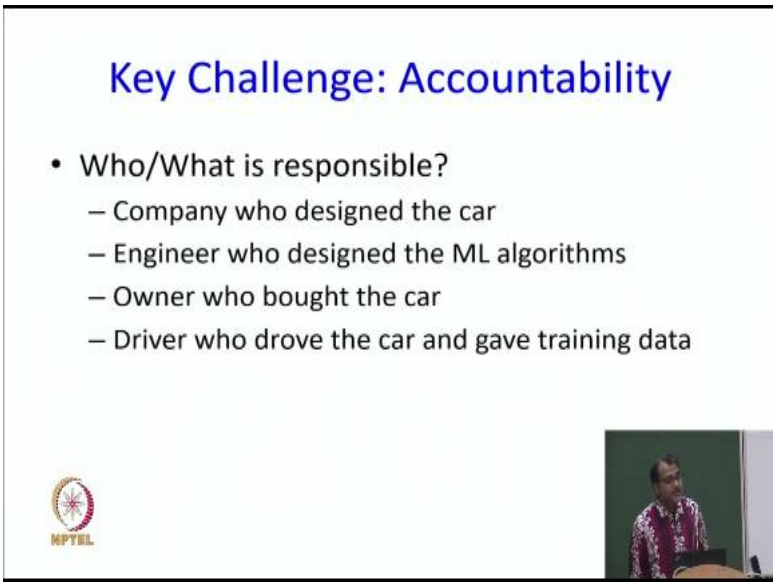
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 95
Ethics of AI: Accountability, Privacy and Human-AI Interaction

(Refer Slide Time: 00:19)



All right, who should you blame is another question. Who is responsible? What is responsible? Suppose you created a self-driving system, lot of people went into there. This is a company which redesigned the car, right? There is a company and the same company, let us say wrote the ML algorithms. There are there is a specific engineer who wrote the ML algorithms or a set of engineers. There is an owner who bought the car.

Now the car is his or her property and there is a driver who drove the car and gave training data. Now these would be learning systems. You may have a style where you go extremely conservatively. You may have a style where you can go really fast and do zigzag. You may have very different styles and the learning system needs to learn your style otherwise you will not be happy with the automated driving system.

So you drive it a few days or a few months. And the system observes you, gets more training data, fine tunes its algorithms. Now we are in the space where there is a company which designed the car, there is the engineer who designed the ML

algorithms, there is the owner who bought the car, there is somebody who gave the training data. Now an accident occurs. Who is responsible? Who should we blame?

Whose insurance should increase? And this is not a joke you will understand because most of you do not have insurances in cars. But you know 10 years from now you will all be dealing with this mess. That if I make a small error or if somebody made a small error, I do not know what happened. There is a bump in my car. I go to the insurance, my insurance jacks up and so on so forth.

So who is responsible in this case, and it is very complicated, because the engineer who may have designed the ML algorithms may say that look, this is the data that caused it. The person who gave the data said I just drived the way I drive, I never made a mistake. I do not know what the car learnt from me, how can I be responsible, etc. And so we need to even rethink the law system here.

Because until we figure out who to blame, we will not or who to be who to hold responsible, we will not be able to bring this into the world. But it is very hard to think about who to blame because we cannot just say, AI system, you go to the jail. That just does not work. Of course Tesla can say that. Okay, we are rebuilding the AI system because we think that the clouds are means white cars are clouds, you know the reference, I hope.

And so we are rebuilding the AI system so that that mistake does not happen. So over time slowly as they evolve, they will become better and better and better and make less and less and less mistakes. But still, when one human makes a mistake, we can say that you made a mistake. If it is a mistake, genuine mistake, maybe only your insurance goes up, that is it.

But if an AI system makes a mistake, it is the AI system that is used everywhere and it is not clear how do we deal with it? There is a question. Right? So that is the other question. Is it not a very dangerous practice to let the driver modify the driving practice? They should be frozen, right? And again, this is a design decision aspect of it. Every each one of us has a style. And yes, you can.

