LAB 6

Ethics

Q1 -- PARiS' lists of suggested applicants closely resembled the lists that would have been drafted by Strategeion's human HR team. To the extent that PARiS was biased towards a particular kind of applicant, this suggests that the human HR workers were as well. Indeed, it can be argued that PARIS is merely an extension of the human biases already in existence at Strategeion. Are computational biases necessarily worse than human biases in a recruitment context or are humans just as bad, what are advantages and disadvantages for each?

Response:
Traditionally, humans have been the primary means of screening job applications and making hiring decisions. Most of society is comfortable with this procedure, as most believe and value that humans can empathize with the applicant by understanding the nuances of the applicant's circumstances and make sound decisions based on this context and the information provided in the application and recruitment process. Sometimes, however, humans make mistakes, and these mistakes are often due to biases. Every human is biased in some way – it's the nature of how we are. These biases are constantly applied to help speed up human decision making. Thinking through all the possibilities of everything encountered would simply take far too long. For the most part, these biases help us to get along in the world without problems, but sometimes, they cause issues. For example, sometimes humans' generalizations about a certain group of people tend to fail. In these cases, humans re-evaluate the situation. This human re-evaluation is key.

In the relatively new digital age, computer algorithms have been proposed, built, and implemented to help streamline the recruitment process. These algorithms are attractive to hiring organizations because they are quick, scalable, and reduce workload for humans; thus, they ultimately save time and money for the organization. Unfortunately, computer algorithms are not free from biases themselves, Its fairly easy to argue that any human developed computer algorithm will inherently contain some human biases – these algorithms are developed by humans and trained using data informed by human decision making after all. Additionally, computer biases are only able to do what they are trained to do (they can't go above and beyond the information they have, they can't deal with unique or new circumstances well, and they definitely don't re-evaluate situations on their own).

Ultimately, both humans and computers can be biased. Although difficult to pick a side, I'd argue that computer biases are worse than human biases. Here is a table of pros and cons of both human and computer biases:

	Computer Bias	Human Bias

Pros	- Easily modified when identified - Consistent and objective	- Speeds up human decision making - Better understood and accepted by society
Cons	- Easily scalable - Viewed as ruthless/cold by society	- Difficult to modify - Difficult to self-identify - Inconsistently applied

While human biases are difficult to identify and even harder to change, they are well understood by society at large and well accepted as a part of human nature. Computer biases, on the other hand, are viewed by society as ruthless, cold, or dehumanizing. Even worse, computer biases are extremely scalable, and therefore can impact multiple individuals very quickly. It is important to acknowledge, however, that computer biases are much easier to correct when identified. For example, if it's known that the training data is causing a bias, the model can be refined and retrained very easily. Still, the scalability and current societal apprehension towards computer algorithms by general society are the main reasons I believe computer biases to be worse than human biases, especially in a hiring context where the decisions have major impacts on human lives.

Q2 -- Biased data pose a problem for ensuring fairness in AI systems. Given the company's demographics, what could Strategeion's engineers have done to counteract the skewed employee data? To what extent do you think such proactive efforts are the responsibility of individual engineers or engineering teams?

Response:
Biased training data contribute to the creation of an algorithm that is great at being biased itself. This is therefore a major problem for individuals developing algorithms to be aware of, and one that should be tackled with care.

For Strategion, the use of biased training data became a big problem. The Strategion developers intended to use their algorithm to identify anyone who would have been seen as a good fit for the company by a human. They trained their algorithm using labelled resumes from successful and unsuccessful applicants to their company. Unfortunately, this ultimately led to an algorithm that learned to be biased against individuals based on arguably arbitrary characteristics. Individuals, such as Hara, were screened out by the algorithm based on their participation in sports, although this is not considered by HR to be relevant to success on the job. Strategion's algorithm was ultimately found to be biased because of its training data.

