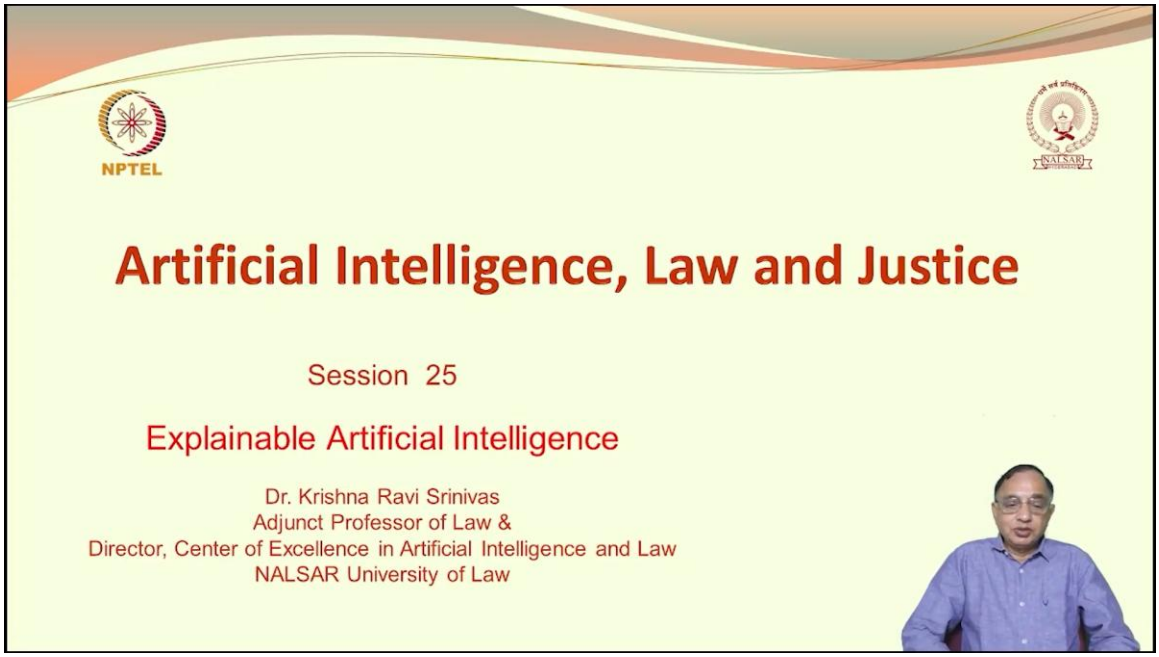
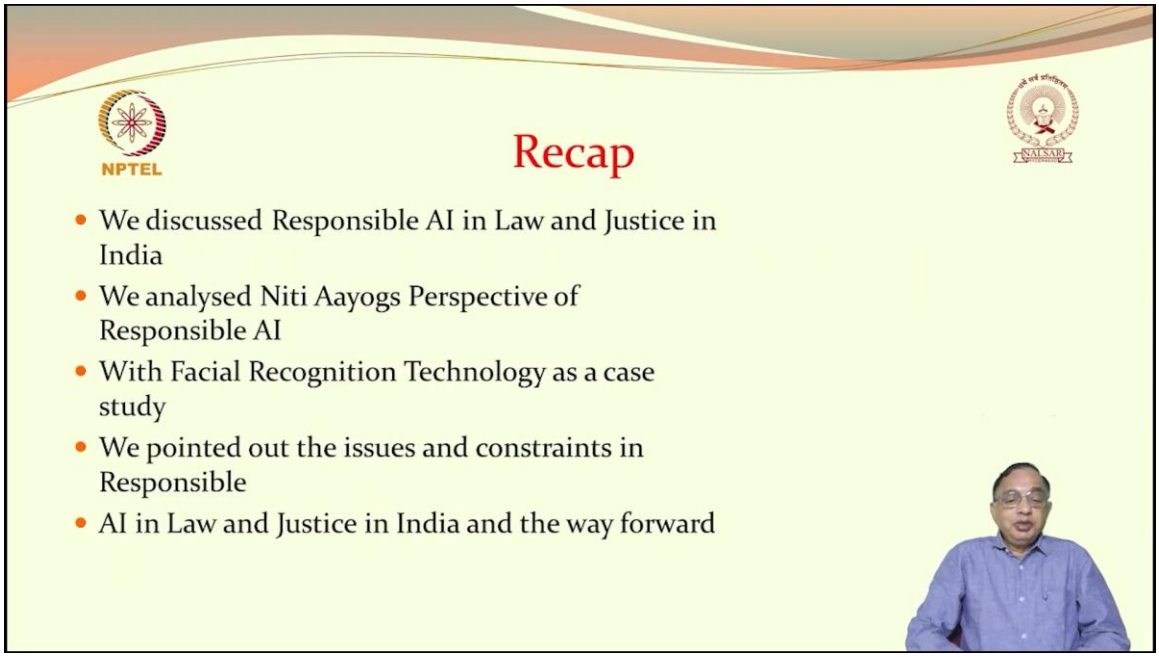
Course Name – Artificial Intelligence, Law and Justice
Professor Name – Dr. Krishna Ravi Srinivas
Department Name – Center of Excellence in Artificial Intelligence and Law
Institute Name – NALSAR University of Law
Week – 05
Lecture – 25