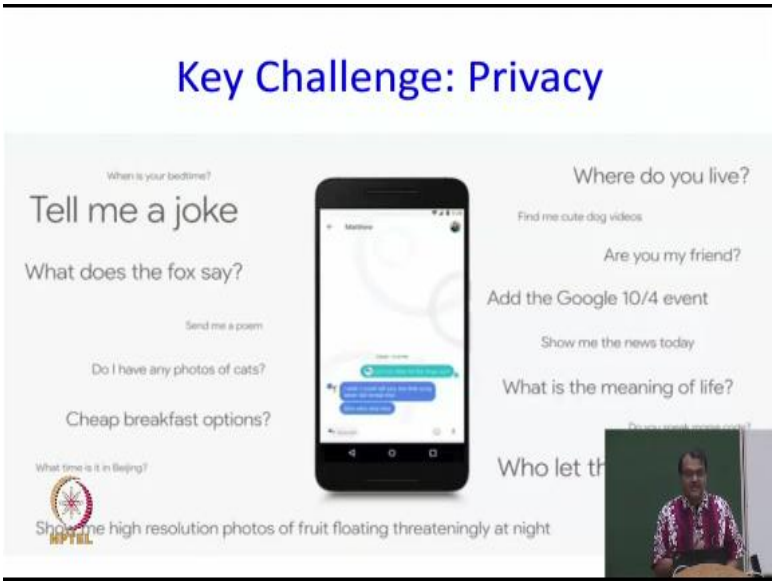
I mean, this is what happens today. Nobody puts in learning systems. But when you think about personalization systems when you think about AI systems that will be you know your closest companion. Have you seen this movie called Her? Nobody? How many of you have seen Her? Okay, this is your homework. I mean, you are IIT students anyway. You will relate to Her, at least some of you.

So please see Her. What happens is there is this guy who cannot talk to women very well who has no date who has no girlfriend and suddenly starts to befriend a personal assistant, personal software assistant. And over time, the software assistant knows him so well that she becomes a friend and more. And you should ask yourself the question, is it ethical? Is it what we want our society to look like?

But you have to first see it before you can start asking those questions, worth asking. So you will want to be in a world where the AI system understands you better than your spouse. Because the AI system hopefully someday will start making decisions on your behalf. You will give your credit card number to the AI system and say, go buy me this book and the system will go and you know do its thing and buy the book.

It will have to learn for you specifically. Now I am not saying that it will have to learn for the car situation. So you have to decide what system to put in there. But eventually you will have AI systems where learning is happening with the particular owner, right. Good questions. Then one of the challenges you need to deal with is privacy.

(Refer Slide Time: 05:12)



What about privacy? Well, we give training data, we give our data to the machine to train it. But that data comes from somewhere and sometimes it comes from a private information like for example, if you want to have a better query processing system or you know keyword search processing system, then you have to give it a set of keyword queries. And what is good or not for those queries in order to train it.

Now what did AOL do once is that it just and this is 10, 15 years ago. So you may not remember this. AOL gave a lot of query logs for public to, you know experiment on for AI to improve. So these are queries that you and I have made on Google, not Google AOL in this case, right? And believe it or not, there were people who just started looking at those queries and they were able to identify a specific person who made some set of queries to be a woman living in a certain location.

How did they do this? Well, there are phone books. There are specific queries that you make, that allows them to learn your age that allows them to learn your location, many things they can learn about you and then they can go and cross reference and come up with who made these queries. Now is it not a breach of private information? This breach became so bad that AOL had to fire the CTO and the two researchers who released the data.

So people, after at that point, became extremely conservative and just stopped sharing query logs, including companies like Google. You cannot get query logs from Google. They have it inside their company, but they will never give it to researchers. Because nobody wants bad press that came out after the AOL data leak, right. Now recently, they have trained a system where you write, start writing a response to email and the Google autocompletes the response.