To counter this, Strategion's engineers could have done a few things. The first thing that they could do is complement their training data with resumes from other, external, sources. This approach may have helped to make the algorithm aware of more diversity in applicants by increasing training data size. To do so, they could have tried to get resumes from other companies, although this would likely have been a challenging due to concerns about competitive advantages.

Another strategy that the engineers could have tried to counter the skewed data is to, in some way, emphasize the importance of minority classes (i.e., unique resumes of individuals that were still successful). This may be done in some way by changing the way the model trains (for example, weighting certain resumes as more important), or could have been done manually by designing the training dataset to contain more examples of minority classes that are still a good fit for the company.

Strategion's engineers could have also tried to counteract the skewed employee data by designing the model to ignore certain parts of the resume (those that typically contain minimally relevant characteristics, like sports, would likely be under a hobby heading, for example) to prevent selection based on "less important" characteristics. Actually implementing this would likely be very challenging, given the fact that resumes are very unstructured and have no standardized format. Additionally, this approach may not actually be helpful as this information may be able to be inferred using other "important" parts of an individual's resume.

Flagging unique/new/uncertain cases for review is a final example of an approach that the engineers could have used to counteract skewed data used in training. With this approach, when an individual's resume is not interpreted or recognized by the model well, then it could be tagged for review by a human. Ideally, this approach would still allow the model to cut down workload with only a few resumes requiring human review every so often.

Carefully testing the model's performance in many ways is a very important part of the development process to ensure that the model is working as intended. The model developers could have better understood the strengths and weaknesses of the model by testing it within different subgroups of data (for example, unique resumes, resumes from minority groups, resumes from non-traditional applicants, etc.). The knowledge gained from this testing would be very valuable to identifying points for improvement before putting the model into practice.

All the above proactive efforts to ensure that the model is performing optimally before rolling out to practice involve the engineers/model developers in some way. This is because these individuals have the technical skills to develop and test the tool. The entire burden of this process, however, should not fall on the engineers alone. This is because engineers are trained to be able to develop tools etc. but are not trained to have expertise in, for example, the hiring process. To ensure that any model is truly effective, it needs to be properly informed. Therefore, it is likely best practice to involve any individuals to which the model is relevant in the development process. In the case of the Strategion resume screening algorithm, this would include involving individuals like representatives from HR, who play a central role in typical hiring procedures and are well versed in the typical challenges, pitfalls, and strategies for fairness. It can even be argued that involving HR alone is not going far enough. Company leaders, for example, likely have valuable input and concerns about a tool that will ultimately directly influence the makeup of the company. All in all, any algorithm should be, ideally, informed by the broadest audience as possible to ensure that diverse perspectives and considerations have a chance of being incorporated.

Q3 -- When it comes to hiring decisions, do you think there should always be a "human in the loop" to make sure that machine decisions don't bypass human qualities? Why/why not? What would this entail, considering that we might be dealing with very large numbers of applications when making hiring decisions?

Response:
In general, it's a good idea for humans to be in the loop and supervise the work of an algorithm. This is especially true when such algorithm is being used to make very important decisions. Having a human in the loop is important because algorithms are not perfect; even the most highly advanced ones still have bias (like humans), and more importantly, algorithms fail to consider the complexities of the world in the same way that humans can. Hiring decisions are a good example of a very important decision because they directly impact human lives in a major way (income, which directly influences multiple essential aspects of life in a capitalist society). When an algorithm makes a hiring decision, it makes it based off the information it has and what it has been trained to do, without any regard to the impact that decision has on the person it is related to. Algorithms are cold, hard, and fast. When a human makes such a decision, they are more aware of the context of such decision and ultimately can empathize with the individual and feel regret for turning them down. These human qualities and considerations are important to job applicants. Having a hiring decision boiled down to a quick algorithm makes job applicants feel like a "number" – it makes humans feel stripped of their humanity. Having a human consistently in the loop, supervising algorithms that make decisions as important to other humans as hiring decisions is a bare minimum essential to ensuring that humans still feel valued.