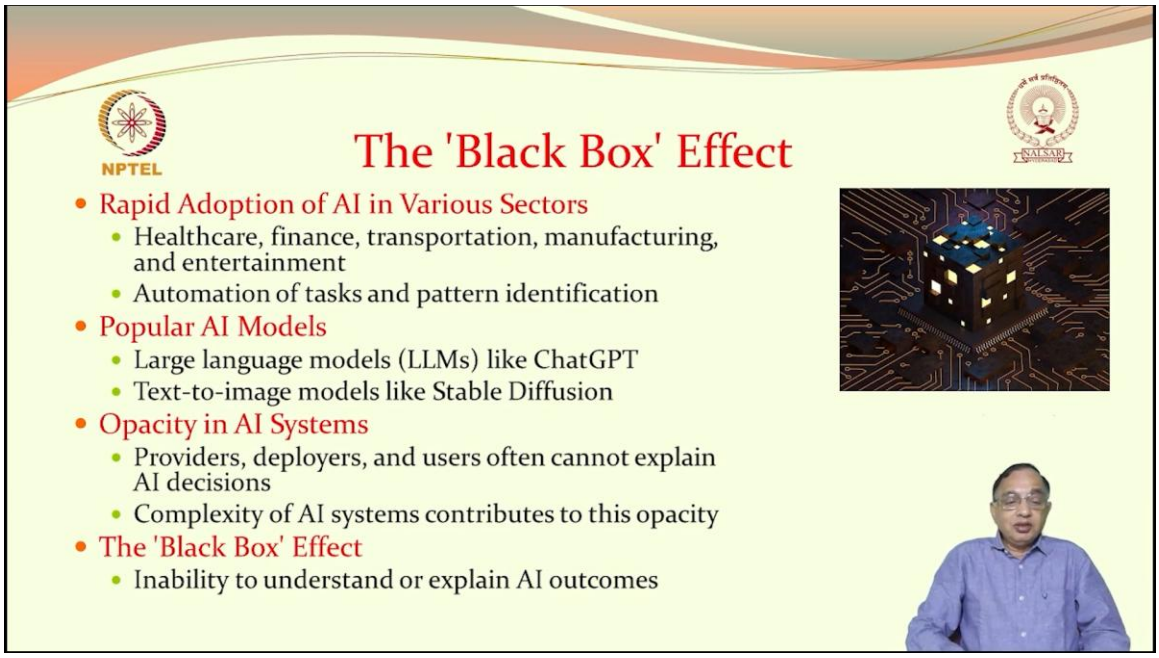


Artificial Intelligence, Law, and Justice Course, Session 25: the topic is "Explainable AI."



Let us do a recap of what we discussed in the last class. We discussed Responsible AI in Law and Justice in India. We analysed Niti Aayog's perspective on Responsible AI,

particularly taking into account facial recognition technology as a case study. Then we also pointed out the issues and constraints of Responsible AI in Law and Justice in India and the way forward.

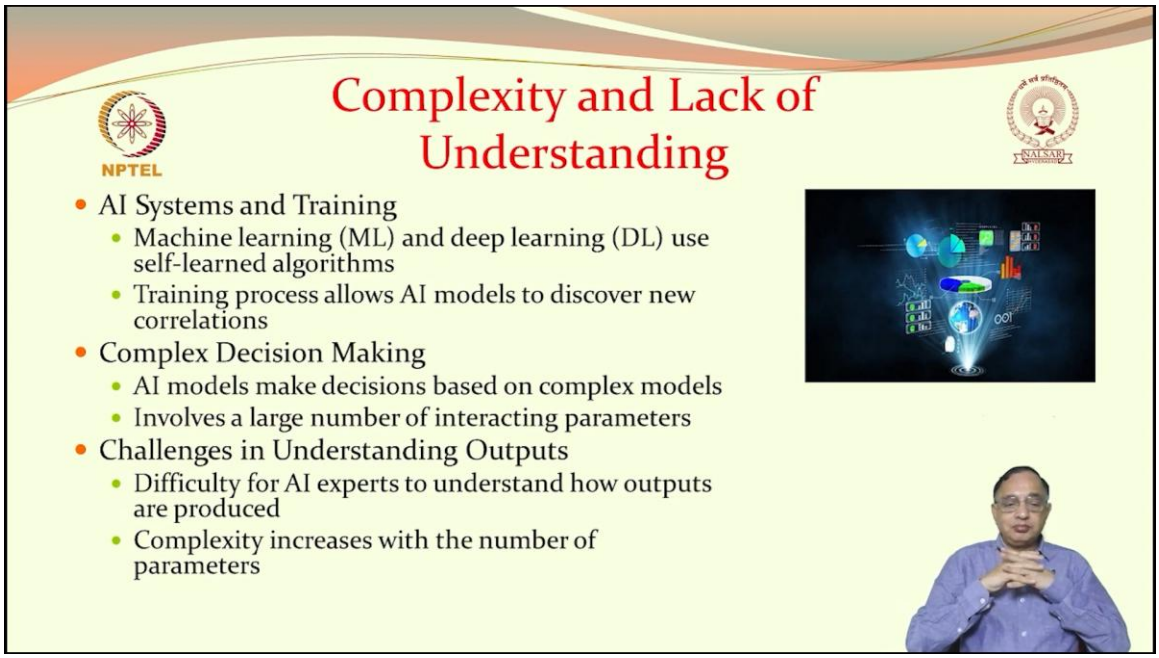


So why are we talking about explainable AI in the context of AI is that there has been rapid adoption of AI in various sectors, including health care, finance, transport, etc. And then a lot of tasks could be automated, and patents could be recognized. So, AI's adoption is much more sophisticated and is growing day by day. And then there are a lot of new models coming up in the sense that ChatGPT itself has many variations available, and not even a single day goes by without some minor variation of some model or some new model coming up in one way or another. Additionally, there are a lot of other applications, including Stable Diffusion, which deals with text-to-image models. So, the popular AI models, on one hand, are expanding, diversifying, and their sophistication is increasing. And then there are a lot of other models that are not popular in the sense that they are not available for everyone to use.

But then very sophisticated models that are inbuilt and tailor-made for certain applications in specific industries are also expanding. What we are seeing is that as the AI systems become more sophisticated and more easily available on one hand when it comes to models like ChatGPT, the other side of the picture is that nobody actually knows what exactly these models will do or what exactly these models are going to explain. Put in another way, we are seeing lots of developments in AI being done by different agencies, different actors, different model developers, and so on. And then there is a lot of convergence in them based on the applications and the users. But still, the systems remain more like a black box, in the sense that nobody today has come up with something that makes a system 100% explainable or that a system is built with explainability in mind. So, this opacity makes things difficult for others to understand, but it also allows

