RACIAL BIASES IN ARTIFICIAL INTELLIGENCE

Artificial Intelligence is, nowadays, already used on a large scale and getting step by step in our everyday life. It becomes influential in our society as they are helping us - humans, with their suggestions and predictions.

INTRODUCTION: THE CASE OF AMAZON REKOGNITION

We can take the example of Amazon Rekognition [1], a software able to identify and track people in real time thanks to facial recognition, done by means of a deep learning algorithm, subfield of AI. As a result, this software has been licensed by law enforcement agencies (in Oregon and Florida for instance) to identify possible suspects and maybe match them to a mugshot of a criminal. However, as a study from the MIT Media Lab pointed out, Rekognition does not perform well when identifying female and/or dark-skinned people, or at least is doing worse than when identifying white people. Moreover, a disturbing proof of the lack of accuracy from the software has emerged: in July 2018, the ACLU (American Civil Liberties Union) showed that among the 535 members of Congress in America, 28 of them were matched to known criminals, which disproportionately affected people of color [2].

Rekognition, if still performing that way, can lead to huge bad consequences for public safety, especially the one of people of color. This possible difference of treatment towards people who are not Caucasian, even without wanting the software to do so, is considered to have racial biases, which can have a deep impact on people's lives.

In this paper, we will first define the formulation of racial biases. Newt, we will focus on how they are incorporated in artificial intelligence algorithms. Finally, we will discuss on how and if we can get rid of them.

Lisa De Campos

I – HOW CAN WE DEFINE A RACIAL BIAS AND HOW MUCH IS IT HARD-WIRED INTO US?

First of all, we need to understand what is a "Racial Bias", and break down the composition of this expression, defining first the notion of "Racism", and then "Bias".

A – DEFINING 'RACISM' BEYOND DECOMPLEXIFIED DISCRIMINATION

Racism [3] is most of the time defined as the belief that a particular race is superior or inferior to another, that a person's social and moral traits are predetermined by their inborn biological characteristics. This common definition denotes a decomplexified form of discrimination with prejudices, in the "best" cases leading to considered funny jokes to hate crimes in the worst ones. However, this formulation feels like it is only exposing the tip of the iceberg called "Racism". Indeed, it is not only summarized to hate and / or despise for another race. That is why we also need to discuss about "Systemic Racism".

Systemic or institutional racism [4] is a term which was developed by sociologist J. Feagin, and is defined as how ideas of white superiority are captured in everyday thinking at a systems level. It considers how society -which also includes people of color- operates, rather than looking at one-on-one interactions. This concept became a popular way of explaining, within the social sciences, the significance of race and racism historically nowadays. It stipulates that racism is embedded in laws and institutions, giving an unfair amount of resources, rights and power to white people while denying it to people of color. Examples of this kind of racism can be noticed in access to sport, education, hiring, and even has roots in beauty standards (having which are considered "more" Caucasian physical traits, especially having light skin, is being more preferred and considered beautiful).

B – RACIAL BIAS' PRESENCE IN SOCIETY

On the other hand, we also have to define what is a bias. It refers to the action of supporting or opposing a particular person or thing in an unfair way, because of allowing personal opinions to influence judgments [5].

Adding both the notions of "racism" and "bias", we end up with the formulation of a "racial bias". It is a form of implicit bias, which refers to the attitudes or stereotypes which affects an individual's understanding, actions and decisions in an unconscious manner.

This racial bias seems to be undeniably present among people at least leaving in Western countries. This assumption can be confirmed when looking at the results of the Implicit Association Test (IAT) provided by Harvard [6], which has been performed on nearly 864,000 people between 2014 and 2015, and can still be taken nowadays. This test aims at helping people to be conscious about their deepest of thinking regarding their association of light and

dark-skinned people to the notion of good and bad. A summary of the saved results shows that 68% of web respondents have a slight to strong automatic preference for light skin compared to dark skin, the highest percentage being 28% of moderate preference for light-skinned people. Moreover, this summary states that only 19% of the we respondents do not have any racial bias at all.

