**The AI Ethics of Fairness (Part 1 of 2)**

Audrey Louie, Jose Neopaney, Ana Paredes Fernandez

**Jose’s Part**

The case study on AI in medical diagnostics reveals both the promise and pitfalls of artificial intelligence in healthcare. A German company developed an AI tool to detect skin cancer by reading images of lesions, assigning a color-coded circle to each spot: white for benign, yellow for uncertain, and red for potentially cancerous. However, the AI model tends to perform better on lighter-skinned individuals than on darker-skinned individuals. Some might jump to conclusions, attributing this disparity to racial bias, but the underlying causes are more complex and rooted in data and visual differences.

One key factor in this issue is **representation bias**. Since lighter-skinned individuals statistically experience skin cancer more frequently, most available training data comes from lighter-skinned patients. Like a football player improving with more reps, the AI model learns more effectively when it has ample and diverse data, gaining accuracy as it encounters more examples. With limited exposure to darker skin tones, the model has fewer “reps” in analyzing these images, leading to lower accuracy on darker-skinned individuals. By incorporating more diverse training data, the AI can improve its diagnostic accuracy across all skin tones.

Another reason the AI performs differently on darker and lighter skin is due to visual and technical distinctions. Lighter skin tones, often having greater contrast with lesions, make it easier for the AI to identify variations, while darker skin, which absorbs more light, provides less visible contrast. This is not about the AI being “racist” but about inherent visual challenges when interpreting different skin tones, a common issue in medical imaging. Just as players in football adapt to new environments by practicing, AI models can be “trained” on data from diverse skin tones to better handle these variations.

The issue extends beyond healthcare; similar problems occur in fields such as criminal justice and hiring, where AI models may favor one demographic unintentionally due to biased data. For example, in hiring, some AI tools favor male candidates over females if trained on historical data reflecting past hiring biases. Likewise, facial recognition technology has faced scrutiny for higher error rates with darker-skinned individuals due to insufficient representation in training datasets. Each of these cases illustrates the need for balanced, representative data to achieve fair outcomes in AI.

Addressing AI fairness often requires a balanced approach to resource allocation. Since skin cancer disproportionately affects lighter-skinned individuals, it makes sense to invest resources in ensuring accurate diagnoses for this group. However, an equitable healthcare approach would also mean expanding AI’s ability to identify skin cancer on darker skin to ensure everyone benefits from accurate, fair diagnosis. By enriching datasets to reflect diverse populations, AI can achieve fairness without sacrificing accuracy.

Ultimately, this case study emphasizes that fairness in AI is less about correcting an “unfair” algorithm and more about understanding the nuances of data and representation. With thoughtful, inclusive data practices, AI can evolve into a more accurate, equitable tool across various applications, from healthcare to hiring, benefiting diverse populations more fairly.

**Audrey’s Part**

**What Is the Definition of Fairness?**

According to ChatGPT, fairness is defined as treating all individuals and groups equitably. However, this definition is incorrect, as shown in our previous case: it is impossible to treat someone equitably in terms of both skin color and vulnerability at the same time.

Aristotle defines fairness as “treating equals equally, and unequals proportionately unequally, within a specific decision domain.”

**Examples of Equals Treated Equally:**

* A hiring manager needs to choose between two lawyer applicants. The only difference between them is height, with no additional information about their backgrounds. Who should be hired?
	+ Height should not influence the hiring decision, as physical appearance does not determine one's ability to be a better lawyer. Therefore, both candidates should be given an equal chance.
* A school principal must choose between two students, one boy and one girl, to represent the school at a science fair. Both have identical grades and achievements. Which student should go?
	+ Gender should not play a role in the decision because both students are equally qualified to attend based on their academic background.

In both cases, flipping a coin ensures an equal and unbiased chance for each individual.

**Examples of Unequals Proportionately Unequally:**

* A hiring manager is choosing a player for a basketball team. One candidate is significantly taller than the other. Who should be chosen?
	+ In basketball, height is a critical factor in basketball performance, it is reasonable to prefer the taller applicant. Here, unequal treatment is fair and proportionate to a relevant characteristic of height that impacts job performance.
* A scholarship committee is deciding between two students: one comes from a high-income family, and the other from a low-income family. Both students have similar academic achievements. Who gets more financial aid?
	+ The student from the low-income family should receive more financial aid. The unequal treatment is fair because it addresses the greater financial need of the lower-income student.