In practice, algorithms are very helpful to companies. They help to reduce the workload of those humans working for the company, and ultimately save time and money (scarce resources). Finding the balance between automation of hiring decisions (which, when managed traditionally can be very demanding and time consuming) and maintaining the traditional human considerations that go into hiring candidates is not simple. I think that having a human consistently supervise the work of a hiring decision algorithm with structure is one potential approach to balancing workload demand and the societal value that is placed on human considerations that traditionally have gone into hiring decisions. For example, a human could be responsible for reviewing every nth application and decision made by the algorithm. Not only would this ensure balance between workload and societal value on human considerations in hiring, but it would also provide opportunity to continuously review the algorithm's work and provide feedback for its improvement.

Q4 -- Job interview companies Pymetrics and Hirevue implement games from psychology research to more accurately determine the qualities of the applicants. They can also use facial recognition software to detect and judge emotional reactions during interviews. Both companies have statements around their focus on fairness in the interview process and work to eliminate human bias. What do you see as the main potential advantages and disadvantages with hiring interviews conducted this way and do you think such companies will improve the hiring process overall?

Response:
Using games to judge human qualities and facial recognition software to judge emotional reactions during the interview process is quite controversial.

For companies, these tools are likely viewed as beneficial. They provide an additional supposedly evidence backed and unbiased way to accurately determine the qualities and abilities of any applicant. This could be extremely helpful for screening out, or providing an additional metric on which to judge, applicants and thus these tools cut down workload for humans and ultimately save time and money for the company. Most hiring teams are very likely to view these benefits positively and believe that they are identifying the best candidates by using them.

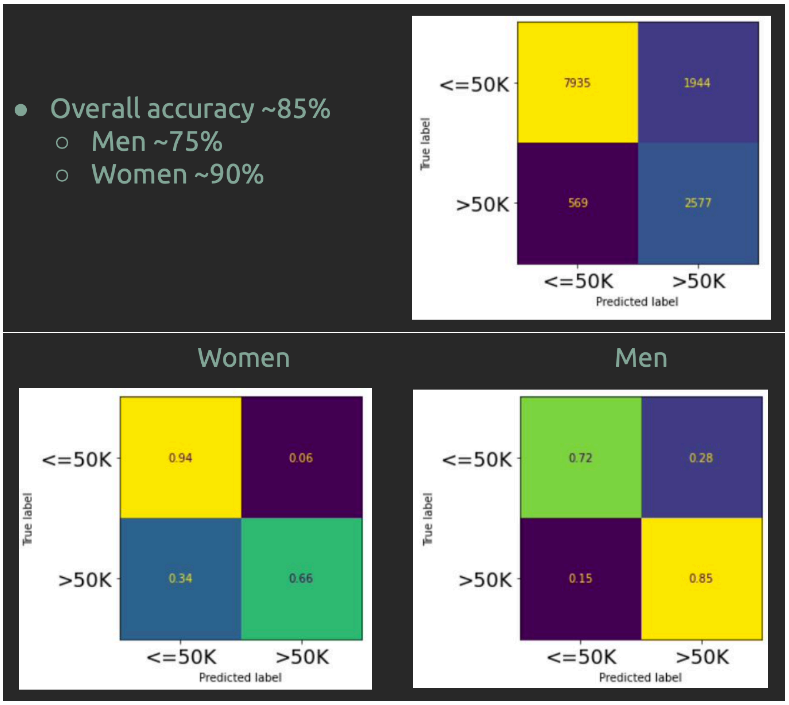
Job applicants are very likely to immediately question and scrutinize these tools, as their fate as an applicant is directly impacted by them. If I were an applicant, I would be concerned about, for example, how the games were developed and validated (the context (cultural, ethnicity, etc.), the training data), and what the games/algorithms are actually designed to do (are the hiring teams using them correctly?). I think that these questions are likely to be asked by other applicants too, and for valid reasons. Games and facial detection algorithms are likely to be based on machine learning models, and as we have learned in MEDI 504A, there are many limitations to using these methods. For example, the training data may have been biased, or incomplete. Additionally, these models are often designed to measure very specific things any may not be a complete measurement of an applicant's suitability for a particular job. Those who decide to implement these models should be completely aware of the limitations of these models, their true effectiveness, and be transparent about their use with those they are used on. Awareness will enable hiring committees who are truly passionate about finding the right candidate to develop a method (that may or may not use these tools) that provides a holistic understanding of the applicant and their suitability.