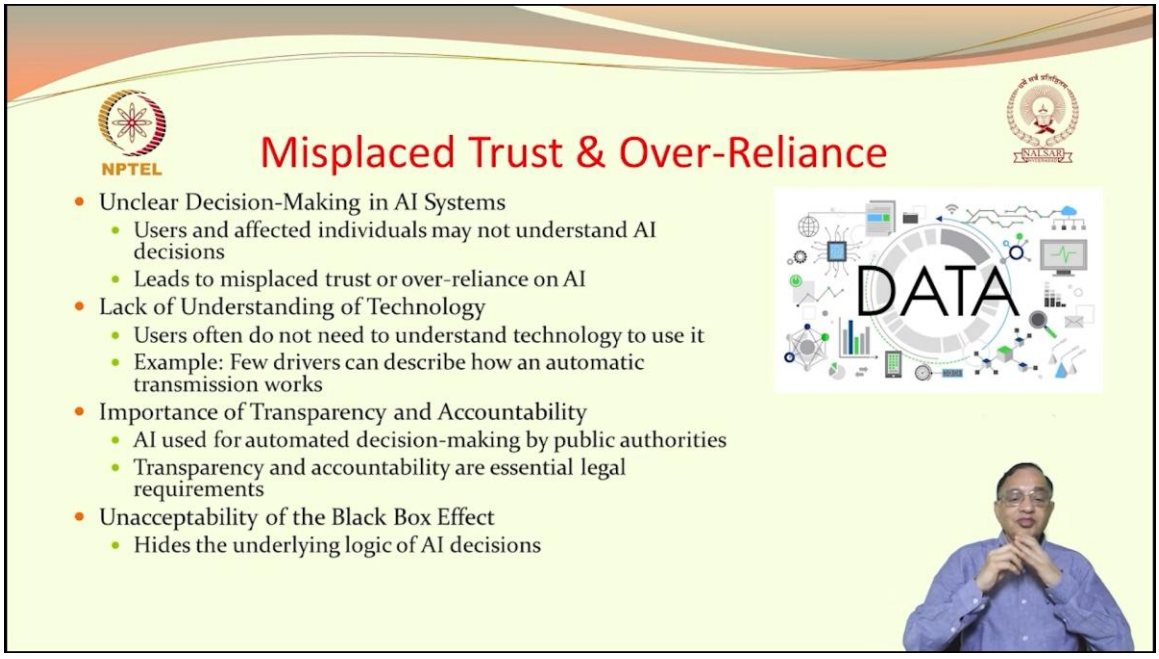
developers to easily explain what the system does, why it does it, and how the outputs are derived. And then the bigger problem is that as systems become more complex, this opacity also increases. So, this black box effect is, in fact, resulting in a situation where we have too many models and too many AI systems with too little explanation.

So, this black box effect, when it cascades or when it increases, will result in a situation where we will know very little about all the models we are using; we will know very little about the models that make decisions for us, and we will know very little about the models that are widespread across different sectors, utilities, and purposes. This is not a very comfortable situation, particularly as AI becomes more and more adopted. We should at least have some clarity as to what these models are up to and whether they can be explained to us in the sense that this is exactly what they're doing.



Machine learning and deep learning use self-learning algorithms. And then the training process allows them to discover new correlations. This we know. But it is also true that models make decisions based on complex algorithms which they themselves often derive or develop. And then this involves a whole lot of interacting parameters. So, when this happens, it even becomes difficult for AI experts to understand how outputs are produced or what exactly the decision trees are that result in the final outcome. In the sense that when algorithms make a decision or the decision is made on account of models that have arisen out of self-learning algorithms, it becomes complex for even them to understand and then explain why this happened, why this did not happen, or why this output occurred but not anything else. And when complexity increases with number of parameters, it becomes all the more difficult for anyone to get convinced that these systems are really safe and reliable. So, complexity and lack of understanding, on one hand, tell us that systems are becoming more and more sophisticated and higher in

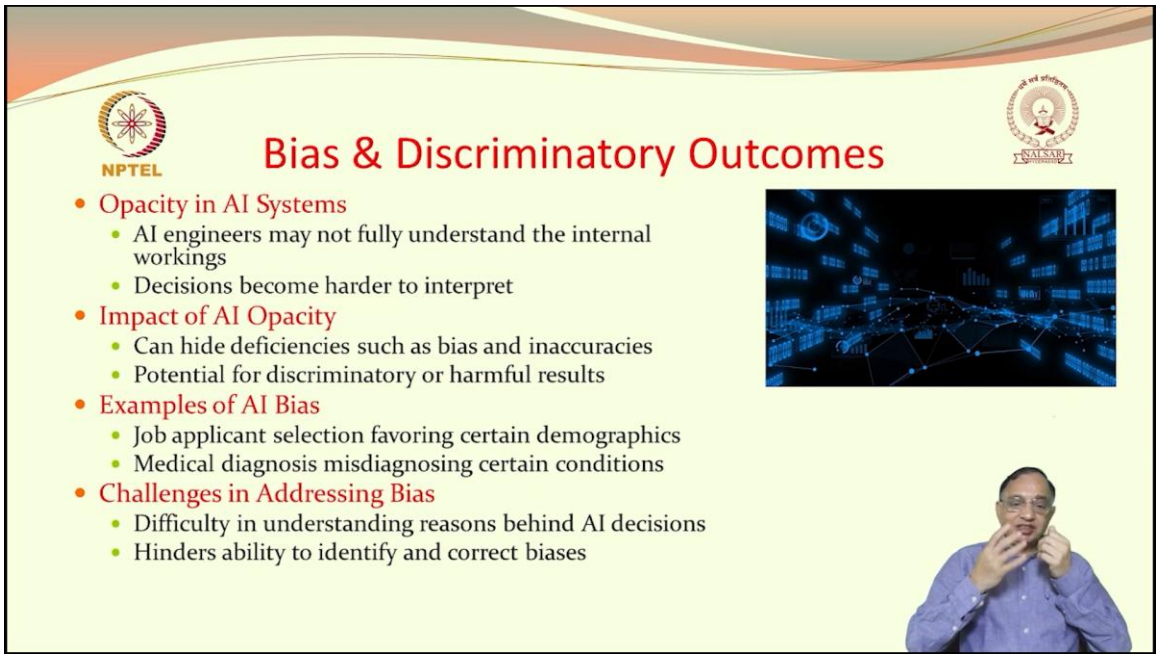
capacity, but they are also worrisome because we really won't know; we won't even be able to understand what they are up to.



But should we go for misplaced trust and over-reliance? This is something we need to really look into because what happens is that an unclear decision-making AI system affects everyone, as affected users may not even understand why they were impacted on account of what decision, so this results in either over-reliance or misplaced trust. Although people may not be able to understand why it has been done this way, they may have a general faith or a general trust that AI systems are reliable and trustworthy. Because they would have been told that systems built for specifications are accurate, that there is no human intervention, and that everything is driven entirely by algorithms, statistical programming, and other factors.

