FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 08

Lecture 36

Lecture36

Good morning. Continuing with what I was talking about in last session Responsible AI Ethics and Digital Trust. In this session I will talk about how to manage some of these problems managing ethics and bias. I will cover briefly how to manage ethics and bias, the practical guide to ethical AI, take actions to mitigate ethical risks and managing ethics. This is the practical aspect of what I was talking about in the previous session.

Managing ethics and bias. Consider some real life examples of AI failures because AI can fail. We will start with that wherever things AI can fail and what can be the consequences and what we can do. A robot designed for grabbing auto parts grabbed and killed a factory worker. These are all real cases; this happened in Germany somewhere.

Image-tagging software classified Black people as gorillas. Medical AI classified patients with asthma as having a lower risk of dying from pneumonia. AI designed to predict criminals acted racist—Black and White. It is a very common problem in the US. AI judged a beauty contest and rated dark-skinned contestants lower.

A self-driving car had a deadly accident. And every day we are experiencing shortcomings of AI, GPS proving providing faulty directions often causing accidents.It happened last year once in Andhra Pradesh during a flood etcetera the car drove into a it was a road, but then it was flooded with water and then it drove into the water because GPS kept on saying that it is a road, but it finally ended to a river and the people died. Machine translations give incorrect results that we know we also have seen biometric systems misrecognize people etcetera even facial recognition systems. It is difficult to find examples of AI that do not fail.

It is very common in most AI applications failures are there. You have to live with the failures. So, to work with algorithms will go wrong.So, have a plan in place what we say is the mitigation plan it will go wrong we know that some things will fail, but are we prepared to handle those failures which we must as responsible professionals we should be prepared. So, system designed to do something will sometimes fail to do that particular thing.We cannot predict that based on this generalization that it can produce something which

was not desired to produce a doctor will misdiagnose some patients in a way a real doctor would not or a video description software will misunderstand the movie plot. Employee screening software will be systematically biased during the hiring, I am talking about that and tax preparation software will miss important deductions or make some inappropriate assumptions of a deduction or whatever. So, ultimately calculate an incorrect tax value. Such examples can be there plenty in our life, but then we need to put some best practices in place, such as controlling user output to the system and limiting training to verified data input.

So, it comes back again and again to verified quality—the quality of the data, the verified data inputs. Am I using those data? If you use good-quality data, then most of these problems can be taken care of. Checking for racial, gender, age, and other common biases—some biases are quite common. So, you can easily detect that. Some biases are not so common. We already talked about that in the previous session, but some of the major ones are actually not very difficult to detect.

And analyzing how the software can fail and providing a safety mechanism for each possible failure. So, do I have a Plan B in place for anything that fails? Can I have a Plan B? I have a communication plan in place to address the media in case of some embarrassing failure. So, if there is a major failure—many customers are unhappy, or it is used for the public, many people are unhappy—then you must have a communication plan already in place.

Do not wait for things to go wrong. I am repeatedly saying this. So, the governance committee should be ready with a communication plan—if it goes wrong—because you can always preempt that things can go wrong. When it goes wrong you are ready with nonsense. That is very important that timely intervention. If you do not immediately respond, if you keep quiet for maybe 1, 2, or 3 days, then that becomes worse and worse.

You should also keep in mind that with increased use of AI, and cost of training and developing models, AI failures are likely to increase. The more and more we develop new models in AI, the cost of training is going up. We will try to have some shortcuts for developing the models. The failure rate also can increase or number of failures, not the rate that is the number of failures.

In 2019 there was a case in Los Angeles the state government federal government Los Angeles sued IBM for allegedly misappropriated data it collected with its ubiquitous weather app. So, whatever data it was collecting from its weather app it had used it elsewhere misappropriation. Company called Optum was investigated by regulators of creating an algorithm that allegedly recommended that doctors and nurses pay more attention to white patients than to sicker black patients. So, these are all medical insurance related thing. Goldman Sachs was investigated by regulators for using AI algorithm that allegedly recommended

discriminate against women when giving loans. So, it was preferring men or males for disbursing loan. If the 50 males or 50 women had applied for loans, they would give to probably all say may be 40 out of 50 for males and may be 20 out of 50 for females. This is the kind of bias which got detected and companies get sued. You have to be also aware that whenever using anything for such you know AI tools you can possibility of getting sued for these reasons.

