FOUNDATION OF DIGITAL BUSINESS

Surojit Mookherjee

Vinod Gupta School of Management

Indian Institute of Technology Kharagpur

Week 08

Lecture 35

Lecture 35 : Common AI Ethical Issues

Good morning, today I will begin module 10 which will be dealing with Responsible AI ethics and digital trust. So, we will discuss some common AI related ethical issues, 5 common AI ethical issues and then we will talk about generative AI responsible innovation and the law. As all of you know that this AI business is a big talk going on at a parallel level. So, what is the development part?

New technology coming in and adoption of the new technologies and scaling up of the AI tools in various operations of various functions of various industries. That is one part of the story. Now all that the popularity, the success of this whole business of AI has challenged many things and one of them is our ethical issues, privacy, racism, bias etcetera. So, these are all being talked about heavily.

So, what is the impact it is going to have on all of these aspects of our human life or career. So to say that about a couple of years back or maybe 3 years back, I think that time Stephen Hawking was alive and Elon Musk, Stephen Hawking and about 100 such senior scientists and executives, business leaders had framed a letter signed by them stating that AI research should be stopped for 6 months. Now why did they say that? Because they were getting worried especially when talks were going about say super intelligence, so super AI.

So, then that was what people are very afraid of that when a computer becomes equal to a more intelligent smarter than a human being, what can it do or what will it do or what are the potentialities it will have to create whatever controls damage etcetera to the extent that people even Contemplate that it can lead to the extinction of the human species and

being taken over by this world of smart, intelligent computers. Far-fetched, but still people are talking about it and people are dreaming about it and current research in Gen-AI going on, you realize that it is gradually progressing toward that AGI.

It is a general intelligence—not that a computer can defeat a chess master, play chess, and then cannot do anything else. So, we are now moving beyond that. We are trying to train computers such that they can do several things. That is general intelligence, not narrow, specific intelligence. That is when it starts getting more and more powerful. So, what this group felt was that research should stop?

The whole purpose is to slow down, review, check, and generate a common opinion on what we should do as a strategy. So, we should limit ourselves to certain things and not cross the limit—like a nuclear bomb, if you take that comparison. Today, there was even a threat a month back that on May 10th, specifically 2025, there could be a potential nuclear war in our subcontinent. So, that fear is always there—that even countries like Pakistan or India could get into a nuclear war. This is what AI is compared to—the damage it can create and the power it has.

So, anyway I will not talk about that because that is as I said is bit too far fetched, but we will get into hand some real issues some real common a business manager, executive, or professional, you should be conscious and aware of. And try to take care of those things in your work or even in your social life, etc., when you are interacting over social media. So, bias propagation is number one—racial, gender, and socio-economic biases. Biases can be anything; I will talk about that later. Any number of things can be a bias. The common ones are of course, racial, gender, religion, socioeconomic etcetera.

There are certain recruiting tools examples use cases Amazon was one example where it could selectively not select certain category of people. So, in that case it was a black women candidates were not preferred and there was lot of hullabaloo about that whole thing. If the facial recognition systems are infamous for disproportionately making mistakes on minority groups or people of color. The error rate for lighter skin males was no higher than 1 percent. However, for darker skin females the mistakes were much more significant reaching up to 35 percent.

So, if you compare white lighter skin male versus dark skin female this is the error percentage 1 percent versus 35 percent. What is the reason? The reason is of course, the data because you are training your module on a data and your data set was biased as

simple as that. So, how do you know your data set is biased or not biased? AI systems today are only as good as the data they are trained on.

The data is not representative skewed towards a particular group or some imbalance the AI system will learn the same and produce the same. So, using a snapshot of the web to train models can mean you have learned the biases in that snapshot. So, be very careful when you choose your dataset, when there is limited oversight in the quality of data is to monitoring various biases are bound to happen. Plagiarism, because the tools are trained on available literature, available works of art, works of creativity. scientific literature everything whatever you take it is trained on all those information.

So, it is actually learning from whatever is there and there is a high chance that it will produce directly from that content. So, that amounts to what we call as plagiarism. If I want to write a paper on particular topic and seek JNI's help that JNI may find out similar material about that topic on that same topic and produce material copying it from that exactly. So, that becomes straight away plagiarism. We are seeing this issue in a lot of artworks, in scientific literature, in various literature, etc., publication because

it is such an easy tool to have and it is so commonly available freely available to everybody and it is very difficult to resist using it. So, we use it and then blatantly to save time effort, etc., maybe use whatever it has produced. We do not bother to go through it, edit it or modify that or do not want to give that effort anyway. The question is who exactly owns the copyright of the Gen AI. So, many issues are coming up many people are suing say open AI companies like open AI all these Gen AI owners etcetera that because of you have used my whatever copyrighted data.