Often, the other problem is that it is very difficult for users to understand the technology. For example, only a few drivers can tell exactly how an automatic transmission works in a car, in the sense that they know that an automatic transmission car exists. However, the way it works is something they would not be able to describe, even if they can comprehend it to some extent. Then what happens is that we start relying too much on algorithmic decision-making, AI-based decision-making in systems provided by public authorities, whether it is in railways, banking, insurance, or the normal day-to-day delivery of services like education and health. Transparency and accountability are far more important so that people know whether these things meet the essential legal requirements in the sense that they have been tested to meet the essential requirements in the sense that they do not discriminate, they do not have any inbuilt biases, and more importantly, their decision-making is in tune with or aligned with the human decision-making process.

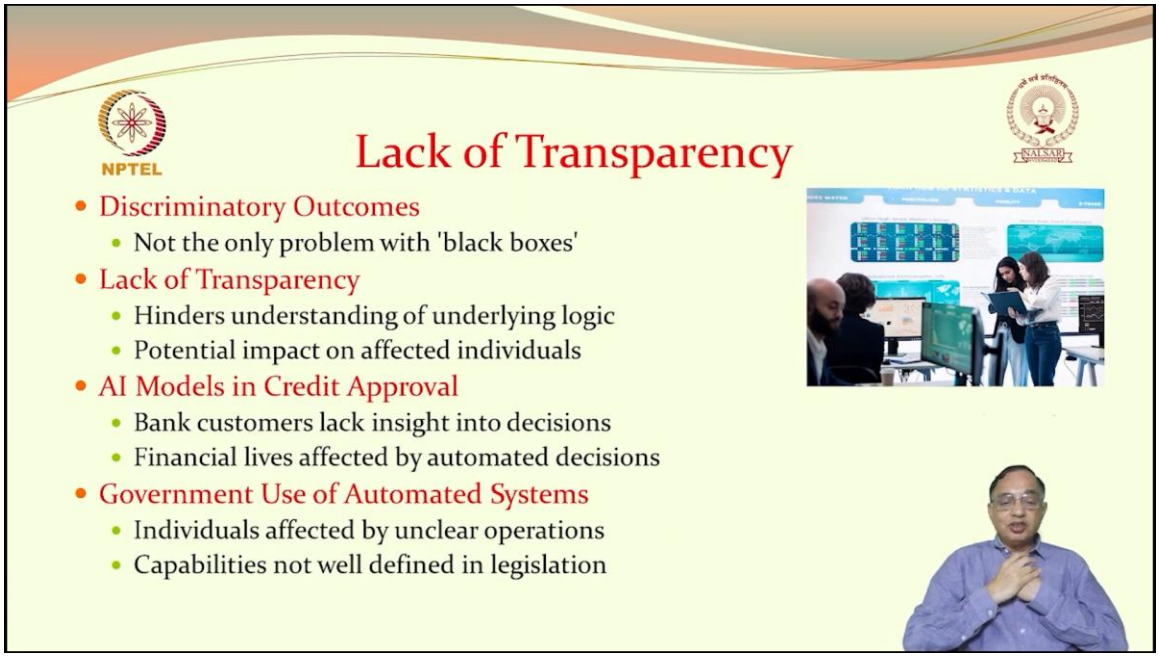
So, the unacceptability of the black box effect actually hides the underlying logic of AI decision-making. So, we should not end up in a situation where either we become overly reliant without understanding or we trust them too much without bothering to know or without even being told that there could be issues.



So, this opacity, as we said, makes the decision harder to interpret. And the thing is that this opacity, in fact, can act both ways. Opacity can also be used to hide certain things from the public and suggest things that you won't be able to understand. But it is also possible that the opacity itself could be a good cover to hide biases and inaccuracies, either in the model, in the data, or in the way the system crunches the data and arrives at a decision. So whether the system has negatives or positives in the sense that it could result in discriminatory or harmful results, or whether it is able to process things in such a way that it does not discriminate and has no biases, is very difficult to understand unless we know that the system is somewhat transparent and that opacity is not used as an excuse to deny information about the way systems work. In some applications, like applying for jobs, if the job applicant selection favours certain demographics or people of certain categories, then the bias would be there; but people should know that it is there so that those who are impacted can fight back or at least know where they should go and ask for justice.

Similarly, as we have seen, even earlier medical diagnoses could involve diagnosing certain conditions wrongly, as in misdiagnosing, and this could have serious repercussions for health. But the fundamental issue here is that, as it is very difficult to understand what exactly the reasons behind AI decisions are, it becomes almost impossible to challenge and address the biases that we have seen in a couple of cases earlier as well. So, when this happens, it becomes a constraint to identify the potential biases and then correct them. So, this bias and discriminatory outcome is not good for the

persons who are going to be the consumers, as well as for those who developed it, as well as for those who deployed it.