So, what do you do? You identify existing infrastructure that a data and AI ethics program can leverage. So, one is of talking about the governance board having a data governance board as part of IT infrastructure, the executive level drives the seriousness of the issues. So, when you have a governance body senior executive C level people sitting on the governance body then the development team will be more serious. They will be when they know that they have to report

on a weekly basis or a regular basis of what they are using in their experiment or development product development, they will be more conscious about it. They would not be very otherwise people become very tend to be very casual fine nobody is you know there to see whatever I get I use and I want to be fast quicker easier etcetera I take shortcuts. But once you know that senior people are there watching me then you would not take shortcuts Data and ethics strategy needs to be orchestrated with general data strategy and this is done at the executive level. So, what is the data related strategy?

What data to be used? And protecting the brand from regulatory reputation and legal risk is ultimately a C suite responsibility. So, if something goes wrong about the organization, so the C suite is responsible. So, when you know that you will get involved right from the beginning, you would not delegate or leave it to a junior executive. The C level executives must sit on the governance board.

Any high stake issue should be immediately brought to the notice of the C field. So, something you can say you get a sense or feel that something is has gone wrong or is going to get wrong. Immediately highlight that, immediately bring it to the notice of your highest executives, do not try to hide it because it will come up sometimes later and if you hide it then the more delay you do the more damage you can that can happen. It may be advisable to have external subject matter experts also in your governance team, because you may not be adequately expert in say AI related technology of course, your subject you will know it, but still if you think having an external SME will help will strengthen the team, please do so. Create a data and AI ethical risk framework,

which is tailored for your industry. So, all data is very industry specific. So, you need to do also things for tailor governance part also consider the industry factor. So, articulation of the ethical standards, identification of relevant external and internal stakeholders, who are my stakeholders, who are going to get impacted when this tool is being used. Then recommended governance structure.

And then establish KPIs and quality assurance programs to measure the effectiveness of the tactics employed. Finally, things have to be need to be measured. So, have this KPIs ready. So that you know what is to be measured to see that things are going well and the right data is getting selected or the wrong kind of data is not getting selected. Then indicate how ethical mitigation is built into the operations giving details of the

So, right from beginning the mitigation thing should be embedded into the overall tool development program, so that if something can go wrong then some backup plan will immediately need to happen. Ensure the processes are in place to wait for biased algorithms, privacy violations and unexplainable outputs. So, somebody has to wait the process should be there the checkpoint. Somebody to red flag that this is not right this data is violating doing some privacy evaluation or this algorithm is doing some output of the algorithm is giving some again some biased report or privacy evaluation etcetera or something which cannot be explained. Sometimes it happens with AI is that many of the outputs cannot be explained they are not understandable by human.

Things which we cannot explain. So, that those should be obviously, be removed or taken care of and to ensure. Go back and try to check which data probably might have been causing this and remove that from the training part or bring in new data etcetera. Frameworks will need to be tailored to industry specific requirement like finance, healthcare, retail each of them industry type will have their own requirements. Governance requirement, privacy requirement, customers requirement etcetera, data security etcetera.

All these are different types of data, different types of customers, different types of requirement, security requirement etcetera. It varies from industry to industry. Your framework will need to be tailored according to your particular industry. Change how you think about ethics by taking cues from the success in healthcare. Why healthcare?

Because healthcare is an industry where ethics as we know plays a very big role. It does in all kinds of industries, but healthcare is a very specific one, many several issues crop up, individual privacy is very much involved. So, we can take that industry and say compare try to compare with what how we benchmark that how we stand vis-a-vis the requirements of a healthcare industry. Many feel that AI ethics is a bit funny and sufficiently not concrete to be actionable, which is true I mean sometimes we feel fine, but what can I do about it. I do not have much control over whatever say data I took from somewhere some data set etcetera.

