Therefore, the output data produced by machine learning systems are simply a reflection of the training data to which they are exposed.

In the light of the above-mentioned considerations, it can be said that if training data (used for training and subsequent updating of the machine learning system) are influenced by human prejudices, consequently, the output data produced in the future will inevitably be subject to biases.

Predictive policing

Digital crime prediction has become the flagship of criminal research in police districts across the United States of America.
With the rise of big data and the ‘Internet of Things’, there has been a growing implementation of an “innovative police force” with a work based on a digital strategy composed of automatic learning systems on which predictive police algorithms rely on.

Predictive analysis sounds like an unbiased futuristic solution to the old problem of crime.

Actually, the certainty of being devoid of human prejudices or inefficiencies is only apparent.

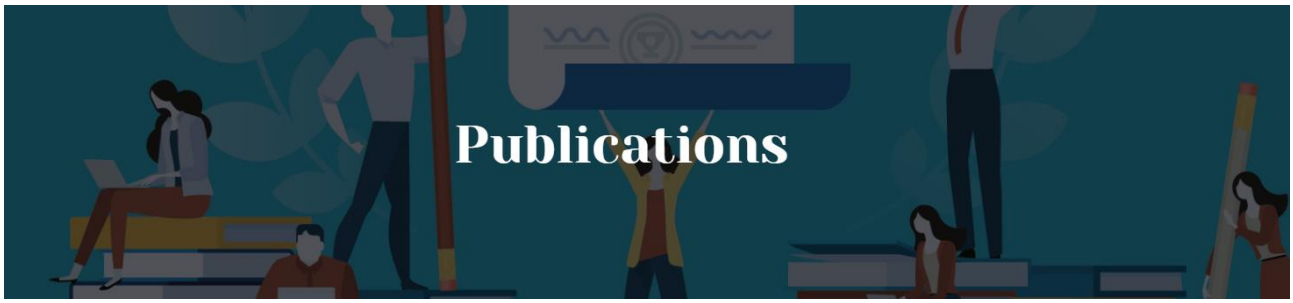
While predictive police and machine learning techniques are gradually evolving, the accusations of algorithmic bias still persist.

The main concern of researchers is that predictive police algorithms based on machine learning can strengthen racist police models in the United States (under the apparent neutrality).

My considerations are as follows

Programmers provide the predictive police algorithms with machine learning systems, a type of artificial intelligence that allows algorithms to identify factors through which they can indicate areas prone to future criminal activities.
The predictive policing based on machine learning starts with an algorithm file.
A machine learning algorithm implements its "training” based on the input data provided by the police forces.
In the context of predictive policing, this means that, during the training phase based on the data received, the algorithm predicts when and where the crime could occur later.

When the algorithm produces this response, “its neural connections and the variables that produced the response strengthen themselves” until the computer teaches itself the characteristics that define "the problem brought to its attention - crime forecasting”.



With each use, the algorithms automatically adapt to incorporate patterns perceived through machine learning systems and become more efficient. In this way, machine learning creates the enigma of the "black box"(previously described) in which the algorithm learns and incorporates new models in its code.

In light of the above-described considerations, it appears that predictive police algorithms are based on historical crime data provided by districts.
The police, therefore, is not only an end-user of algorithmic outputs but, above all, a provider of the information that those same algorithms use.

Historical data on crime are often intertwined with racial biases.
Data on criminal contacts, for example, show that American blacks and Hispanics are more likely to have contacts with the U.S.A. police force.

New York stop-question-and-frisk policy is a relevant example of the disproportionate contacts the black and Hispanic people have with the police forces.

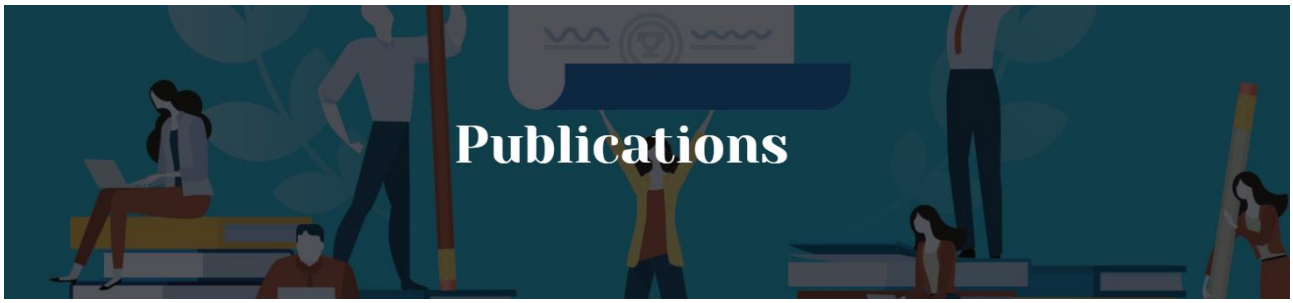
Between 2004 and 2012, the New York City Police Department earned about 4.4 million stops, out of which over 80% involved people of different races. More precisely, 52% of these 4.4 million stops involved American blacks and 31% involved American Hispanics.