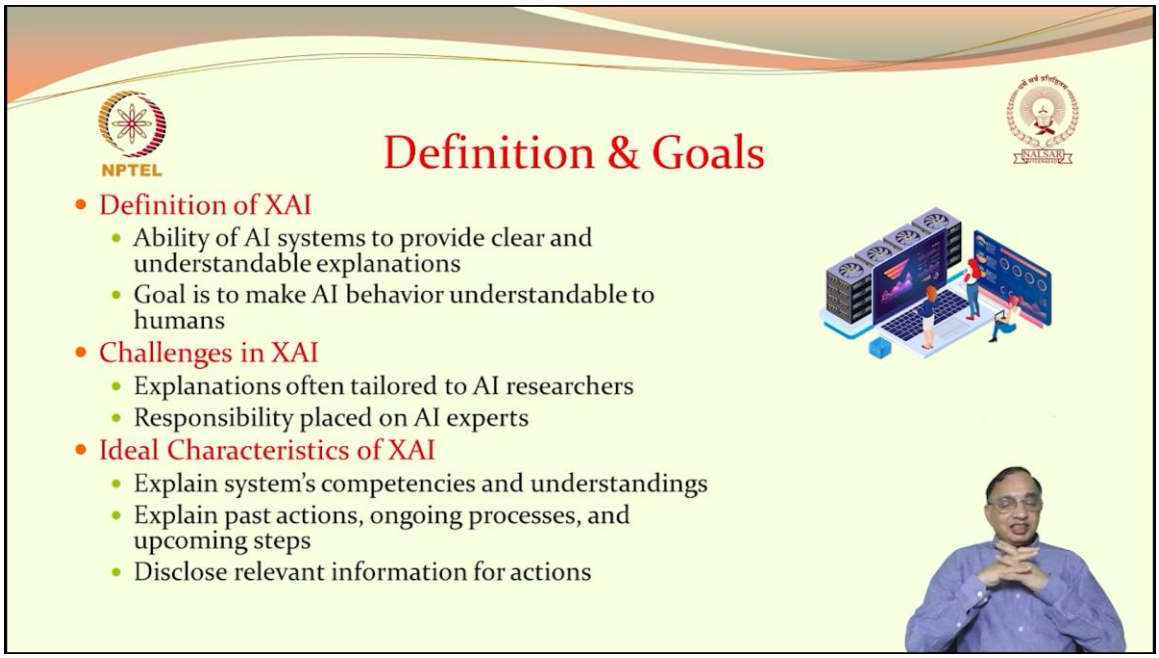


But the discriminative outcome is not the only problem with black boxes. The lack of transparency hinders understanding the underlying logic regarding the basis on which this decision was made and what the conditions and parameters that were used or chosen to arrive at this decision are. The potential impact on affected individuals could be small, large, or something in between life and death, such as denial of bail, denial of services, or denial of a loan.

Particularly when bank customers lack insights into decisions, because the customer would have known or would have thought their credit score is high, they have cleared all the dues, they have nothing to pay, and their financial condition is also in reasonably good shape, so they should be getting the loan. The outcome is that the person is not getting the loan or is getting the loan with conditions that the person did not anticipate or that are harsher; then the problem would be what exactly the AI models decided. So, automated decisions can affect people's financial planning; they can affect people's purchasing power; they can affect people's access to many services, including education. That becomes all the more acute when governments use automated decisions, and then individuals will have the expectation that they will be treated fairly and that the system will give them the results as any other human would have done. But then when that doesn't happen, it becomes very difficult to even understand why the system is behaving like that.

But the real problem is that when these automated systems are being used, the capabilities may not be well defined in the legislation in the sense that systems might have been put to use without clear-cut capabilities, meaning that these systems will be used only for this purpose and then will not be used for this purpose, or these automated

systems will be used only for the exclusive idea of allocation of rations, allocation of educational loans, or for deciding on giving housing loans. When that is not clearly defined in legislation, the potential problem is that ill-defined legislation can also result in some other system, which is developed for some other purpose, being used without a clear-cut understanding as to why this should be used here and why this system should be refined to meet the requirements of the legislation. So, the lack of transparency can occur at different levels, and the problem could also occur at different institutions.

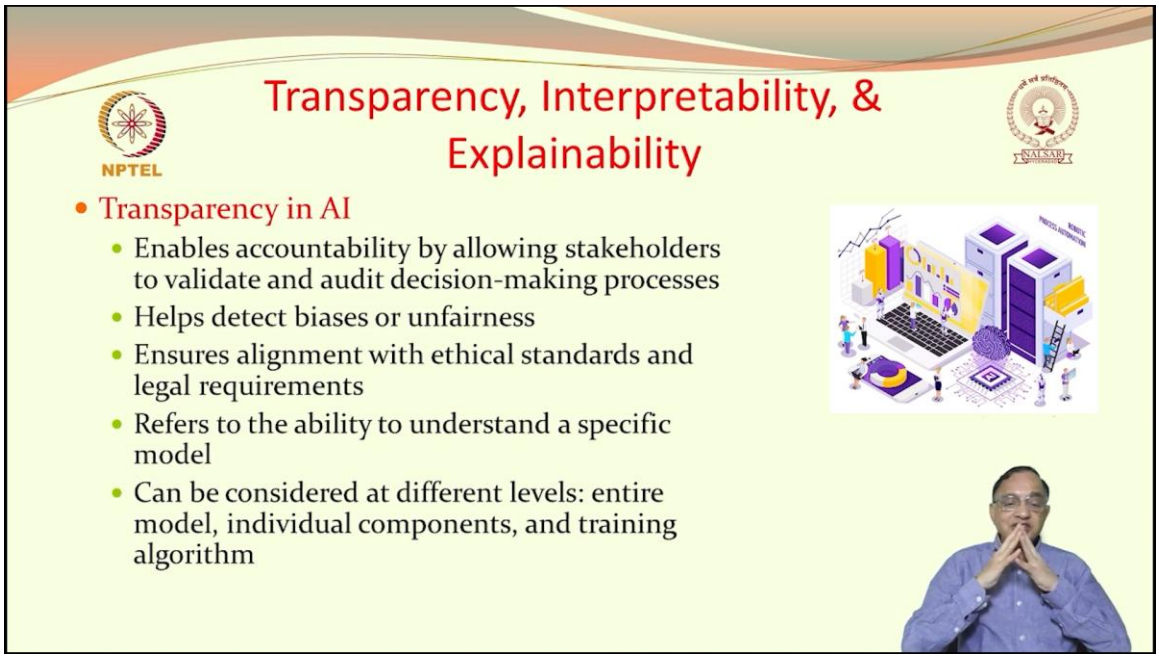


So we need Explainable AI, which is also shortened to XAI, as something that can help us overcome some of these issues, and more importantly, Explainable AI (XAI) will also help both the developers and the deployers, as well as the public or the users of the system, in that the system's functioning, its output, and its working rationale can be explained to anyone in an understandable way. So, to define it, it is something very simple: the ability of the AI system to provide clear and understandable explanations.

When we talk of explanation, we talk of explanation as it is understood by humans in the sense that you tell categorically and unambiguously why this person did not get it or why this person got it. The explanation should either justify or provide some acceptable, justifiable reasons for the decision. It is not that the explanation is a very crisp one or two sentences; 'you didn't meet the criteria, so it was not given to you'. So, an explanation is something that we would see in the court's order or the human judgment, where it is thoroughly explained, taking into account the circumstances, facts, and everything, and then the decision arrived at is explained. So, explanations can also be of different levels and different categories.