So, I can sue you for copyright violation. So, this is becoming a big issue. leveraging web and public datasets for developing models can result in an unintended plagiarism. However, due to little AI regulation world currently as we have it is not very regulated environment, we do not see any enforceable solution in the near future. So, this will need lot of studies, lot of regulations, lot of discussions at the government level, political level,

country level and scientist level, user level, the AI scientist level to come up with some reasonable solution to prevent such misuse of copyrighted material. Technology misuse is the defect which can create videos, images, morphing, I can morph anybody's head or anybody's body and etcetera and can put words in your mouth, false speech, hate speech or whatever speech. And AI event generator tools like DALI and stable diffusion can be used to create incredibly realistic depictions of events that never occurred. You can create

a scenario for example, a crowd problem somewhere you can create that and publish it, false it was it did not happen at all.

So, intelligent tools can be used as weapons in a war to spread misinformation, propaganda, to gain political advantage, manipulate public opinion, commit fraud and what not. So, these tools have given us a very powerful weapon so to say. So, while the business or technology is not exclusive to AI because you can misuse many technologies, but the AI tools are so adept at replicating human abilities. And it is also so easy to use anybody can use it you do not need any training so to say not to that extent any person can learn that use those open source tools and create whatever they want to create.

The abuse of AI could undetected and give a lasting effect on our view of the world. It creates an uneven playing field. So, algorithms can be easily tricked or you can game the system. So, the AI's decision making process in whatever it is being used let us say hiring as a use case your recruitment. that that particular tool or they are using some tool and they have some for affiliation for some particular keywords or when you search for some features in a CV for a particular job.

So, in advance in your CV you can intentionally put those keywords in many places so that your CV gets selected or getting ranked highly in the Google's search engine. you know SEO search engine optimization. So, that also can be tricked, we can game the system such that my website gets a higher rank. So, people know what we using this search things of SEO that what can make Google attracted or get higher scores for the website. It could be through programs, it could be with a certain keywords, it could be certain images or it could be how it is structured.

So, various techniques can be used so that that particular website gets higher score from the SEO of Google. So, gaming AI algorithms is one of the easy way to gain an unfair advantage in wherever in your business career, in your influential ship and politics. People who figure out how your algorithms operate and makes decision can abuse and game Spreading misinformation that is the easiest thing you can do. Without proper citation it is easy to verify facts and decide which answers to trust and which not.

So, I ask something a prompt and it gives me answer, but I do not know whether it is right or wrong. So, it generally tend to accept, but we know also that they hallucinate. So, that is also quite well known. So, we need to verify or vet the output of a tool. If you want to write a script for example, so you can employ articles written by say ghost writers and you can get it even published.

For example, they get the story written in your name, but actually it has been written by some AI tool, but you have not been able to verify whether some of those parts etcetera have been paraphrased from some other stories or written by some other authors. So, that becomes a copyright violation, but you would not be knowing because it is not possible for you have to read all those books and remember what exactly those paragraphs were or chapters were etcetera. So, over reliance on AI generated content without the human verification element of the facts will have might create problems, but again it is not.

as I said in some cases to verify because if you want to write a story and then suggest some prompts and it creates a story, but you never know whether it had picked up certain portions from other already published stories. So, difficult to verify this. So, one of the best solution is to avoid genii to write stories, if you have to write stories write it from your brain using your brain not do not use genii. So, in summary the five real ethical implications of AI, bias propagation, unintended plagiarism, technology misuse, uneven playing field and widespread misinformation.

They can propagate racial, gender, age and socio-economic biases, they can infringe on copyright laws, they could be unethical, they create unlevel playing field because somebody knows how to game and people do not know how to play the game and then the trusting answers blindly from AI systems can cause widespread misinformation. With that I move into for topic of generative AI being responsible innovation and the law. So, development and deployment of generative technology raise legal and ethical concerns because of obvious reasons, it is so powerful, it is so human like.

Now put the responses that we tend to use it to and it might cross the boundaries. Responsible innovation starts with inclusivity. So, I will discuss some few principles of what is made by responsible being responsible innovation. The idea is that you innovate. So, you try out you experiment obviously, that is why this tools are there, they have to be used for your business benefit for your advantage etcetera, but do not ignore certain factors.

So, one of them is inclusivity, we generally tend to tend not to think about these things, hence I am talking about this and showing it to you. So, it must be developed in a manner inclusive of all communities. including those who are traditionally underrepresented in the technology sector. The development process must take into account the unique needs and perspectives of diverse communities to ensure that GNI is not discriminatory or biased. Now, it is easier said than done.

