**Question 1 of 1**

AI systems are only as good as the data we put into them. Bad data can contain implicit racial, gender, or ideological biases. Though AI is capable of speed and capacity of processing that’s far beyond that of humans, it cannot always be trusted to be fair and neutral. AI is used to identify people, objects and scenes. Many AI systems continue to be trained using bad data, making this an ongoing problem.

These are just a few example:

1. [Camera misses the mark on Racial Sensitivity](https://gizmodo.com/camera-misses-the-mark-on-racial-sensitivity-5256650)
2. [Machine bias](https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing)

Submit your proposal in a minimum of 1000 and a maximum of 4000 words to mitigate bias from an AI system.

**Proposal**:

**Proem**

AI is the study of the design of intelligent or rational agents. In normal words, it is an exciting new method or art of creating machines that perform functions that require intelligence when performed by people. AI is used to identify people, objects and scenes.

But AI systems, due to the Bad data, may become biased and show discrimination against race, nationality, gender, skin color, etc. Though AI is capable of speed and capacity of processing that’s far beyond that of humans, it cannot always be trusted to be fair and neutral.

**Example** (for better understanding)

By having an AI read text on the Internet, it can learn about words, and you can ask it to reason about analogies. So, you can quiz the AI system now that you've read all this text on the Internet, in the analogy, man is to woman as father is to what? So, the AI will output the word mother. If we step on to the biasing part, the researchers also found the following result, which is that if you ask it, man is to computer programmer as women is to what? That same AI system would output the answer, woman is to homemaker.

If we want our AI system to understand that men and women can equally be computer programmers, just as men and women can equally be homemakers, then we would like it to output man is to computer programmer, as woman is to computer programmer, and also man is to homemaker as woman is to homemaker.

Second, there're also some facial recognition systems that seem to work more accurately for light-skinned than dark-skinned individuals. If an AI system is trained primarily on data of lighter skin individuals, then it will be more accurate for that category of individuals to the extent that these systems are used in, for example, criminal investigations, this can create a very biased and unfair effect for dark-skinned individuals.

**Given Defendant’s Risk Factor example:**

It is mentioned that “the Rating a defendant’s risk of future crime is often done in conjunction with an evaluation of a defendant’s rehabilitation needs”. There are 137 questions that the software asks the defendant before it yields the Risk Factor. Yet, the score proved remarkably unreliable in forecasting violent crime: Only 20 percent of the people predicted to commit violent crimes [with high Risk Factor] actually went on to do so. While the ones with less Risk Factor were involved in violent crimes later.

The considerable points are –

* The company does not publicly disclose the calculations used to arrive at defendants’ risk scores, so it is not possible for either defendants or the public to see what might be driving the disparity.
* The questionnaire also asks people to agree or disagree with statements such as “A hungry person has a right to steal” and “If people make me angry or lose my temper, I can be dangerous.”, which could actually help to understand the temperament of the defendant better.
* “A guy who has molested a small child every day for a year could still come out as a low risk because he probably has a job,” Boessenecker said. “Meanwhile, a drunk guy will look high risk because he’s homeless. These risk factors don’t tell you whether the guy ought to go to prison or not; the risk factors tell you more about what the probation conditions ought to be.”

**Mitigation**

Less bias <= More inclusive data

For example, if you are building a face-recognition system, and make sure to include data from multiple ethnicities, and all genders, then your system will be less biased and more inclusive.

Having transparent systems as well as systematic auditing processes increases the odds that will at least quickly spot a problem, in case there is one, so that we can fix it.

If you have a diverse workforce, then the individuals in your workforce are more likely to be able to spot different problems, and maybe they'll help make your data more diverse and more inclusive in the first place. By having more unique points of view as you're building AI systems, I think there's a hope all of us create less bias applications.

If we consider the Risk Factor example, it is the CRIMINOLOGISTS who were behind the analysis or research of such calculations of the software. An object may be towards right for a person and left to the other. That says that the perceptions of every person may not be same always.

If the accuracy of the Northpointe’s software is low, as stated that only 20 percent of the people predicted to commit violent crimes actually went on to do so, then it should not be considered as an erroneous software, because, the risk factor is “quite a prediction”. It is just a statistical estimation of the future risk caused by the defendant. Instead, it could be considered as an erroneous software if it had taken some inputs from the defendant which is related to his race, nationality, skin color, etc. These kind of inputs may effect the risk factor in a negative way, resulting a wrong assessment during criminal sentencing.

To manage bias in AI:

* Choose the right learning model for the problem
* Choose a representative data set
* Monitor performance using real data.

**Peroration**

NO MACHINE IS PERFECT.

Yes, no system can be perfectly perfect. Every machine has its own defects. What matters is the impact created by that defect, whether it should be overlooked or to be fixed. If there is an adverse impact due to that defect, we should look for a remedy that could fix it.

All systems are human made objects. Their response to a given stimulus is what they have acknowledged from their input analogies. They are not the ones which intentionally discriminate. Bad data is behind this.

AI systems are making really important decisions today, and so the bias or the potential for bias is something we must pay attention to and work to diminish. One optimistic thing about this is that we actually have better ideas today for reducing bias in AI than reducing bias in humans. So, we should never be satisfied until all AI bias is gone, and it will take us quite a bit of work to get there.

In addition to the problem of bias, one of the other limitations of AI is that it can be open to adversarial attacks like generating fake comments, DeepFakes, spams, frauds, undermining of democracy and privacy, etc. Because of this, even though the AI community still has a lot of work to do to defend against these adverse use cases.

Identifying and mitigating bias in AI systems is essential to building trust between humans and machines that learn. As AI systems find, understand, and point out human inconsistencies in decision making, they could also reveal ways in which we are partial and cognitively biased, leading us to adopt more impartial views. In the process of recognizing our bias and teaching machines about our common values, we may improve more than AI.

**Crux:**

* Every machine has its own defects. What matters is the impact created by that defect.
* Increase inclusive data.
* Prefer transparent systems as well as systematic auditing processes.
* Improve diversity and mitigate diversity deficits.
* Be aware of proxies and Technical limitations.
* Prevent discriminatory effects [Right to an explanation (for defendants)]
* To manage bias in AI – choose the right learning model for the problem, choose a representative data set and monitor performance using real data.