I really cannot control all my data. In the healthcare industry has taken been deeply involved with ethical risk mitigation and some of those examples like it is an essential tenet of respect for patients is that are treated only after they have granted informed consent based on the full understanding of the risks and consequences. So, how do we start an operation? So, the patient or the patient's relatives or parties will have to be informed in detail about what the operation is going to be, what are the possible risks. For example, the age factor or the health condition etcetera many things have to be considered before finally, deciding to whether to go or not go for the operation.

But once you decide to go then the doctor's responsibility is to tell the patient and the patient's family that details of the operation what it is he is going to do and what could be the possible things which can go wrong. And before every operation the relative of the patient has to sign a bond. Why do the hospital authorities make them sign a bond? That is to take care of the legalities. When you sign a bond it means that you have understood all the implications of the operation. Whether actually I have done or not that is a different story, whether actually the hospital or the

doctor has explained to you or not in details that is again on the different story varies from hospital to hospital, doctor to doctor, country to country, region to region and we have you know we have lot of horror stories for that. But the basic design principle behind signing a bond is that the things have been explained to the patient party and they know the risk behind the operation and due to the operation something can go wrong. So, they are aware of it. Or they were made of it. The same principle you can think of when you are using developing AI tool.

Tell your customers in advance that educate them that we are proposing this tool for disadvantage, for your benefit, for your ease or for your better customer experience etcetera benefits. However, this is a tool, not a human being who will be responding to your query. The tool has been adequately trained. However, no training can be 100 percent correct or perfect or no software IT product can be tested for all possible variants, it is impossible.

So, there could be some failures and when you are using it you should be aware that yes there could be some failure. So, whenever you install or you know any software you have to sign finally, tick mark you know I accept etcetera. And there will be a huge long list of you know points all written in very small prints which you probably hardly read or understand also probably because you are not a software expert now you are a legal expert. But finally, you have to sign accept otherwise it will not allow you to install the software.

So, once you have installed the software means you have accepted the whatever terms and conditions or whatever the disclaimers have been given by the software developer or the company. Later on you cannot come back and tell see I was not told or I was not aware etcetera. Let us say you had read it and you had signed then only you could install otherwise you would not have been able to install the software. So, it depends on how you want to handle your product you want to make a long list pages very small prints. So that people are not going to read or they get bored and they find

they come they want the tool find all the advantages and they just accept or you can do it in a much better way or take them into confidence spend some time maybe give some more documentation online or real time. So, that they can read and find out more information about the tool, how it works, how it was developed etcetera and that there are chances that. So, they should kind of not really blindly trust output of the tool, use it with some element of doubt. If you need to have some doubt then cross check with a human being or a human operator or one of the sales services. You have you should mention that when you are not

convinced with the response or you are not satisfied with the output the task done by the tool, contact customer service care or give some phone number or a chat bot or a mail id or whatever. So, that you can send your problem to somebody to analyze and tell you what could have been done better or whether whatever has been done has been done right or not. The same requirements can be applied on how patients data is collected used and shared. So, this is very important how you are collecting data and how you are storing it and how you are using it and how you are sharing it. All of these things you have to consider the ethical angle and the privacy angle.

The bottom line is to break down ethical concepts like privacy bias and explainability into infrastructure processes and practice that realize those values. This is what you have to ensure privacy bias and explainability and you have to do it through the right infrastructure, the right process and the right practice which will help you to realize these values. So, that is why governance comes in a big way because governance decides the processes, the practices, and even the infrastructure, as that is where investment and budget decisions are taken. Explainability in AI—so that is another interesting topic. It refers to the ability to understand and interpret how an AI model makes decisions or predictions.