Numerous studies, therefore, argue that the human nature of input data combined with the nature of machine learning algorithms lead to continuous biases.
The identification of areas subject to future criminal activities is possible through a series of machine learning systems, one of which is Nearest Neighbors (KNN).

Nearest Neighbors (KNN).

KNN algorithms incorporate new variables based on the "nearest neighbor" of the original variables that the programmer encoded to use the algorithm. The KNN algorithm autonomously learns which variables are closest to the original variables (the nearest neighbors of the original variables). The algorithm then incorporates the nearest neighbors of the variables into its code and relies on the new variables in its subsequent decision-making process.

A KNN algorithm is capable of discerning that race is a “near neighbor” of some of the variables that the programmer originally trained it to select. For example, if a KNN algorithm relies on historical crime data and is programmed to predict where a future crime will happen based on where it has already happened before, the algorithm could find that there is more crime in predominantly Black and Hispanic neighborhoods and thus determine that race is a “near neighbor” of the location variable that the human programmer originally programmed it to select for. A deep learning neural network does not rely on humans in order to be told what to look for in discerning criminality. Instead, based on the examples it is fed with, it makes associations and defines criminality on its own. So, a deep learning neural network that is trained on historical crime data could examine countless rap sheets and determine that race is a good predictor for crime in



the United States, because of the overrepresentation of people of color in the criminal justice system.

Crime is everywhere, but the police only find it where they are looking for it

The traditionalism of investigative strategies based on "human" decision-making mechanisms resulting from the intuitions/assumptions of individual police officers, has undergone a clear change over the years.

To date, the attempt to spread the protection of public safety through a process of "preventive localization of crime" and "futuristic detection" of areas at risk of crime is entrusted to digital survey programs equipped with machine learning systems.

Having obtained the statistical forecast regarding the places of future criminal perpetration (through the elaboration of the historical data present in the database, described above), the machine learning systems begin a process of designing territorial districts that are classified as dangerous (crime mapping), where they will send police patrols to carry out the supervision activity.

Many studies show that the mapping of criminal geography elaborated by machine learning systems leads to a social stigmatization of minority groups.

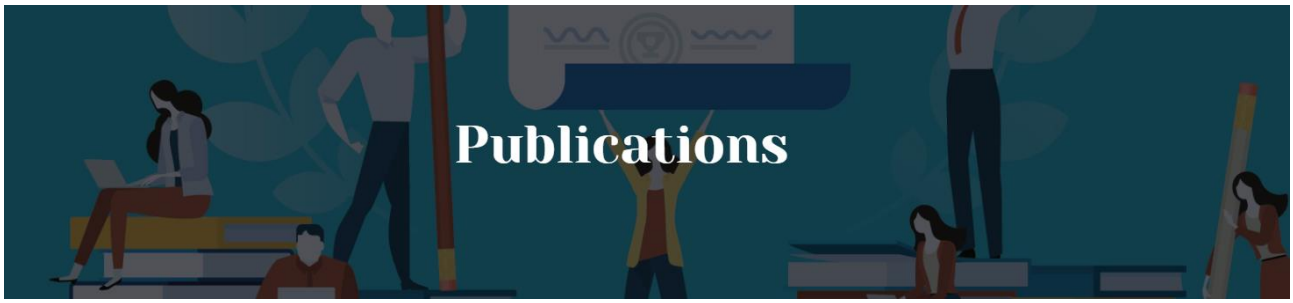
Recent researches on the most widely used preventive police systems in the United States, such as Predpol, Compstat and Hunchlab, show that the detection, provided by these systems, on the places of major perpetration of crime is statistically distorted.

The historical data used to train the algorithm are affected by a partial perception of the statistics regarding places of perpetration of crime. This is due, on the one hand, to an over-representation of data concerning crimes perpetrated in areas inhabited mainly by minorities (Blacks and Hispanics) and, on the other hand, to a sub-representation of the data concerning the crimes perpetrated in the areas inhabited by the majority groups (white).

The digital programming of machine learning systems, therefore, even if not explicitly codified to the identification of race as an analysis variable, can (based on the processing of historical data) continuously direct police patrols to the places where social minorities live.

The process by which predictive policing algorithms create feedback loops which reinforce deleterious and biased patterns happens over time. These algorithms are also known as “go with the winner".

This means that even if crime rates in two neighborhoods are remarkably similar, if region A (predominantly inhabited by white people) has a crime rate of 10% and region B (predominantly inhabited by black and Hispanic people) has a crime rate of 11%, the update process will settle on region B with a 100% probability.



