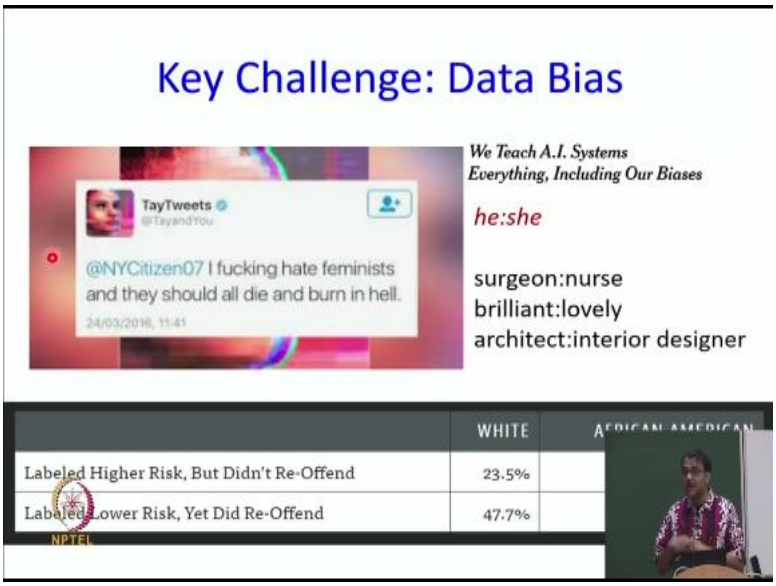
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 94
Ethics of AI: Data Bias and Fairness of AI Systems

(Refer Slide Time: 00:17)



Now this is a New York Times article just came a few days back. A beautiful catch line, we teach AI systems everything including our biases. All data sets are biased, right because they have been collected by some mechanism. Remember we had this whole discussion about sampling, if you want to figure out you know whether what is the average age of somebody right?

Like somebody who has in the city, right how do you figure this out? When we discussed this in the time of Bayesian networks that you know you can call. Then there will be a bias for people who have phones, you will call in the day. Then there will be bias with people who stay at home etc., etc. So we give all those biases right? We give, we create the data in because of that we create bias.

We give labels and sometimes because of that we create biases. This is a quote I believe from Professor Emily Bender at University of Washington a very well-known linguistics, computational linguistics researcher. She thinks about all the languages in

the world. The dying languages in the world, how do they connect with each other? What are their common properties across many languages?

What are their unique properties etc. a fun person. So I believe this quote comes from her. And some very surprising observations of current AI system. So for example in an LP system was asked he is to she then what is to what? And if it has a good understanding of words, then hopefully it will be able to come up with some you know things like father is to mother.

You know that will be a good answer or daughter is to son is to daughter that will be a good answer. But it came up with answers like also those but also came up with answers like surgeon is to nurse. And brilliant is to lovely. An architect is to interior designer and so on so forth, right.

And all the women in the class and also many of the men in the class, hopefully all men in the class should be disappointed and unhappy at this particular observation of the AI system that it thinks that surgeons are men and nurses are women and sort of propagates our biases further. Surgeons are or men are lovely, sorry, men are brilliant. And women are lovely, right? It is just very sad.

And is it what the AI system learned completely by itself without any human bias? Of course not. We gave it the text. Who wrote the text? We wrote the text. Is it bias that exists in our society? Of course it exists in our society. But what it did is that it said okay, I will take the systematicity in there and get rid of the noise.

So maybe the 10% or the 20% of the examples where surgeons were women or women were brilliant or whatever it is, men were lovely it just removed that as oh, that is noise. And basically just accentuated the biases further, just expanded on the bias, exaggerated them further. And we absolutely do not want that. We have biases, we have to get rid of our social problems, we have to correct for them, but there are people who have you know done better.

There are people who are not sexist, right or not that much sexist. There are you know those ideas should also come into the AI system but because they are smaller in

number, and because it is a majoritarian view because it is always looking at systematic patterns and getting rid of the noise which it cannot model it is sort of only going to take the majority view. There is a very beautiful experiment that got done.

And Microsoft in all their wisdom came up with this idea that why do we not, why do we not have humans all over the world teach a robot how to speak or how to talk? Not the speech part, but how to the words the language. So of course, where would they put such a system? They would put it on Twitter. Do they not know that people on Twitter do not really speak English first of all?

And then we all know that you know people who are on Twitter are different than the same people who meet you face to face. We all know this, I mean, you know yourself. So we have this beautiful phenomenon called trolls. I am teaching to many of them and they can take any idea and you know take it to the extreme. And what would happen if Microsoft released robot which learnt the speech or speak of Twitter users?

Not surprisingly, it will become highly racist, highly biased, highly so it became a Nazi. It became sexist, it became racist. It became everything possible. It started using curse words. It is basically all of that. Here is one of the nicer tweets that I can show you from Tay where other tweets I cannot even show you, right. And so what does that mean? That means that Microsoft is going to get really bad pressed.

