Machine Learning and Artificial Intelligence Assignment 1

Question 1. Does the COMPAS system discriminate against individuals. If so is this discrimination unfair. Why? If not why is it characterised by some as unfair?

NOTE this is not a trick question but it does require you to think about the whether and how COMPAS represents individuals. And to define and defend a theory of fairness.

The discrimination of COMPAS

The COMPAS does discriminate against individuals, and it's unfair. According to a study by ProPublica, 2016, when COMPAS evaluate criminal defendants, the black defendants were often predicted to be at a higher risk of recidivism than they actually were, while the white defendants were often predicted to be less risky than they were. Which could tell about the discrimination between races does exist and it's not in accordance with the facts. And to be honest, it's obvious unfair to the black people, the reason might be that the COMPAS has treated the defendants with unfairness and the COMPAS is based on a data base with historical bias.
The COMPAS might treat each individual unequally, as if the system evaluates the defendants' risks of recidivism based on their races or individual finances. The system is just dividing the defendants into different groups by races and generate unfair discrimination to the black or other color, which is not obeying the fairness. On the other hand, the COMPAS may predict the risks of recidivism by mistake and lead the defendants suffer excessive punishment, as the algorithm of the system is not transparent to the public.

The reason of people characterised by as unfair

The reason of some people thought it's unfair is that the lack of transparency and the bias in the data of history. Although the COMPAS may not show its unfairness in the evaluation, but it's decision-making is affected by the algorithm and the data base. When the algorithm is closed source and people would challenge its accuracy, fairness and objectivity. On the other hand, the data of history may contains bias from the whole society, which could also lead to discrimination in the COMPAS' evaluation.

The fairness theories

To avoid this unfairness, we could emphasize outcome fairness or opportunity fairness. The outcome fairness should be when it comes to some similar scenarios with different defendants, each defendant should also has the similar outcomes. It shouldn't be different from black people to white people. And to the COMPAS, the evaluate procedure should be the very same under the similar circumstances. For opportunity fairness, it is necessary for each individual defendant to share equal opportunities of judicial justice. When the COMPAS is used, it should not let some groups of people being treated with bias and discrimination, especially those who were treated like this before.
So as my words, the COMPAS does show the discrimination against individuals, it needs to be more transparent and further comprehensive evaluations for the criminal defendants to eliminate the unfairness.

Question 2. Review the case of COMPAS and also the artificial job hiring case described in the paper by Linus Huang and collaborators. What do

these cases tell us about the role of 'Proxy'features in generating unfairness?

These cases tell us about the 'Proxy' features played a significant role in the emergence of unfairness. The 'Proxy' features are preventing the exposure of sensitive information like genders or races, but still they can represent these information to some extent, like criminal records, naive language or degrees. Then the 'Proxy' features may lead to bias and cause unfairness in making decisions.

The cases study

- In the case of COMPAS, although there's no evidence says that races are as a factor being input to the system, but the system still could speculate the suspects' background by criminal records and financial status which could be connected to the races of the suspects. With these factors the unfairness could be easily generated by individuals in the rating of risk scores.
- On the other hand, the case of AI hiring case described in the paper by Linus Huang and collaborators, the 'Proxy' features like home address of education could represent the races or gender of the candidates, although they weren't asked to provide these sensitive information. But these data of 'Proxy' features could still mislead the decisions full of unfairness which occurred in the social discrimination history.

Both of these cases demonstrate that reliance of ‘Proxy’ features could lead to significant differences between different demographic groups. When these algorithms are used to decide whether individuals are hired it might be especially a huge problem. Also the ‘Proxy’ features’ could create a cycle of feedback which perpetuate historical bias. "For instance, a job that was initially open to both genders could be affected if a female employee were to cause a serious accident in her position. Following this incident, the algorithm might take this factor into account, leading to a reduced likelihood of hiring women for that job in the future.

The conclusion

So these cases show that although we could prevent the sensitive information which may lead to the discrimination directly being analyzed by artificial intelligence, the system could still read the information by using 'Proxy' features and make a decision with unfairness.

Question 3. What is the solution if we discover that a ‘Proxy ‘feature is producing an unfair outcome.? How does your proposed solution relate to the distinction between algorithmic and moral fairness?

The solution to prevent misleading 'proxy' feature

To address the issue of a ‘proxy’ feature producing unfair outcomes, we could take these actions as a solution.

- First we could develop an algorithm to audit the bias in the input, and we could use human to oversight the whole process.
- When the algorithm identifies that there might be proxy features related to sensitive attributes like gender or races, we could remove the features from the AI model to prevent biased decision-making based on unfair factors.
- Of course we should make the whole algorithm transparent by open source or documenting its decision-making process with features it uses.

- To ensure that any questions regarding the fairness or justice of the model can be traced back to the root of the issue, we should establish a clear process for identifying the source of these inquiries.