Have you seen the system? What do you think it was trained on? It was trained on Gmails or the email conversations on Gmail. Somebody wrote a message. They took that message and took the follow up answer and that was the training data. But of course, they had to be incredibly careful. Lawyers were before anything else. They would make sure that you cannot look at any private information of the person.

You do not know where this came from? Who was the sender? What was the email address? You do not know names of people. But of course, if you do this completely automatically, you will make mistakes. You cannot know exactly all the names. You cannot know all the named entities. They just so they have to be very careful, they cannot release these kinds of training data.

So you want to make sure that you can train machine learning systems. Like we want to train machine learning system where you give it an X-ray and it will tell you what disease you have. But where is the system be trained by? What is the system going to be trained by? Your x-rays. Now is it not private information to most people? Yes, everybody thinks their health information is extremely private, right etc., etc.

So here is a fun story. Targets figured out that a teen girl was pregnant before her father did, right and in this case, I am not sure what is right. But the story goes, that target sent a lot of coupons for baby products and this, that and the other to an address and the father took it and said, what the heck are you sending my daughter? She is still in high school. And the manager of target was visibly embarrassed.

And he said, no, sorry, you know we will fix it this that. And then later by chance, because the manager really felt guilty, called the father again a few days later and said, I am very sorry. I hope you know you do not take it personally. And the father said, No, it was my mistake. My daughter is really pregnant. I did not know. And actually, it is not very surprising.

Google knows what disease you have before your doctor knows. Because what? You have some symptoms, you first Google it to figure out oh, what could this disease be? What should I do? You know I am puking and coughing and blah, blah, blah. Okay, what type? Google knows flu outbreak or twitter knows flu outbreak before anybody else knows. Why, because whenever you have flu, you tweet.

And when lots of people start tweeting from a region, you know there is a flu outbreak there and then you know that it can you know go. So now you are careful about the surrounding areas and then you can see how it was. So all these companies

that you are interacting with on a daily basis know more about you than you can possibly imagine. They know things about you, you may not know about yourself.

And right how do they use this information becomes extremely important. Now I want to say here, that a lot of the questions that we are discussing, some of them are related to the AI system specifically but some are related just to which products are kosher to build and which are not, which are ethical products and which are not. And this question is not a general question. It is a very case by case question.

Do we want automated doctors? Do we want a doctor that you know talks to you and tells you, you should take this medicine, which is an AI system? Do we want it? When you want to answer this question, you have to take the whole sociopolitical story into account. Imagine the villages of our country, which have no access to healthcare. If there we put in this AI system, and it makes five mistakes out of hundred.

Is it better than not having a doctor or worse? And you can always say depends on the mistake or kinds of mistakes it makes. Maybe I am not willing to have the doctors do surgery on me just yet. Maybe I am not willing to have the doctor prescribing medicines which can have catastrophic side effects. Maybe I do not want this AI system to give me medicines for life threatening diseases yet.

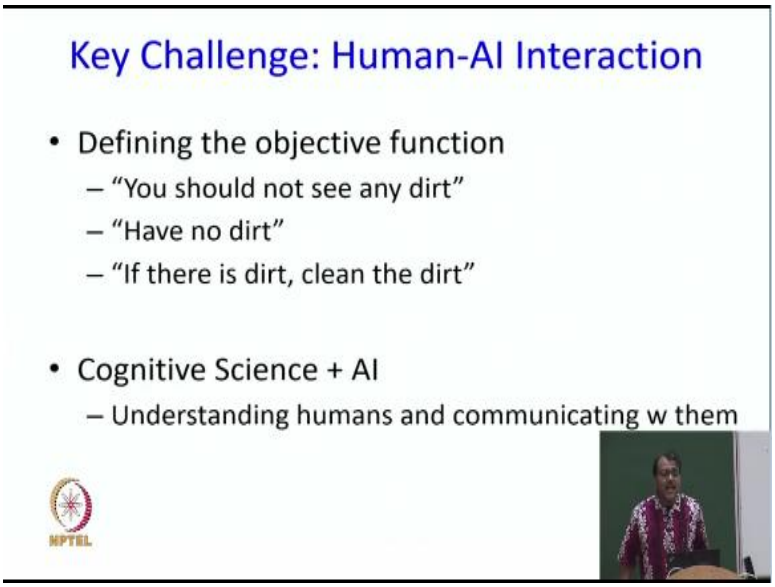
Maybe we should have a pipeline where the AI system says you have flu take this Crocin. You have, you know a common cold, do this. You have small disease, do that. And you may have cancer, please go to the doctor. It is possible that the AI system does that. It does that triaging for you. Right. But in a different scenario do we want an automated utensil cleaning and a house cleaning robot?