Making the process transparent and comprehensible to humans—how did it come to take this decision? The step-by-step analysis. So, the logical flow—did it think logically, like we do? We think logically. When you want to travel immediately from here to somewhere there, then I have to start thinking I have to book a ticket or train ticket or book a car or an aircraft or a plane ticket. I log in and see train tickets available, if not then what is the alternate plan etcetera. You go step by step by step and then I have to apply for leave and things like that and I have to pack my bag.

So, all of those things. This is crucial for building trust ensuring accountability and identifying the bias or errors in AI systems. Explainability AI it provides aims to provide insights into the reasoning behind the output often required in high stake domains like healthcare, finance or legal systems. So, legal systems—yes, extremely important—how it comes to the conclusion. If the outputs are subject to regulations that require explanations—for instance, regulations in the banking industry that require banks to explain why someone has been turned down for a loan.

I applied for a loan it was refused by the tool. So, the tool must tell me why my application was refused. Same is true say like when we do say examination or admission procedures in education system somebody does not or a job for the let us say admission somebody does not get admitted. So, they often send out an RTI, they would like to know why I was not selected, why my interview performance was so good, the written test was so good etcetera. And then the university concerned will have to give an explanation that what went wrong or what was not found for the candidate not to be selected.

Doctors can verify the models focus aligns with medical knowledge increasing trust in the diagnosis. So, the doctor should study the tool I said whether the tool is aligned with the known medical practices. Then the output will be as per the known medical practices and not something very awkward. And, the doctor will trust the diagnosis output of the tool. So, the doctor needs to understand the tool itself how it is the explainability function, how it is deriving coming to that conclusion.

In financial institutions they use a tool called SHAP to explain fraud alerts to customers or auditors ensuring the clarity and fairness. Suppose there was a fraud in detected then it has to explain why it was a fraud and why it is saying that this is a fraud and not a genuine transaction. Because I am a customer I did a transaction and the tool tells that you did a fraud transaction. Then I need an explanation from the tool and the tool should be able to give me that instruction. Build organizational awareness, cyber risks have become of prime importance to anyone who works with data and AI products.

We all see news regularly said educating and upskilling employees and empowering them to raise important questions at crucial junctures and bring key concerns To appropriate deliberate model. So, how do you prevent that? Build this culture, educate an upskilled employees, so that they get empowered to raise important questions, they know when which question to ask and when to ask that question, what signal triggers a question at the right time, so before it is too late. All of this goes to develop a culture in which data and AI ethics strategy can be successfully deployed and maintained.

This is at the organization level also formally and informally incentivizing employees to identify ethical risks. If you incentivize them if you do not incentivize them on the other side they will ignore, but if you incentivize them that is a way of getting more policeman actually little to be saying to give an analogy you are getting more policeman to do the policing job for you, because it can be cannot be done by a one or two or three four people from your IT team or serial team. So, have your entire employee pool trained in this basic the AI tool what we supposed to do and what is not supposed to do etcetera. So that they can detect anomalies and they are incentivized to report these anomalies.

They will keep their attention they keep the eyes and ears open. And then you have to look out for anything that goes wrong. So, there will be a big strength, a big police force for you. Monitor the impacts and the engage the stakeholders. One was creating organizational awareness, the ethics committee and the informed product managers, engineers and data scientists all in the part of the involve them from the right from the development process.

So, they know what is being developed. Limited resources time and a general failure to imagine all the way things can go wrong it is important to monitor your data and AI products used commercially and involving customers. AI products can be ethically developed, but unethically deployed one example is for example, airbags in a car it does not guarantee safe driving at a speed beyond 150 kilometers per hour, but do we know we do not know. Of course, maybe in India we do not drive at such speeds, but in the other countries you can drive plus with all the new highways which are coming up in India people are driving at higher speeds. So, have they been told that your airbag will not actually save you at speeds beyond a certain level?

So, it is an ethical product, but it is deployed unethically. So, well-intentioned algorithms can have detrimental effects. A classic example of this is Airbnb, which developed a product to help homeowners attract more customers. It was observed that white homeowners in the US were getting more customers than black homeowners. When they conducted research, they found that white homeowners were using this AI tool—which was voluntary, not mandatory—much more than black homeowners.