As stipulated before, racial biases are not generated on purpose by the one who has them. As a result, the owners of these unconscious thoughts probably do not want to target people of color and add discomfort or suffering to their daily life, all of this due to their work. However, systems, such as some artificial intelligence algorithms, supposed to be neutral in their way of treating individuals and insensitive to the notion of racism, actually end up adding to people of color's life more discomfort, and discriminate them. As a result, we are now going to see how racial biases are incorporated in artificial intelligence algorithms.

II – HOW ARE RACIAL BIAS INCORPORATED IN ARTIFICIAL INTELLIGENCE?

A – LACK OF REPRESENTATION BIASING DATASETS

As feared, even algorithms that we can consider as well designed can be subject to exhibition of discriminating factors, especially racism, and so not on purpose.

First of all, bias in data used for a machine learning algorithm can occur because of lack of representation from each racial group. Indeed, the amount of effort an artificial intelligence puts into learning about a race is proportional to its frequency in the exploited dataset. This kind of bias in the data is said to be explicit.

An instance of the results of these kinds of skewed dataset is the one exploited by Nikon for one of their products [7]. One of their digital cameras, equipped with a blink detection feature, would not snap photos of many of their Asian users because of the software deducing their eyes were not open. Having a low frequency of Asian faces among their dataset is the reason of such discomfort.

B – MISTAKING CORRELATION WITH CAUSALITY

On another hand, even when the dataset seems fair and unbiased-looking, we can still deal with some issues, the main one being that the artificial intelligence do not have any idea of what its inputs mean in terms of real-world implication. As a consequence, the algorithm has to learn possible causalities on its own. However, it can turn out to be wrong, as it can mistake causality with correlation, the last one not implying the first notion. This second type of bias in the data is said to be implicit.

An article named *Dissecting racial bias in an algorithm used to manage the health of populations*, published in October 19th in the scientific magazine *Science* analyzed an algorithm widely used by the U.S health care system which affects 200 million patients in the United States [8, 9]. This prediction algorithm's aim is to identify and help patients who have complex health needs, focusing on the ones who will benefit the most from social dedicated care (dedicated nurses, extra primary care, appointment slots and so on). This algorithm also allows the optimization of resources -in this case money, allocation. However, the bias lies in this problematic prediction's consequence: healthier white people will actually receive the same treatment as the one dedicated to less healthy Blacks, as they are scored at similar risks of needing more care. As a matter of fact, at a same level of risk, black patients appeared to have 26.3% more chronic illness than the white one, and are considerably sicker than them.

Even though the patient's ethnicity is excluded in the data in order to avoid explicit bias, the source of bias in this example is more subtle than this, or the lack of representation in a dataset. To know where it comes from, we have to understand what are the features -the variables the algorithm relies on in order to give a prediction, and what is actually predicted by this artificial intelligence. Indeed, the dataset is composed of information such as the age, sex, insurance type, diagnosis and so on over the past year for each patient. These variables are accompanied by a numerical label expressing the total medical expenditures in the following year. As a result, if we give to this algorithm a new patient as a datum for prediction, it will return a value with the same metric as the labels, meaning the health care future cost for this patient. Knowing what the algorithm actually returns brings us closer to understanding why it is skewed. As a matter of fact, it does not predict the illness of the person and their health care needs, but it had built a causality between high future health care costs and illness. However, even if there exists a high correlation between these two notions, the causality does not hold here. The algorithm, not having an understanding of real-world implications, does not know that given a level of health, black people tend to generate lower costs than white people. As a result, due to the fact that the metric used for prediction is not the right one to use in order to have a causal relationship with the illness of the patient, the artificial intelligence ended up having a racial bias by accident. Consequences of such a bias applied on a huge population is deeply affecting people of color's life. Indeed, the article stipulated that if this algorithm did not have this implicit bias, the percentage of black patients receiving additional care needs would increase from 17.7 to 46.5% and improve their condition, at least for health care.