And moreover whatever English it will learn is not the English Microsoft wants robot to learn. So what did they do? What would you do if you were Microsoft? Come on, this is obvious. If Tay was your project, what would you do? Yeah, take it down. Come on. It took them 24 hours to take Tay down. It was clear that this is a failed experiment. We cannot get the collective power of the humans especially on Twitter.

Maybe we can get it on Mechanical Turk because we are paying them some money and maybe they will do a good job, but not necessarily on Twitter, right? These kinds of things show again and again and again. So for example, here is a, a tool that was created or that was developed where you take an offender who has come to court, and you come in front of a face recognition system.

And the AI system analyzes the face, now you can think about what would that mean? And predict whether this offender will repeat offend, right? Will make a do another offense in a year or two or something like that. Is this person a repeat offender? Or is this person a one-time offender? And what would you expect this system to be really doing? It would be trained from our biases in the US. So what would it say?

Well, it will make more mistakes. It will make mistakes because this is a hard problem. But what kind of mistakes will it make is a very important question. And so somebody did this analysis where they took all the white offenders and the African-American black offenders and they said, which subset of people were labeled higher risk, right. But they were actually nice people. They did not offend again.

They did not make an offense. They did not repeat. They lived lawfully, somewhat. And the answer is, yes it is a mistake, but it is a mistake, where we are saying you are sort of worse than what you are. And then there were people who are labeled as lower risk, but they did commit another crime. And again, the system will make those kinds of mistakes also, right. The system is saying, you are fine.

You know give I can give you bail, but then you go back and do something bad. So of course, our system will make mistakes. But it made more mistakes for the first time for African-American folks and more mistakes of the second kind for Caucasian folks. Basically, it was using skin color as a big proxy or proxy, right? It is not exactly using skin color.

We do not know whether it is using skin color, but it is using something that is correlated with that at the least in order to make these kinds of judgments. Yes Parth. Yes, good. Parth says what the heck is this? In not in those many words, but why would I look at somebody's face and classify this kind of a thing? And I do not know the answer to that. I mean, whoever did this, they must have some wisdom of some kind.

But basically, we can see what kind of wisdom is coming out. So therefore, this particular system was labeled as racist. And I do not think it was being used. It is being used for bail prediction, but you can imagine. Then these US police folks were

given this face recognition system, which will just stop a car if somebody something is going on, put this in front of them.

And the system will say whether this person is found in the criminal database or not. It had 5% error rate in US lawmakers. So 28 US lawmakers were quote unquote found in the criminal database. So imagine if this is happening in the US lawmaker, then what would be happening to public at large, right? So many such examples. There is a face recognition system from Google, where black people were labeled as gorillas.

The people have found that women are less likely to be shown ads of high paying jobs than men. And by the way, Google knows what you are. So what is going on? Well, what is going on is the data bias. We are training these systems with a lot of bias data. And if we are training these systems with bias data, because we have biases, those things will get exaggerated and the systems will not be very usable or we would not want to use them.

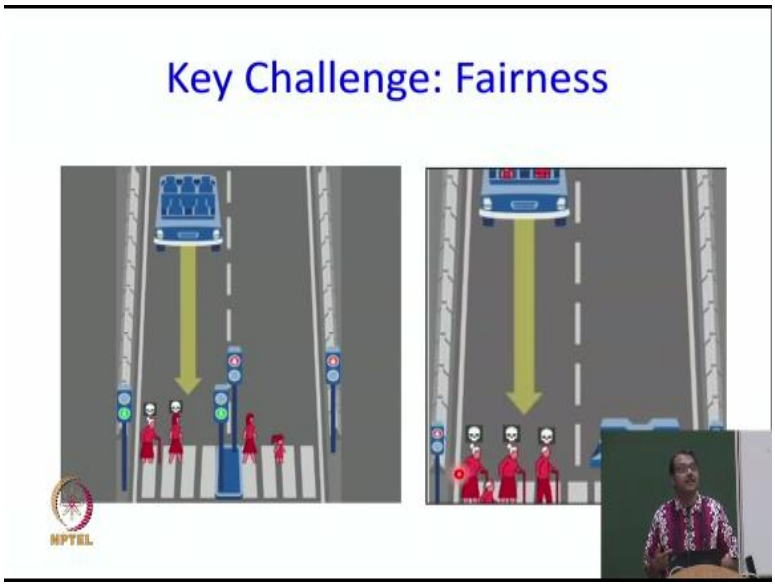
It will give us bad press. By the way, there is one other thing that people have found that we expect more from our AI systems than we expect from individual humans. That if an AI system comes in, then it should be somewhat I mean it should be good everywhere across the board. It should not make any kind of obvious mistakes, or it cannot make any kind of this mistake, that mistake.