When police are sent to a particular region repeatedly, they are more likely to see crime in that region. This predisposes the police to collect more crime from region B than from region A. This means that the algorithm will, over time, consistently learn to send police only to Black and Hispanic neighborhoods. Furthermore, the lack of observations about the under-policed region "prevents the system from learning" that the crime rates of two regions are actually very similar. Under this analysis, predictive policing algorithms will learn less about crime in predominantly white areas and will report that there is less of a risk of future crimes in those areas; instead, they will learn more about predominantly Black and Hispanic neighborhoods and indicating that more police personnel should be sent to those areas.

A notorious example is drug crime. Survey data suggest similar rates of use of illicit drugs among Blacks and Whites, although Black drug users are more likely to have criminal justice contact compared to White drug users. One explanation for this phenomenon is that police forces have prioritized enforcement actions on open-air drug markets, primarily used by African-Americans, rather than residential transactions of Whites.

All this mechanism generates a self-fulfilling prophecy. If policy-makers expect ex ante to find more crime among group A than group B, then it is possible that they will find this expectation validated ex post, but only because they have spent more time looking for crime among members of group A than among members of group B.

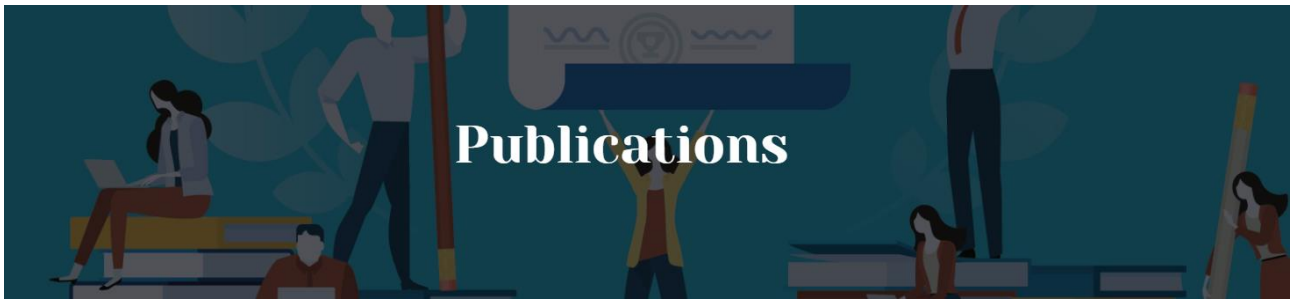
The Public Oversight of Surveillance Technology (POST Act)

Law enforcement, over the years, has been so tight-lipped about how it uses these technologies that it is very hard for anyone to assess how well they work. Even when information is available, it is hard to link any one system to any one outcome. And the few detailed studies that have been done focus on specific tools and draw conclusions that may not apply to other systems.

For these reasons, one of the most influential cities in the United States (New York) has decided to take a significant step in the matter of "digital crime prevention".

On June 18 2020, in fact, the New York City Council voted (44 against 6) to enact the Public Oversight of Surveillance Technology (POST Act). The bill will increase transparency and oversight over the New York police department's ("NYPD") use of sophisticated new surveillance technologies and information sharing networks, by requiring the NYPD to disclose basic information about the surveillance tools it uses and the safeguards in place to protect civil liberties.

The POST Act was actually introduced in March 2017, and its success builds upon three years of persistent advocacy from civil rights groups and community activists, including the Brennan Center, the Surveillance Technology Oversight Project (STOP), the New York Civil Liberties Union, CAIR New York, the National Lawyers Guild, and the New York Legal Aid Society, among others. In 2020, nationwide racial justice protests, motivated by ongoing police violence, spurred the overdue passage of the bill.



The bill requires that by January 12, 2021 (180 days after the POST Act came into effect), the NYPD must publish impact and use policies for all existing surveillance technologies describing how the technology will be used, the limitations in place to protect against abuse, and the oversight mechanisms governing the use of technology. Impact and use policies are also required at least 90 days before a new surveillance technology is used. Following the publication of each policy, the public has 45 days to provide inputs, which will be incorporated into a final impact and use policy brought before the Mayor and the City Council.

By requiring transparency and periods for public input, the POST Act takes critical first steps towards establishing community oversight over the use of surveillance technologies by the NYPD.

Bias mitigation strategies

Many bias mitigation strategies for machine learning have been proposed in recent years. The different approaches can be divided in the following three distinct groups.

Pre-processing

Efficient bias mitigation starts at the data acquisition and processing phase since the source of the data as well as the extraction methods can introduce unwanted bias.