Sometimes people may want a very detailed explanation as to at what stage they failed in the sense that it was felt that they could not meet the conditions. At what stage of the process was their application not considered favourably and rejected? They would also

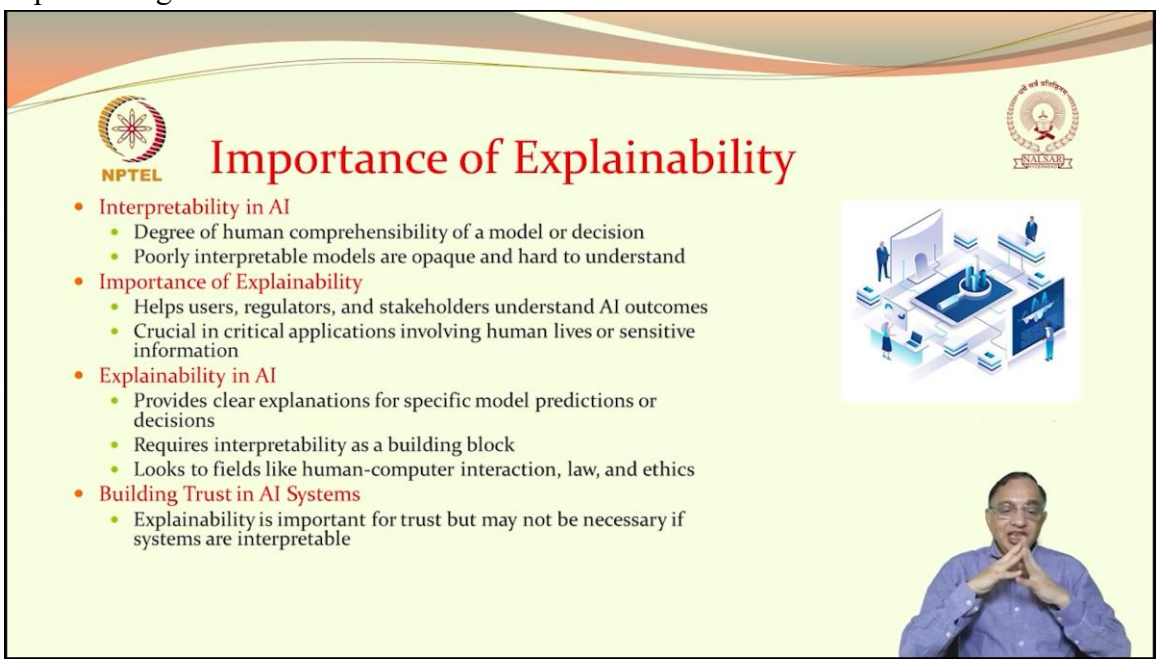
like to know at what stage of the processing the decisive turn to say yes or no was taken and on what basis. What is the precedent, or what is the decision that was made? Was it made on the basis of any previous decision made regarding the same person but for a different purpose? For example, someone might have applied for a housing loan earlier. The earlier decision to deny the person a housing loan -was that really taken into account when the person applies for an educational loan? And then, since that denial gives a negative impression, was it also factored in when the decision was made? People would like to know. That is very fundamental. So, the issue here is that the responsibility is on the shoulders of the AI experts and developers to make people understand how these decisions were arrived at by the system. In an ideal situation, we should be able to understand how the systems work, demonstrating competency through Explainable AI (XAI). Then, you should also look at what the past actions were, what the current situation is, and what the next steps are. So, disclosing the relevant information, correct information for all actions is important.



But there is a catch. When we talk of transparency and accountability by allowing stakeholders to validate and audit the decision-making process, we presume that the stakeholders have an idea about it and that they are not blindly having faith in it. So, we need to first educate the stakeholders as to how the systems arrive at decisions in plain language or in language understandable to them, that this is the way the system works and this is how the decisions are made. So digital literacy should be part of educating the users and the stakeholders, where they should also be trained in some of the fundamental ideas of Explainable AI (XAI). Otherwise, what would happen is that the stakeholders having accountability, in the sense of having standing to know that the system does something and it is being explained without getting a fuller understanding, will not result

in fuller transparency. But when stakeholders themselves are aware of it, it will also help them understand the defective biases and unfairness.

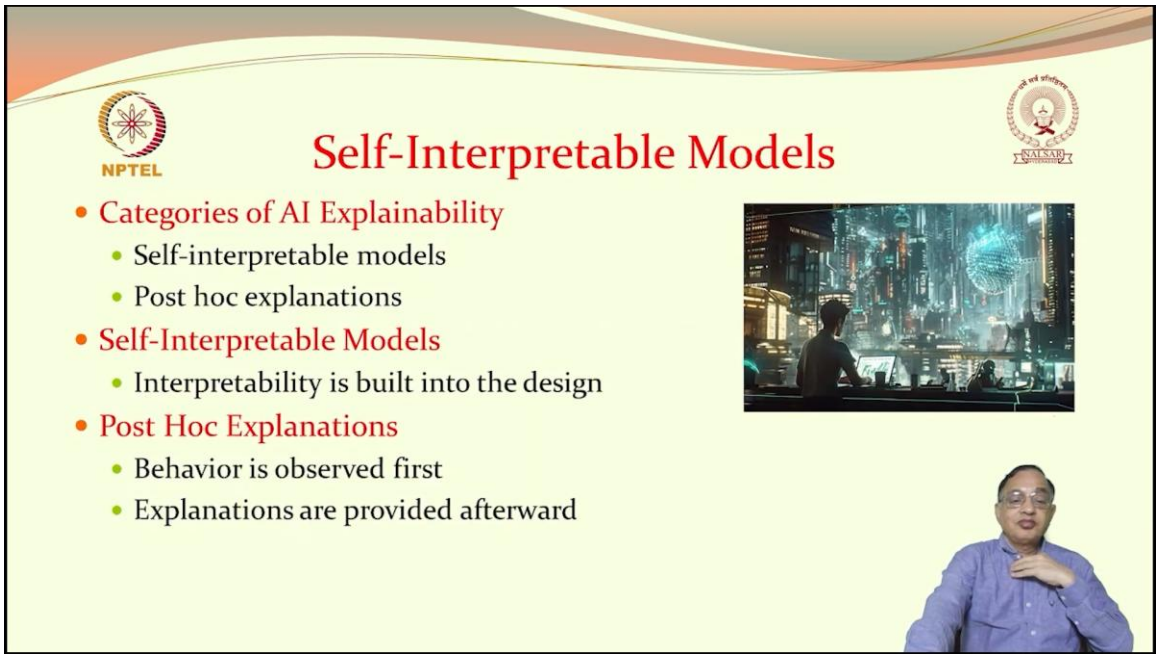
And then the system can be aligned with ethical requirements and the standards, or it can also be fine-tuned in such a way that it fulfils the key criteria for Responsible AI (RAI). Transparency should not be limited to a larger system; it should also be extended to a specific model, which again could be a standalone one or could be part of a larger one. At different levels, we need to look at the explainability of the entire model: how it works, what the flows of the data and information are, what the decision-making steps are, who decides which step, what is decided, and how the final decision is arrived at. Then, we need to examine the individual components, whether they are algorithms or data, and how they are linked together in the way the processing is done. Finally, the training algorithm is equally important because it ultimately plays a key role in the way machines understand things and then make decisions. So, transparency, interpretability, and explainability should be seen in a broader way than looking at them as mere technical issues or issues that could be solved by bringing in some technical experts who can explain things in technical terms.