Now people in the US absolutely do. Because the labor is extremely expensive. You know they have to pay somebody once in two weeks to come and do a massive cleaning. And for that one and a half hours of work, they charge like 150, 200 dollars. It is a lot of money. In India, things are opposite. Too many people and the labor costs are low. We may be happier in some ways with an automated utensil cleaner, like a dishwasher or whatever.

But we do not have many dishwashers in this society. Dishwashers as an automated dishwasher today. We can, some people do. But we prefer to, you know work with the maid or the cook, to please clean the utensils because you also realize that we can give them some money in the process. And that sort of helps in the larger good.

So you have to really think about which AI system makes sense in what scenario taking the whole sociopolitical situation into account, socioeconomic situation, not political sorry.

(Refer Slide Time: 13:10)



And then you will have Bots that will work with you. Not only do they have to communicate in your language, they also have to understand what you said. And if we give them incorrect goals, then they and then they do the wrong thing for us, then who is to blame is a big question here. So let us say we have a vacuum cleaner. Bot we tell the bot, hey, when I come from office, you should not see any dirt.

And when I come back from office, the robot closes its eyes because the robot does not see any dirt anymore. So we say no that is not what I meant. Have no dirt in the house. So the robot throws everything out of the house. Because now there is no dirt in the house. Did they satisfy the role? Absolutely. Then you say no, this is not how I wanted you to do.

If there is dirt you clean that dirt and only then I will give you some reward for cleaning the dirt that exists. So the vacuum cleaner bot goes around and you know cleans the dirt got a lot of reward is very happy. And then there is no more dirt to clean. So what does it do? It just throws the vase on the floor. Now there is dirt to clean. Now it can you know clean that dirt and get its reward. So you say dude, you cannot create dirt.

You cannot throw the vase, okay. Okay. So I will just put my leg in front of you when you are walking so you will fall down. When you will fall down, you will hit the table. When you will hit the table, the vase will fall on the floor. If the vase falls on the floor, it will break. Then I would have created dirt, I can clean the dirt and get my reward. So the issue is that when we start to design these systems, you and I have a lot of quote unquote common sense.

We live in a physical world, we have shared context. We have made you grow up you know slowly. And by the way, these shared contexts change completely. It is funny how these things change. So one of interesting stories is know there is a tonic. And the ads was pictorial and the pictorial ad was, you know you were lying down on the bed. You drink the tonic, you rise up and you start running.