Here are some pros and cons of facial detection and games for hiring purposes:

	Facial Detection	Games
Pros	- Supposedly judges human emotion responses accurately - Reduces workload burden on humans - Supposedly eliminates human bias in judging emotional response - Provides additional metric for hiring decisions	- Supposedly detects human qualities accurately - Reduces workload burden on humans - Supposedly eliminates human bias in measuring human qualities - Provides additional metric for hiring decisions
Cons	- May not be broadly applicable (different genders, ethnicities, cultures, age groups) - Applicants may be frustrated, distracted, or pushed away by use of this tool - Specific disclosure of the use of such tool likely required	- May not be fair for different groups – groups with more experience playing games may perform better - Specific disclosure of the use of such tool likely required - Applicants may be frustrated by tool or not believe it is fair.

If used appropriately by a team that is aware of the limitations of each method, I think that these methods have the potential to improve the hiring process overall. From the perspective of those hiring candidates, these methods may help to cut human workloads, save time, and provide an additional metric on which to rank applicants while also ensuring that a candidate truly has the qualities a company is looking for (assuming the tool is accurate, generally). Applicants could also benefit from the utilization of these methods - they will only end up being hired for positions in which they are truly a good fit, and this is likely to result in higher job satisfaction. Proper use of any tool, especially when involved in important decisions, is key.

Q5 -- We are building a model in order to determine who to approve for a loan based on their predicted income. We are using US Census data that includes information about education, marital status, sex, race, etc. We want to predict who makes >50k / year and ensure that our model doesn't have a bias towards either men or women in the prediction.
On the next few slides you will find the model accuracy and confusion matrix, both overall and for women and men separately. Do you think this model is fairly treating the two groups? Why/why not?



- Overall accuracy ~85%
  - Men ~75%
  - Women ~90%

Response:
Based on the provided confusion matrix and model accuracy information, this model is not treating the males and females fairly. Here's why:

1. There is a relatively high chance (34%) that women who make more than 50K will be predicted to make less than 50K by the model. In contrast, men who make more than 50K are less than half as likely (15%) to me mislabeled as making less than 50K by the model. Ultimately, there is a high (2 times higher) false negative rate for women. Individuals who make more money are more likely to be approved for a loan. Because this algorithm is more frequently mislabeling females as making less money than they truly do, it is biased against women for loan approval.
2. The model performs very well when it comes to identifying women who make less than 50K. Women who make less than 50K are identified as such 94% of the time. Men, on the other hand, are quite likely to be mislabeled as making more than 50K when they truly make less than 50K (28%) (False positives). Again, for the purposes of loan approval, women are again disadvantaged by this model; men are more often assumed to make more than 50K even if they don't, and those who make more money are, in nearly all cases, more likely to be approved for a loan.

Based on the above comments, it's clear this model favors men for loan approval – they are less likely than women to be mislabeled as the lower income category when they are truly part of the high-income category, and more likely than women to be accidentally labelled as the higher income category when they are truly part of the lower income category. This model should be modified.

Additionally, this model is reported to be far more accurate in females. That being said, the overall model accuracy is closer to the accuracy among men, and this is indicative of there being many more males in the training data. The model may perform better overall with more balanced training data.

All in all, the model should be modified, in some way, to ensure that females are not disadvantaged as they are in the current state.