III – IS IT POSSIBLE TO GET RID OF RACIALY BIASED ARTIFICIAL INTELLIGENCE?

From the previous example, we could see that a biased artificial intelligence can lead to dramatic consequences, and might in some cases lead to death due to not benefiting from dedicated care, or to unjustified altercation with the police, if we reconsider Amazon's Rekognition software. Therefore, it is crucial to get rid of these bias in datasets used for artificial intelligence. For the last part of this essay, we will try to discuss how it is possible to do so.

A – BACKWARD CHAINING FOR BLACK BOXES IN ARTIFICIAL INTELLIGENCE

Nowadays, artificial intelligence is making automated decisions regarding important turn points in our life such as who will get a bank loan, who will get a job interview or into college, and as we discussed previously, who will benefit from further dedicated health care. All of these major industries are using artificial intelligence algorithms which are actually used as black boxes [10]. As a matter of fact, we can take for instance deep neural networks which are key components to many artificial intelligence applications, mimicking to some degree the way the human brain is structured (including layers composed of 'neurons' which are interconnected). Deep neural networks are commonly used as black boxes in a sense that their inputs and the operations these algorithms perform are not visible to their users, as they are largely self-directed in finding the data features which are correlated to a produced output. The problem with such algorithms is that they are generally difficult to interpret since they do not highlight the most important features which led to their output. As a consequence, getting to know the element which may have caused the artificial intelligence to be biased is fastidious, even for data scientists and their programmers: the operations and processes within the algorithms cannot easily be understood, nor viewed, leading to unnoticed errors.

What can be possibly done to identify such societal harms is to get to know how such biases appear. An interesting procedure to do so is called 'backward chaining' [11]. As a matter of fact, this logical process infers unknown truths from known conclusion by moving backward from the prediction found by the algorithm, which allows us to find the initial conditions and features that led to this particular output. An example of this kind of application can be imagined on these artificial intelligences used in the major industries that are supposed to help us in our daily life: we can consider an artificial job recruiter. Let's imagine that this algorithm is tackling with a classification problem, where some features about the job candidate are entered in the artificial intelligence, which would classify the candidate as someone who would benefit from a job interview or not. Among a set that is used to train the algorithm, we notice that the artificial job recruiter tends to allow strictly more job interviews to white people than people of color. This would be where the use of backward chaining is interesting, as we can

select the outputs which are the ones not allowing a job interview, and consider the features that lead to this conclusion. Therefore, it is easier for data scientists and programmers to find the presence of a racial bias or not in their algorithm.

B – CHANGING THE WAY WE WORK ON ARTIFICIAL INTELLIGENCE ALGORITHMS

However, even if adding the backward chaining process to try and identify the bias in artificial intelligence helps the data scientists in providing nonracially skewed algorithms, it does not mean that they are able to identify all of them. As a matter of fact, data scientists are not aware that some features leading to biased conclusions are actually socially harmful, as they do not have a sociologist's knowledge. Therefore, it is not sufficient to only add a technical procedure to the development of the algorithm, and it is essential to also change the way and who data scientists work with.

According to artificial intelligence reporter at MIT Technology Review and data-scientist Karen Hao [12], a feedback loop is nowadays widely held when an artificial intelligence is deployed. This loop has two sides: the technical experts and the social experts. The technical experts engineer an artificial intelligence, which is being deployed. This product happens to have some impact on society, in our case racial harms due to the algorithm's racial bias. On the other side of the loop, there are the social experts who are evaluating the impact that has this artificial intelligence and raise concerns about it, in our case the racial bias. These concerns lead to debates which are going on until a fix is proposed to this specific problem, that is a phase of regulation. Afterwards, the feedbacks are getting back to the technical experts, who will modify their artificial intelligence to improve it with respect to the raised issue. However, the possible debates about the algorithm should not be done after it is deployed, but while it is being engineered. To Karen Hao, the engineers should not wait for the impacts of an artificial intelligence, but should anticipate it. As a result, sociologists, anthropologists, philosophers and technical experts should not be on the opposite sides of the loops, but together, at the very beginning, as social problems are becoming technical problems, and vice versa.