The proposed solution's relation to algorithmic and moral fairness

Algorithmic fairness: this refers to an algorithm based on statistical or mathematical criteria could produce the outcome with absolute fairness. To ensure that, we should adjust the algorithm without bias and makes sure that this algorithm should be tested under all circumstances with different demographics group. And also we should make sure that the AI models data base is consistent with all demographics groups' data. The unfairness couldn't be part of the fact that affects the decision-making in this procedure because we only focus on adjusting the algorithm itself.
Moral fairness: We should focus on the ethical impacts of the decision-making process and its outcomes. The moral fairness does not just emphasize on the fairness of statistic but also in terms of social values and norms. The actions of human oversight and transparency could ensure that public could see the decision-making process is under the guidance of ethic considerations which could solute the moral fairness issues.
We should find the balance between algorithmic and moral fairness. By implementing technological solutions and human participation could reduce the harm of unfair, while we should focus on the ethical implications of solutions.

Question 4. Are there any situations in which it is appropriate and morally justifiable to use autonomous weapons. Is there a valid analogy with the use of medical technology?

The situations in it is appropriate and morally justifiable to use autonomous weapons could be complex and multifaceted. It could be in self-defense, a battlefield, some other environments with high risks of sacrifice and so on. When it comes to self-defense, it could be some certain situations that human lives are under immediate threaten. For example when the police are facing to murders with heavy power, they could use autonomous weapons to avoid unnecessary injuries or sacrifice. On the battlefield the troopers could use autonomous weapons to eliminate the risk of sacrifice themselves in attacking military target when the enemies are significantly more powerful than they are. On the other hand, some military work may expose them to extreme environments where there is a threat of human fatalities, they could use autonomous weapons to make their missions accomplished. But also we should pay attention to that the using of autonomous weapons may not put the operator into ethical issues. The autonomous weapons may avoid the operator to kill the target directly but still it's the operator's order to make the weapon kill. As it is different to the medical technology, the use of autonomous weapons is just killing while the medical use is for saving lives. There is a huge moral difference between these two kinds of uses.

- The similarities between autonomous using in medical and weapons are complex decision-making which involves significance implications of human lives. The Medical technology, such as robotic surgery, can enhance precision and reduce risks associated with human error, analogous to the potential benefits of autonomous weapons.
- The differences are much more important. The stakes of the war is not the same as those in medicine. The goal of medical care is to improve the patients' well-being and to save for lives, while the warfare would be destroying the target with less casualties as possible. Besides, the autonomous mechanics used in medical field are under the oversight of human ethics, which could ensure that the using would not be harmful to any humankind. But the autonomous weapons are only responsible for the military and there are less moral issues in the procedure.

Question 5. Does a natural language processing system understand the language it is processing? Does the difference between statistical and classic forms of processing make a difference to the issue?

The language comprehension of nlp

I don't think a natural language processing system does understand the language it is processing. As we input the natural language into the system, the system would start to learn about how to process the text by the pattern of humankind. They could identify the language structure and read the context with grammar, but they don't understand the meaning of the whole picture with its own awareness. Human's comprehension is involved in the tiny difference of the motivation of speaking, the emotions, the association of previous speaking and the meaning of words. The natural language processing systems are hard to tell the differences and not even to make a process of them.
The difference between statistical and classic forms of processing does make a difference to the issue.

The difference between statistical and classic forms of processing

- As we known that the statistical forms of processing are intended to collecting data from the real world, by using technologies like machine learning and deep learning techniques. The systems would analyze plenty of datasets to learn about how to process natural language and imitate the patterns of human language. The analysis could make the systems have more flexible and context-aware processing. Despite of the systems impressive outcomes, they still couldn't be smart enough to give their own opinion. They are just processing the natural language by learned patterns. They could answer your questions buy reading the context and searching for appropriate answers with references but not understanding your questions.
- While the classic forms of processing are simpler, the systems are coded with natural language processing patterns and they don't learn from the sessions. The accuracy of their answers depends on the completeness of the algorithms' logic. They could be effective in processing certain tasks but there would be more difficulties in other processing. These systems don't learn from the data and just rely on the coding which make these systems have low adaptability in processing with human language.

Question 6. Could there be a moral version of the Chinese Room argument?

Yes, there could be a moral version of the Chinese Room argument. The premise of the Chinese Room argument is that the person in the room do not understand Chinese at all and he is with a Chinese dictionary for answering the Chinese questions.
Imagine that we replace the dictionary with a moral principles, the questions would be transferred into ethics related, the person in the moral room would generate the answer to the questions by using the principles. The person, which indicates artificial intelligence could give the answer but he still don't understand what moral is. The moral involves including but not limited to compassion, responsibility and empathy. And also the moral criteria could be very different in different cultures. We should take questions of all respects in moral to test whether the artificial intelligence could really understand what moral is but not just take actions by the principles coded in advance.
Besides, the artificial intelligence could assert that whether the moral issues are correct or not because of they could not really understand what moral is. We should consider the hypothesis that the ability to perform morally relevant actions is equivalent to having moral understanding. Under this hypothesis that if the artificial

intelligence could adhere to ethics does that means the artificial intelligence shows its moral awareness like humankind?