So, they were benefiting. So, it was a kind of bias, but an unintended one, because it was a free tool—optional, not compulsory. So, they explained that white owners used the tool more, got better results, and secured more business, while black owners suffered. So, what is the solution? The solution is education.

So, Airbnb could need to encourage the black hosts to adopt the new algorithm by rewarding them in some way or explaining to them the advantages of the tool. So, either you tell them the what are the benefits you are missing out guys use this etcetera or even you can initially incentivize. Something very similar to what we do incentivize or say our caste system here the lower caste, the reserved candidates we give them incentives like scholarships, priority in admission etcetera. So, that they can be made to go for higher education etcetera and better jobs etcetera. So, they can come into the real position.

where they can be at par with the rest of us that is again inclusivity example of inclusivity. So, here something very similar you can since you can find out most of the black house owners they are not maybe they are not you know smart or literate enough etcetera. So, they are avoiding all those things. So, incentivize them if you use this tool I will give 1 dollar more or 2 dollars or more etcetera. So, once so, the best thing in works in life is for people is monetary incentive.

So, focus on building trust and help users understand what the algorithm is meant to do and how it works and also incentivize whenever possible for increased usage. So, this was this Airbnb case. Take actions to mitigate ethical risk. Who needs to be involved? Data scientists, legal complex experts, ethicists and senior business leaders we have talked about that and the main work for this group is to identify the source of the risks generally for the industry to which they belong and of course, and then the company your organization in particular.

Managing ethics, see in this box it is the responsibility of the user to ensure that any text generated by the model does not infringe on any copyright laws. So, what it giving a disclaimer as a language model chat GPT does not have the ability to infringe on any copyright. However, it is possible that the text generated by the model may contain copyrighted material. So, responsible user to ensure that whatever text is generated use it carefully do not use it just like that because it might infringe some concrete loss. Here if you see the same thing example I talked about the education system here is this a student's work ICB for me or just a readymade paid for download from the internet.

It is a question starting to bother lecturers and one that one go away. As I said it is here permanently sustained and might increase. What is the AI Bill of Rights? As I told earlier that many such bills are required to regulate the industry, regulate this AI technology. So, one of those is a US government 2022 brought out something called AI Bill of Rights.

It comes up with five principles that should guide the design use and employment of automated systems to protect the users in the age of AI. So, these are right to safe and effective systems, protection against discrimination by the algorithm, protections against abusive data practice, right to know and the right to opt out. So, if I have a choice I can opt out from the AI tool and prefer to have interact with a human for example. I should have that right.

So, I should not be forced that you can only do through the use of the AI tool etcetera. Just a small cartoon towards the end. So, excited about this first art sterling plagiarism machine encoded see it steals art you can copy the art somebody's famous Mona Lisa you can copy and you cannot make a difference between which was the original which is a duplicate. And now we can have art that looks exactly like it was done by any artist without having to pay for that artist and then say everybody knows that artist get paid too much

then you pay too much for the art may be a crore million dollars etcetera. So, let us copy it and you know whatever do some business. And then say next we go after the programmers. So, once you have solved this thing that we can create copy art probably now let us change the programmers to do something else do something else another sort of inappropriate thing. So, with that I come to the conclusion of the regulated AI, regulation AI and responsible AI section in this course of study. This is a very important aspect though it is not too technical, but the governance part, the responsible part and the legal implications,

or the damages it can do or it can do to you and as an individual and to the organization which is most important for you. So, you should be very extremely careful for whenever you are developing any tools, think about the organization, think about what damage it can do. So, have the trust in the governance, ensure that the governance committee is there right from beginning, this is very important so that it can oversee the entire development process. So, do not ignore any of the development process especially the data part. So, 90 percent of your effort should be on having an oversight on the quality of the data.

This is a list of references I have been using for this module. You can go through them for more knowledge, plus much information is available on the internet. Thank you very much.