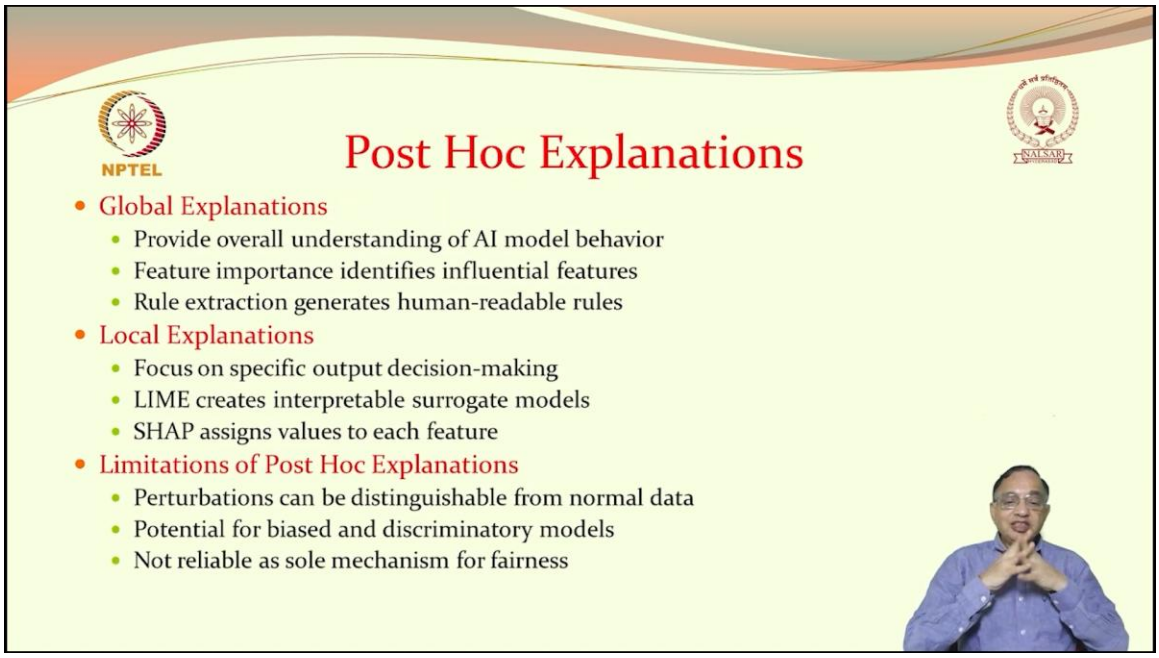
The importance of explainability is that the degree of human comprehensibility should be taken into account. Unfortunately, the poorly interpretable models are very opaque, and they are very hard to understand even for experts. Then, when explainability is built in, it helps the regulators, helps the stakeholders, and helps everyone to understand the outcomes and appreciate them as well. But when it comes to crucial applications, this is all the more important. So, one way to look at it is that you shouldn't interpret interpretability as something that is different. Consider that as part of the explainable AI framework. So, when that is done, the clear explanation for specific model predictions or decisions should be given, and then this interpretability can be something that is

technical, but to understand it in its fullness, it is better to look at the way human-computer interaction happens and then what are the things that they can learn from law and ethics. For example, human-computer interaction can happen in various ways; sometimes the user input could be wrong, but then the user may not even be aware of it, and the outcome could be totally different. So, we need to also look into the fundamental issue of whether the humans who are in the system or who are going to use the system really understand how they are going to interact, what they input, what is being sought from them, and then what the expectation of them is.

So explainability is important if we want the systems to perform reliably and for people to have faith in the system. But the problem is that it may be important for trust, but it may not be necessary for systems that are interpretable. Not all systems need to have explainability as a factor if they can fulfil the criteria for interpretability.



Because some models can be self-interpretable in the sense that they can tell us what exactly they are doing; they can explain, "I am doing this, I am taking this step, and then I am taking data from that. After doing the data crunching, I am going to the algorithm, and then based upon the algorithmic decision-making, I am feeding the data; then the next step happens." So, in the self-interpretation models, interpretability is part and parcel of the design. But in some other models, the decision could be taken, observed, and then explained. So, first what happens is that the system arrives at a decision. It is not self-interpretable, and then it observes its own behaviour and explains why it took the decision or why this decision was handed over to you. So, if we can build interpretable models, it will be really good, but then technically it may not always be feasible to build them in all applications.