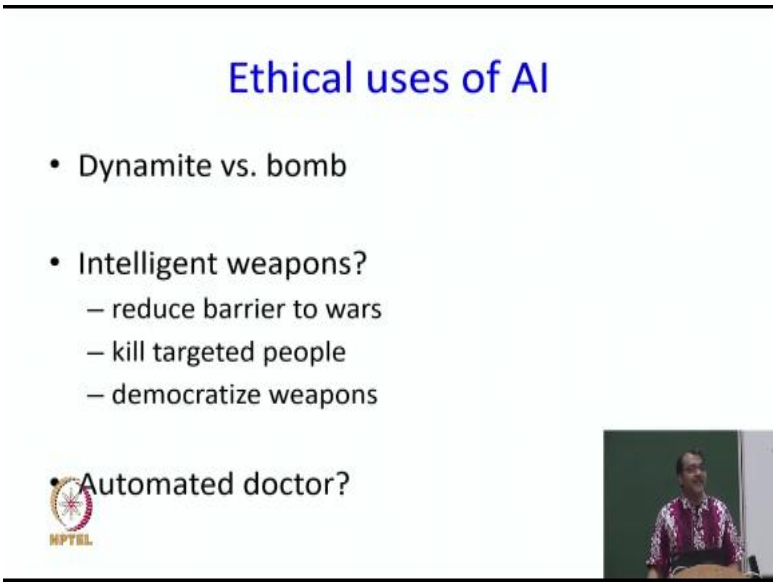
Good ad, it was doing good sales. This tonic ad went to an Arabic country. And there were no sales of the tonic. Can you guess why? Because Arabic languages are read right to left. So it read, you were running around, you drank the tonic, you go on the bed. So what shared context you and I have is what you and I have, we understand each other. Because we live in a certain society.

We have developed in a certain society. If we have AI systems working around us in the speak that they have, in the language that they have, and we are trying to communicate with them somehow, and they have only a fuzzy understanding of what we mean this can lead to lot of side effects that we do not want, right?

And so there is a big area of research that has come out in the last couple of years, it is called human AI interaction or human in the loop AI. Extremely important because

the AI systems need to understand human model, and also to some extent the humans need to understand the AI systems.

(Refer Slide Time: 16:59)



And I have already talked about ethical uses. But here is the question that lot of people ask. You know if we have we are on the process of making intelligent weapons. Imagine these intelligent weapons you will not fire a gun randomly in the air, the weapon will release a bullet. The bullet knows who it needs to kill. The bullet will go around finding that person and when it finds that person boom, killed.

Let us say we could build such an intelligent weapon. Is it ethical to use it or not ethical to use it? Right. Now this is a difficult question. There are many sides to the coin. One sided is that well, at least I am not killing random people I am killing targeted people. The other side says yes, but now since you yourself are not going to the war, you are just firing it from somewhere and then the bullet is doing the rest of the thing.

It has reduced the threshold to go on wars. So now you can go onto a war because there is no risk to you. Usually we avoid wars because we defeat the other country but we also kill our own, right. Now we do not have to, that is not very good. And then what one could say is okay, but guns and these heavy machineries were very hard to find. But now AI system is what, it is a software.

We put in some hardware right, some simple hardware. Most of the magic is in the software actually more often than not. That would be easily exchangeable. People may hack into your system and get the software and so everybody will have this intelligent weapon. And so then anybody who is unhappy with their neighbor will justify fire it and then the bullet will go and kill them. Not very good, right?

So there are a lot of questions that you need to ask before you can think about ethical uses. AI folks on the other hand, try to remain a little aloof on specific ethical issues that exist in specific domains. That is for the domain expert to figure out. See, a domain expert has to figure out that here an AI system makes sense. In an education setting, we do not want to replace the teacher.

But if you can do automated grading, then the teachers you know gets helped. And that is good. I would love an AI system that grades all your papers. I do not like grading. But I do not want them to come teach, because that is the part I enjoy, right. And if you could learn all the AI sitting at home, then that is not fun. Right? So as we go into an era where we will be interacting more and more with AI systems, you already do.

Google is an AI system at the end of the day. There are many AI systems that you do not recognize that you work with. But you will have more and more of those. You know the face detection system on Facebook, which automatically creates a boundary around the face and says do you want to tag your friend here? That is an AI system at the end of the day.

So we have to ask all these questions on an individual case by case basis and study the ethics of AI much more for AI to be super successful in the future. So I think this is sort of a good point to say that you know this was a good, this was a course that came together to some extent.