**Special Case: Manager Hiring**

* A company is hiring a manager. There are two applicants, and one is taller than the other. Academic research suggests that taller people perform better than shorter individuals as managers. Should the company hire the taller person? This case is challenging because height should not generally impact the ability to manage. The question remains: is it fair to choose taller people based on this evidence?
	+ There is no right or wrong answer to this question, it depends on if the company wants to use the research or give the applicants an equal chance.

**AI in Emergency Calls**

Another interesting case highlighted in the video involves AI technology used alongside human operators to respond to emergency calls. The AI indicates when a patient is likely having a heart attack, requiring immediate attention. Studies show that AI can identify heart attacks more rapidly than humans. In this case, the decision domain involves the language spoken.

**Definition of Fairness in Academic Papers**

Based on academic papers the definition of fairness is “the absence of any prejudice or favoritism towards an individual or group based on their inherent or acquired characteristics.” This definition is wrong, as seen in the example of the basketball hiring because height is an inherent characteristic. If fairness always required ignoring relevant characteristics, all decisions would need to be made randomly, which is impractical in many scenarios. This prevalent definition suggests there is much work to be done in understanding fairness.

**Definition of Fairness - Conclusion**

Overall, the definition of fairness depends on how you set your decision domains. What is considered fair can vary based on the specific context and the relevant characteristics being considered. In some cases, treating everyone equally is appropriate, while in others, proportionate and unequal treatment is justified based on meaningful differences.

**Ana’s Part**

Ai’s usefulness has increased so much that it has reached some companies' hiring process. In other words, companies use the Ai skills to help them cut down the number of applicants they can hire.

**Case #1:**

* Retorio is an Ai trained to measure the applicants qualities based on the video they submit.
* The qualities measured are: Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism
* The issue: it has been observed that some qualities like headscarves, glasses or backgrounds, have been affecting the score of some participants.
* Is this a fair rating system? Should external factors decide if a participant is more qualified than others? Is the Ai making mistakes?
* Computer scientist Katharina Zweig states that: “The fundamental problem with face recognition (in hiring processes) by machine learning is that we never know exactly which pattern in an image these machines are responding to”
* Based on tha statement we can assume that the creators of “Retorio Ai” have no biases over what factors is the Ai focusing on to pick the “better candidate”. Is it the candidate's words or external contributors? (We are inclined to believe it is probably the external elements since in one of the examples is the exact same video but one has a different background)
* This Ai was trained based on 2,500 people’s response to who they would hire. Therefore, another interpretation of their words can be that they do not guarantee the best candidates, they guarantee candidates a human would consider good. So is there still a fairness problem when the Ai has been trained to mimic human decisions? If the Ai purposely commits a mistake, is it not doing its job?

**Case #2:**

* Amazon also tried to implement Ai to their hiring processes. They used their worker’s profiles and trained the machine to look for candidates with similar or identical characteristics as them.
* Like other software companies, Amazon has more male workers than females. Therefore, the Ai took in negatively the word “female” since it didn't match the current workspace Amazon has.
* The company tried to take away this gender bias method the Ai had, but it was a difficult task since the machine kept figuring out the gender using the information of the candidate.
* When using an Ai machine, past data bias will create future data bias. This means that if the Ai senses a bias in the past, it will keep on recommending on using that same bias for the future hires.
* Technically the Ai is still somewhat effectively doing its job, since it cannot understand the wrong in the discrimination, the machine’s objective is clear on following Amazon’s current workers characteristics as the guide to hire more workers.

**Questions:**

1. Is the problem technology or the way humans establish society? Should developers get involved in “re-balancing” the machine if it has a preference that society would consider “unfair” to rate on? Or is that outside of their scope? Is it even possible to “re-balance”?
2. What happens if there is an imbalance based on a factor that is invisible to common sense? Ex. People’s last name in the company starts with the letter “W”, therefore Ai hires more people with this letter.
3. Can the machine be used if unbiased results cannot be guaranteed? To what extent should the AI allow these biases? Are invisible biases fair?