If a certain police officer or somebody was racist, you know we will just discard it. But if the AI system was racist, that is not good. And you can question you can argue whether that is a good point of view to have, but it is a reasonable point of view in some ways, because this one AI system will be used many places, whereas this one police officer is acting only in a small district or whatever, right.

So you know there are issues and challenges that we need to resolve before we can get to a real successful deployment of AI systems, right. And a new field has come out because of this. Is called FAT ML fairness, accountability and transparency in machine learning. So today is the time to work on ethics of AI. What is an ethical behavior and what is not an ethical behavior?

How do I make my systems less racist? How do I make my systems less sexist? How do I make my systems fair?

(Refer Slide Time: 11:23)



Now here is the question on fairness. We all know that you know someday there will be automated, there still are, but more and more people will start using automated driving cars. Now when we are in a dangerous situation when our car brakes have failed, you know there is an accident that is going to happen. At that time, we are in a moment of panic. We do not know exactly what is the right behavior.

We do the best we can in the circumstance. It is possible that people may die because of that. It is possible that you know some cars may be injured, but it would be okay because we were in a moment of panic. Humans are not expected to behave properly in an accident like danger panic situation. If you did, you will be awesome. But if you did not, we will not take you again, take it against you.

But an AI system can do very fast processing. At the time, it knows that the brakes have failed it may have one second, but that one second is a lot of time for an AI system or five seconds that is a lot of time for an AI system to make a deliberative decision. It knows that the brakes have failed, it knows it needs to crash. Now it needs to figure out who to kill. And then you can start asking this question who to kill?

So if you are you can keep going straight and you can kill these two women, one older and one younger. Or you can go into the other side and kill these women, actually one

woman and one girl. Who would you want to kill? The first group of people or the second group of people? Pretty much everybody will be happy killing the older woman. Because those are the only two choices we have right?

Now this thing starts to get interesting. Suppose you knew that this particular woman here is a doctor. Does your answer change? Suppose you knew that this particular woman was a man. Does your answer change? There are many questions like this. Now here I have two people. Suppose there was five older women here and then there were two people here or four older women here and two people. Then who would you kill?

And when you start asking these kinds of questions it starts to get to the core of your biases. Who do you think is more valuable in the society? And pretty much everybody says that if I if in one case I can kill three people. And in one case, I will end up killing myself. Sorry, this is not how I should say this. So suppose there is only one driver in the car or one passenger in the car no driver, this is an automated car.

And suppose the car could go straight and kill three people and go to the left and kill itself. And this one person inside the car dies. Pretty much everybody believes it is okay to have the person sitting in the car be killed. But let us not kill three people because three is much better than one. And then they were asked, will you buy such a car? And the answer was, no.

If I am buying the car, I want to make sure I live, this is my first priority. So what we believe others should do and what is the right thing to do for the society has nothing to do with how we think of ourselves. Again, this may or may not be surprising to you. And then there is one group of people from MIT, there is something called moral machine. You can go play it online, you can give them some data, moralmachine.mit.edu.

And you go there and basically they start asking you these kinds of questions. And they start giving you more information. This one is a engineer, this is studying in class second, whatever it is. And then they ask you, who should you kill? And they keep

doing this. And over time, they just do this collectively for societies and start to figure out okay, what is India? Does it value men more versus women more?

Or does it value young people more versus old people more? Or does it value engineers versus whatever? And you realize that different countries have very different value systems? Extremely different value systems. So thanks now currently, the Tesla cars do not have the ability to figure out who to kill. But it is not very complicated to imagine that tomorrow they will.

And then they will all these ethical questions are going to come up again and again. Yes, Parth. The question Parth is saying is I do not understand why you want to add the bias and who to kill because this can happen, you have the time to make that informed decision. What would you do if you are the designer of the automated car and your brakes have failed and you have two choices, what would you do?

You will just keep going straight irrespective of what is in front of you, you will take a toss a random coin and go 50% somewhere, you may have to make some choice. In fact, here is an interesting question. The choice can come from the probability of saving people, but we are assuming in this case that those probabilities are not there, you have to kill somebody, because the car will keep going, the car cannot stop because there are brakes.

Now of course, when you are deployed in the real system, there will be a point where you have some probability of saving versus or braking versus not braking. Maybe you can factor that in also. But then again, how does probability factor in the number of people, etc. would become even harder a question and people may not be very good at that.

If you have a car or a car on one case that you can bump into and a motorcyclists on the other case that you can bump into, who would you bump into, the car. If in one case you can bump into one motorcycle, where the rider is wearing a helmet and another motorcycle in the other case where the rider is not wearing the helmet, who would you bump into? Sorry. That is a good question. You guys can think further.

Most people will say oh, I will bump into the person who is wearing the helmet because as Parth says, there will be more probability that I can save them. But then what are you doing? You are disincentivizing people to wear helmets. Life is not very easy when these AI systems come into the real world, right. And these are questions we need to deal with.