I am writing in something or for a to be used by certain customers, but I may not be knowing that my customer reach is very large. or will become very large. So, today it is small, but tomorrow it can become grow and become much large. So, today probably it is encompassing few communities, but tomorrow when it grows it might be encompassing much larger number of communities, different races etcetera regions. Just take the case of example we will say the country like India.

So, I am sitting in say West Bengal and I develop something and I know my customer base is in and around this state, but when I grow my customer base gradually increases spreads to all over the country. But the thing which I had developed had some reference to certain characteristics of people of say this region. Now, we have people from all other regions. So, they might feel impacted that we do not have this characteristics we have something different characteristics. So, this product does not whatever suit us or it is probably giving us some wrong information or making us do something which we do not it is not palatable to us.

So, that is the thing which is being what it refers to as when we are saying inclusivity. So, when you develop a tool or a product which has to be used by the public or the customer, try to broaden it as much as possible in the context of communities, males, females, gender, religion, race, language. take all of these things as much of things as you can consider age group, the language they speak, the education level, the economic level, it can go on and on, but you must consider that and normally we tend to miss out all those things. So, an example of inclusivity is, suppose you have a question: how can we debias AI?

So, we can debias AI by training the AI on a diverse set of data. Now, what constitutes a diverse set of data? A diverse set of data consists of data from a diverse group of people. Now, what constitutes a diverse group of people? A diverse group of people consists of individuals from diverse backgrounds, such as

Now we are coming to the backgrounds—finally, we have arrived. So, we need people from diverse backgrounds. So, what are these backgrounds? such as people from different countries, people from different ethnicities, people from different genders, people from different sexual orientation, people from different religion, levels, knowledge levels—this list goes on.

So, now you can understand what we mean by inclusivity and how challenging it can be from case to case. But as a responsible citizen, you must keep these things in mind—that

is very important. The second principle of responsibility is sustainability. It must be developed in a manner that is sustainable and environmentally friendly. The technology should not contribute to climate change, and development should be guided by principles of ecological stewardship. So, whatever we do, whatever we produce, should not cause any major impact on the environment.

The third principle is safety, just look at this flow chart. whether my tool is the AI tool I made is safe enough or meets this principle of safety. So, we ask a question does it matter if the output of the tool is true? It could be yes, it could be no, if it is no then it is not safe to use this tool. If it is yes then we ask do you have expertise to verify that the output is accurate?

If it is no then you should not use it, if it is yes then you can ask are you able and willing to take full responsibility legal moral etcetera for missed inaccuracies that by chance there is some inaccuracy and then you are you know whatever people are not happy the user was not happy etcetera. So, you need to take the full responsibility of managing this situation if no then do not use unsafe to use and if it is yes fine possible to use chat GPT. So, this is how you do a check. whether it is safe to use this tool or not to safe.

So, these questions you need to ask and of course, answer them. So, it must be developed in a manner that is safe and secure for users, the technology should be designed to mitigate risks associated with its use and to prevent harm to users and other stakeholders and if something goes wrong then as a owner of the tool as the organization you are owning the tool which you are released for your use for your customers or whatever you should take a whole responsible of the mistake which is being done by that tool. The fourth principle of data protection whose data is AI trained on does it respect sources.

So, we have already talked about this plagiarism copyright and things like that. So, you must be very aware that whatever has been produced by the tool it must comply with data protection legislation to ensure the privacy rights of individuals are expected because you are using huge amount of data many of them could be your internal customer data. companies internal data source must be customer data. So, data privacy has to be protected.

So, we need to ensure that when you are using this data to train your tool. So, lot for European Union law Brussels effect like Brussels effect to come like GDPR is one which we use in IT industry you might have heard about it you often read in the newspaper that Google and Amazon are sued. for so many million dollars etcetera because of data

privacy violation. So, similarly in AI currently we have these rules and acts like data act, data governance act, digital services

act, digital markets act and e privacy regulation, but much many more such regulations and acts are required to protect customers. The fifth principle is the intellectual property fraud again very similar to the other one. Whether you are misusing intellectual property, they often generates output that are highly original and creative which raises questions around copyright ownership. So, this example I have shown here is Ghibli art by Coppola. So, you can use this tool from Coppola to create such art forms, but they can have copyright violation because you might

be unintentionally copying some other characters for example, may be some cartoon characters or some published characters. So, you must be careful to ensure that the outputs generated by AI technology do not infringe on the rights of third parties. Example of deepfakes we can create various art forms deepfakes just copy the existing art and then try to say that this is the real stuff. The sixth principle of sectoral regulation for is the sensitive data. medical education and crime.