If we reconsider the example of the artificial job recruiter, where only data scientists were working on the algorithm, we can rethink of this as a work of both technical and social workers. As a matter of fact, the results provided by the backward chaining process that were not understood in terms of bias by the data scientists can be with the help of social workers. Further than this, the sociologists working on this project can also prevent the use of a specific feature ending up with racial bias thanks to reports made on systemic racism and its impact on job recruitment. Working this way would prevent the data scientists from reviewing several times their algorithm, and also would prevent user's indignation towards the deployed artificial intelligence.

C – POLITICAL FRAMEWORKS FOR AVOIDING RACIAL BIAS?

In order to further reduce or get rid of the existence of racial biases in artificial intelligence, another consideration can be to have a defined environment for the project to be deployed. As discussed before, racially skewed artificial intelligences, if deployed and used on the population, can lead to dramatic consequences for people of color, as it maintains and takes part in systemic racism. Therefore, having a political framework for artificial intelligence algorithms and the datasets used could help regulating the presence of racial bias. For instance, a condition regarding the datasets can be to have an equal amount of racial representation in them, or to not use the skin color or ethnicity as a feature. These kinds of political environment dedicated to artificial intelligence ethics can be thought of specific to the company, the country, the continent or all of them combined in the best case.

In April 2019, a guide to artificial intelligence's ethics made by the European Commission with 52 experts was released to guide artificial intelligence projects based on the European soil [13]. As a matter of fact, this guide consisting of seven key requirements is provided on their website in order to call an artificial intelligence trustworthy, one of them being about the diversity and non-discrimination fairness. This key requirement stipulates that the artificial intelligence systems should be accessible to all, regardless of any disability.

However, the composition of the 52 experts that worked on this guide from the European Commission seemed to be unbalanced, as the majority of them were industrial stakeholders, and the near absence of sociologists, ethicists, philosophers or anthropologists could be noticed. We are ending up with the same issue that was stipulated in the previous point: technical and social workers are not working together, especially here when they are even more needed when talking about ethics [14].

On the other hand, even if such guides and political frameworks are applied to artificial intelligence projects, they are applied on artificial intelligence companies which have their headquarters based in this area (in the case of the European Commission's guide, Europe). However, these companies can have their algorithms exploited in this specific area, but can be based in another location, were these requirements are not necessary to be told as trustworthy. For instance, we can imagine an American artificial intelligence company which has deployed its artificial job recruiter for a French service, this algorithm asking for ethnicity as a feature to predict the possibility to give a job interview or not. Regarding what is related to the content of a dataset to train the machine learning algorithm, America is allowed to use data about ethnicity, whereas it is forbidden in France. As a result, a guide to ethics can experiment a loophole and a contradiction with another political environment, as it can be avoided with the company's basement.

D – RAISE AWARENESS THROUGH DATA SCIENTISTS' SKILLS TRAINING AND RACIAL DIVERSITY

As a consequence, is there no way to have a full consensus on ethical frameworks related to every company, state or continent regarding racial bias? Probably not, as each stakeholder may have their own economic or political interests or liberties that does not constraint them in the dataset used. Yet, requiring skills training about ethic to data scientists is worth considering. Nowadays, students in data science often follow courses about artificial intelligence or computer science ethics, and giving them a baseline in this domain to raise awareness about the consequence of what they will, when finally becoming data scientists, produce.

Another issue to consider is not only the possible lack of racial representation in artificial intelligence's datasets, but also in the project coworkers who work on them. Indeed, these teams are composed of mainly white men, where cultural diversity is important in a modern workplace. Hiring engineers coming from different cultural, economic and sociopolitical backgrounds helps bringing diverse viewpoints and perspectives in these kinds of project, and could also help pointing out the racial biases we want to get rid of.