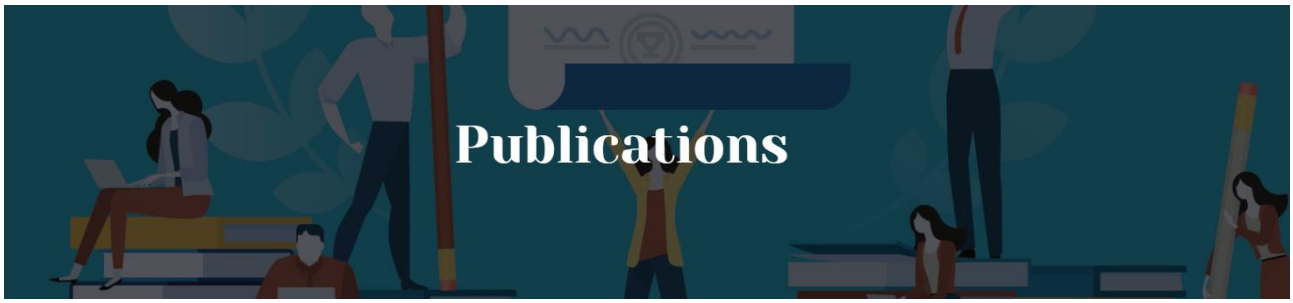
Pre-processing techniques try to transform the data so that the underlying discrimination is removed. If the algorithm is allowed to modify the training data, then pre-processing can be used.

Pre-processing algorithms are used to mitigate prevalent bias in the training data.

This can be achieved by suppressing the protected attributes, by changing class labels of the data set, and by reweighting or resampling the data. In some cases, it is also necessary to reconstruct omitted or censored data in order to ensure that the data sample is representative. There are plenty of imputation methods to achieve this objective.

The idea of many experts is to apply one of the following techniques for pre-processing the training data set and then to apply classification algorithms in order to acquire an appropriate classifier:

- **Reweighing**: reweighing is a data pre-processing technique that recommends generating weights for the training examples in each group/ label combination differently to ensure fairness before classification. The idea is to apply appropriate weights to different tuples in the training dataset to make the training dataset discrimination free with respect to the sensitive attributes. Instead of reweighing, one could also apply techniques (non-discrimination constraints) such as suppression (remove sensitive attributes) or massaging the dataset — modify the labels (change the labels appropriately to remove discrimination from the training data).



- **Optimized pre-processing**: the indication is to learn a probabilistic transformation that edits the features and labels in the data with group fairness, individual distortion, and data fidelity constraints and objectives.

- **Learning fair representations**: the idea is to find a latent representation that encodes the data properly while obfuscating information about protected attributes.

- **Disparate impact remover**: feature values are appropriately edited to increase group fairness while preserving rank-ordering within groups.

In-processing

In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model either by incorporating changes into the objective function or by imposing a constraint.

The idea of many experts is to apply one of the following techniques:

- **Adversarial de-biasing**: a classifier model is learned to maximize prediction accuracy and simultaneously reduce an adversary's ability to determine the protected attribute from the predictions. This approach leads to a fair classifier as the predictions cannot carry any group discrimination information that the adversary can exploit.

- **Prejudice remover**: The idea is to add a discrimination-aware regularization term to the learning objective.

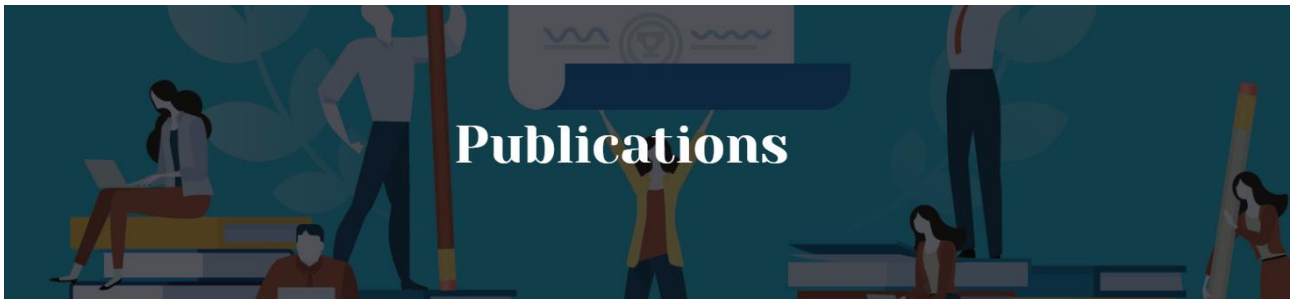
Post-processing

The final group of mitigation algorithms follows a post-processing approach. In this case, only the output of a trained classifier is modified.

Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used where the labels assigned by the black-box model initially get reassigned, based on a function during the post-processing phase.

The idea of many experts is to apply one of the following techniques:

- **Equalized odds post-processing**: The algorithm solves a linear program to find probabilities with which to change output labels to optimize equalized odds.



- **Calibrated equalized odds post-processing**: The algorithm optimizes over calibrated classifier score outputs to find probabilities with which to change output labels with an equalized odds objective.