So, those are legal frameworks that govern the use of AI in specific context such as health care and criminal justice where you have to protect the privacy of your clients in for legal justice and protect the privacy of your patients in medical cases, because you are handling very personal information of individuals as a doctor or as a hospital. So, the Gen-I technology must be developed in a manner that complies with relevant laws and regulations. So, there are existing laws already for protecting their privacy.

So, your Gen-I tool also must comply with those laws. The education sector, Gen-I has proved a very big challenge for us teachers like us. So, high tech gadgets by increasing the students you know what they cannot produce anything any answer for any questions they can get very easily in their equipments like phones or tabs or laptops etcetera. So, because extremely difficult to give say the home assignments or projects for that

because you never know it is very difficult to find out whether that report which has been produced or been submitted whether originally written by the student or created by a gen AI tool. So, this is one of the real challenges the education sector is facing and it will be there it will increase the challenge and we will have to work around this challenge because deny that access to the students I mean you cannot do that it is not physically possible to stop them from accessing these tools. So, you have to live with it, but you have to find out a solution work with the tool and see to it that the tool is used and not

misused. I do not know I am not very clear also what it means also being a teacher I am not very sure what exactly it means.

or how best to use the tools. So, many thinkers, many academicians from Harvard and MIT and leading institutes are all many of them are struggling with this and they are thinking and trying to find out solutions to the problem. Generative AI has potential, but it must be developed in a manner that is responsible Developers must consider the principles of inclusivity, sustainability and safety and then ensure that this technology complies with relevant legal frameworks. But the thing which you will have to think that when you are from a business angle when you are trying to develop a tool this is too big a task really.

So, you have to be very conscious. Number one, number two you have to have a team in place because it cannot be done by a single person like a CIO. So, you have to have an ethical committee what we call as a governance. So, I think I mentioned earlier also that when you start a AI project. So, you set up a team a multidisciplinary cross functional team from various from software analyst to design engineers to AI programmers coders.

etcetera and at the same time you must also set up your governance. This is something which we tend to ignore or tend to delay fine let the tool develop, let the work, let the trial start, let the experience start, then we will see etcetera before we deploy etcetera. But that will be too late because they also have to do extensive work and they also have to oversee whatever design parameters are going into the development of the tool. what are the various data sources are being accessed to train the tool. So, you have to start from there.

So, when your team starts your governance team should be up and running and they should be monitoring on a regular basis regarding the development of the work. The agenda is very simple, you have to keep a track of what is the data set be used. But the question is how are people qualified to check the quality of the dataset. For example, if you are handling a 1000 page documents or many several files, many past records etcetera and they are getting digitalized and being used as for training media. go through all those tools and find out the biases or the types of biases.

Like I was I told you about that previous slide that what here if I go back to this slide this is a I find a very interesting slide for me. So, here what is the diverse set of people and then such as it goes on goes on goes on goes on goes on. different countries, religion, language, age group, education level, your economic level, your intelligence level,

whether physical disability level, physical and mental. So, all of them are forming a particular cohort. So, we talk about gen y, gen z, gen alpha, gen whatever, beta etcetera.

So, each of them have different ways of thinking and different ways of reacting. So, I have to consider that. what z y will react, what z z will react, what z alpha will react or z beta will react etcetera. So, do catering to all of this requirement or to ensure that there none of the requirements are impacted, none of the feelings are impacted is not an easy task like this. You have to ask yourself will the output be accurate?

Now, how do I answer that? So, I have to be a real SME trying it out and then try out various prompts or whatever inputs and check the outputs to get an feeling and idea. And so, I have to go on testing large number of outputs to find yes this is accurate and it is not hallucinating because hallucination is one of the biggest fear. from this JNI tools and if it starts hallucinating then it is not ready enough. So, coming back to the last portion again I will repeat the education sector segment which is seen the biggest

challenge and it is involving it is going downwards towards from education in post graduate then it is I think it will go down to the school level very soon and the school children also because now your open AI chat GPT's are available on your phone handset. So, you can do whatever you want find out whatever you want write whatever you want write report thesis etcetera. So, how will that impact learning is a big question and we do not have any answer to that. So, we might have to change our definition of ethics in education is copying

very unethical or we can tolerate copying to some extent and then augment maybe that with some other tool or techniques. So, that students can ultimately a motto is for the students to learn or for us whether you are being able to make students learn. So, we need to copy something from a whatever tool and they learn then probably you should be happy with that. I will continue on this in the next lecture. Thank you very much.