CONCLUSION

Artificial intelligences are not racist, or at least these algorithms are not intentionally racist as they do not have any political agenda, and they reflect what we actually teach them. These technologies, which are increasingly significant in our daily life, actually reflect the presence of racism within our society, as a system. Moreover, despite the fact that the presence of racial bias can sometimes be fixed thanks to some engineering technologies or the diversity in collaboration for a project, it is probably not possible to fully get rid of all of them. Racial biases taught in artificial intelligence are going beyond the engineering problem, and are becoming with extension social issues, which are even less easy to solve.

However, we do not have to be pessimistic about artificial intelligence and never touch it ever again, and it may be important to have an algorithm, supposed to be neutral, exposing to our faces the errors humans are committing.

REFERENCES

- INTRODUCTION

[1]: *'Gender and racial bias found in Amazon's facial recognition technology (again)',* THE VERGE, updated on January 25, 2019. [Website]. Available on: https://www.theverge.com/2019/1/25/18197137/amazon-rekognition-facial-recognition-bias-race-gender

[2]: *'Amazon's Rekognition software lets cops track faces: Here's what you need to know',* CNET, updated on March 19, 2019. [Website]. Available on: https://www.cnet.com/news/what-is-amazon-rekognition-facial-recognition-software/

- I - A

[3]: *'Racism',* ADL. [Website]. Available on : https://www.adl.org/racism

[4]: *'Definition of Systemic Racism in Sociology',* ThoughtCo., updated on June 11, 2020. [Website]. Available on : https://www.thoughtco.com/systemic-racism-3026565

- I - B

[5]: *'Bias',* Cambridge Dictionary. [Website]. Available on : https://dictionary.cambridge.org/fr/dictionnaire/anglais/bias

[6]: *'Project Implicit',* Implicit Harvard. [Website]. Available on : https://implicit.harvard.edu/implicit/takeatest.html

- II - A

[7]: *'Are Face-Detection Cameras Racist?',* TIME, updated on January 22, 2010. [Website]. Available on: http://content.time.com/time/business/article/0,8599,1954643,00.html

- II - B

[8]: *'Millions of black people affected by racial bias in health-care algorithms',* Nature, updated on October 26, 2019. [Website]. Available on: https://www.nature.com/articles/d41586-019-03228-6

[9]: *'Dissecting racial bias in an algorithm used to manage the health of populations',* Science, updated on October 25, 2019. [Website]. Available on: https://science.sciencemag.org/content/366/6464/447

- III – A

[10]: *'The 'Black Box Problem of AI',* Data Driven Investor, updated on May 9, 2018. [Website]. Available on: https://medium.com/datadriveninvestor/the-black-box-problem-of-ai-33d261805435#:~:text=The%20%E2%80%98Black%20Box%E2%80%99%20Problem%20of%20AI.%20Artificial%20intelligence,inboxes%20from%20spam%3A%20they%20are%20our%20invisible%20workforce.

[11]: *'Backward Chaining in AI: Definition, Uses & Efficiency',* Study.com, updated on October 26, 2019. [Website]. Available on: https://study.com/academy/lesson/backward-chaining-in-ai-definition-uses-efficiency.html

- III - B

[12]: *'Why We Need To Democratise How We Build AI | Karen Hao | TEDxGateway',* TEDx Talks Youtube Channel, uploaded on June 9, 2020. [Website]. Available on: https://www.youtube.com/watch?v=D28aL_5LH2Q

- III - C

[13]: *'Ethics guidelines for trustworthy AI',* European Commission, updated on April 8, 2019. [Website]. Available on: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[14]: *'L'Intelligence artificielle peut-elle être éthique ?',* FigaroVox, updated on February 25, 2019. [Website]. Available on: https://www.lefigaro.fr/vox/societe/2019/02/21/31003-20190221ARTFIG00141-l-intelligence-artificielle-peut-elle-etre-ethique.php