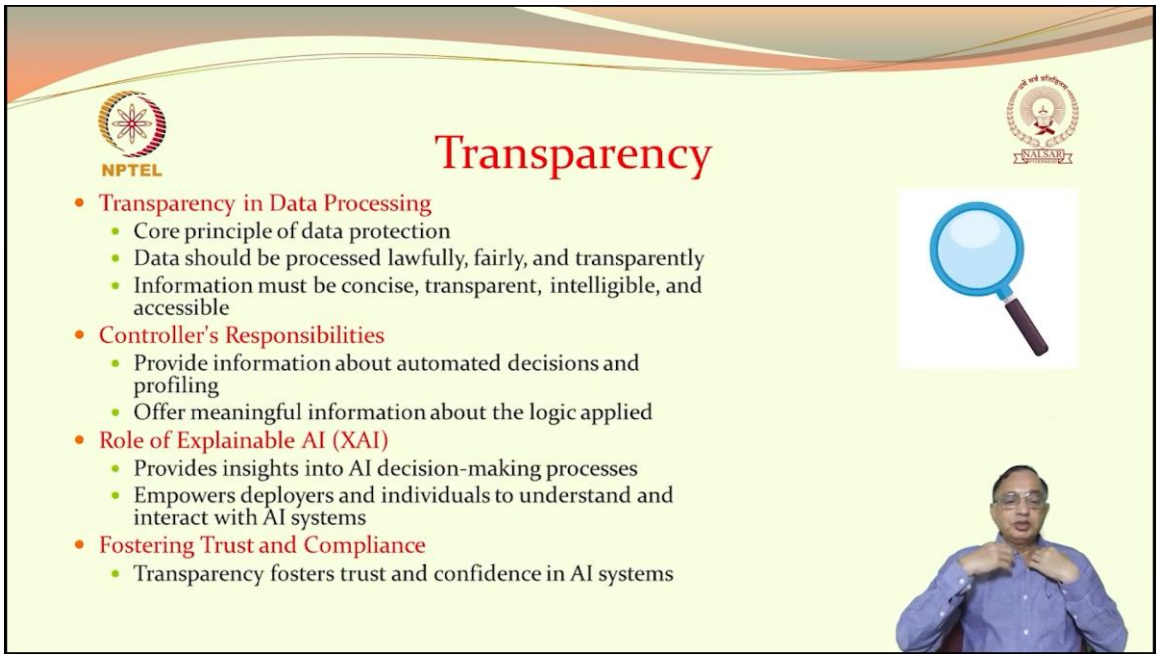


The post-hoc explanations, which come as an afterthought, or we would say that after the system has made a decision or handed down a decision, are also important because they help us understand how the overall system is functioning; and then, by looking at them, we can understand what the major key factors are that influenced them. For example, when it comes to creditworthiness, whether the key factor is the repayment capacity, the experience with previous loans, or something else, it will be worth knowing for the person who applies for a loan. So, rules can be extracted from these explanations, which means that you would know rule number 1, rule number 2, and rule number 3. If rule number 3 is not fulfilled, go back to rule number 2. So, these things can be understood by humans, and then this could be made part of the other frameworks where you know this could be built upon and then made better.

So, when we talk of global explanation, we talk of the whole system where everything goes top to bottom: 1, 2, 3, 4, 5. But within the bigger system, there could be something very specific. For example, the bigger system can be a health-based one where the diagnosis could be for a specific disease or for a specific organ, whereas the bigger system could be for overall health diagnostics, taking into account all the parameters and all the findings. So, the specific output-decision making is also possible. LIME is a technical one that can create interpretable surrogate models which can tell us how the system works and then SHAP: assigns values to each feature.

So, these are technical terms that we need not go into fully in detail, but we should understand that local explanations are also possible, desirable, and feasible. The problem here is that post hoc explanations may be more likely to justify in the sense that the perturbations cannot be distinguished from the normal data, or whatever changes or actions the system had undertaken would be the very same as the normal data. And then, if the systems' models are biased and discriminatory, post hoc explanations may not be

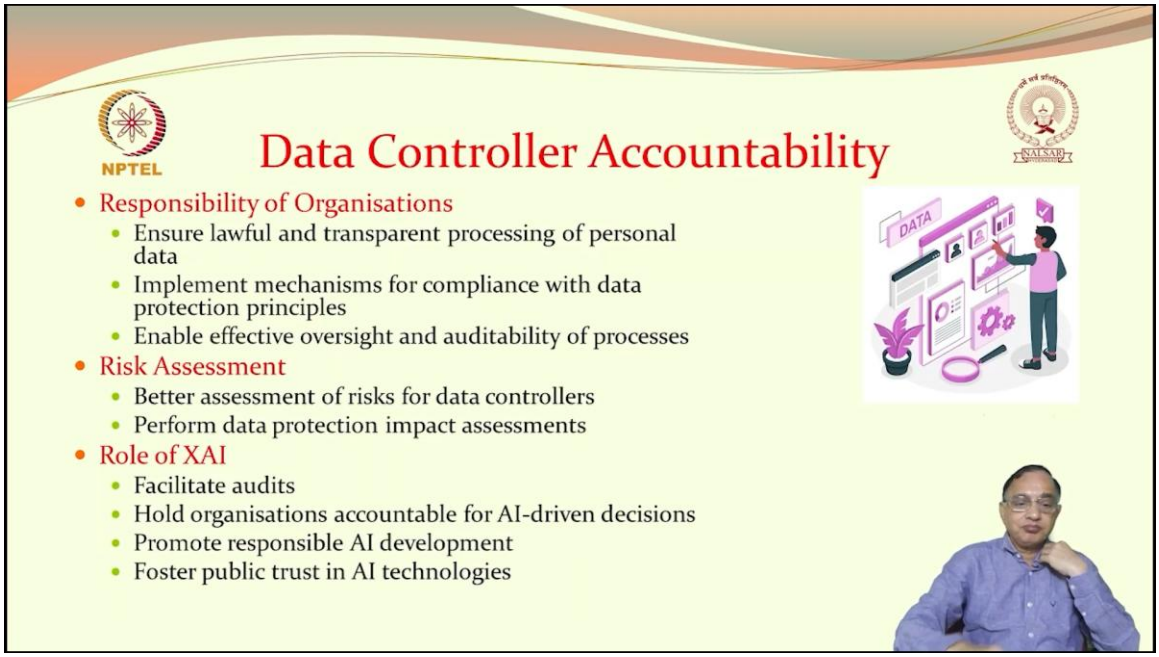
able to capture it fully or may just cover it up, or they may come up with an explanation that seems convincing, but at the core of it, it is not true. Then we need to look into this post hoc explanation as one of the ways of understanding and one of the ways of using Explainable AI (XAI) in the way we want to examine systems. But it may not be the only one that we should rely upon. So, we should try to look at explanations not only from post hoc analyses but also from other perspectives; in other words, Explainable AI (XAI) should have the space when we talk about Explainable AI (XAI) systems. It should have the space or the criteria to provide more explanations than just post hoc analyses because post hoc is more like doing a post mortem, which is important but could also be misleading.



So, we need to improve transparency. Transparency could be an issue in data processing. The data process should be lawful, fair, and transparent. This again is a core aspect which the Data Protection Act try to implement or make it possible. So, the information that goes into it must be transparent, intelligible, and accessible. So, this, again, as we have seen earlier, is the problem of data interpretability and the availability of good quality data. Then the system could be controlled either by a human or by an inbuilt controller, where the controller will be monitoring the system's various stages and its overall functioning. So, the controller's responsibility is to give information about automated decision-making and profiling.

It should also tell us, in a way we understand, about the logic applied by the system. So Explainable AI (XAI) actually helps the user and the stakeholders to understand and to gain insights into the way AI decision-making works in the real world. So, it empowers them and then it also facilitates their closer interaction with the system, than they thinking that it is a plain black box which they will never be able to make sense of. So, this level of transparency through Explainable AI (XAI) will foster trust and confidence

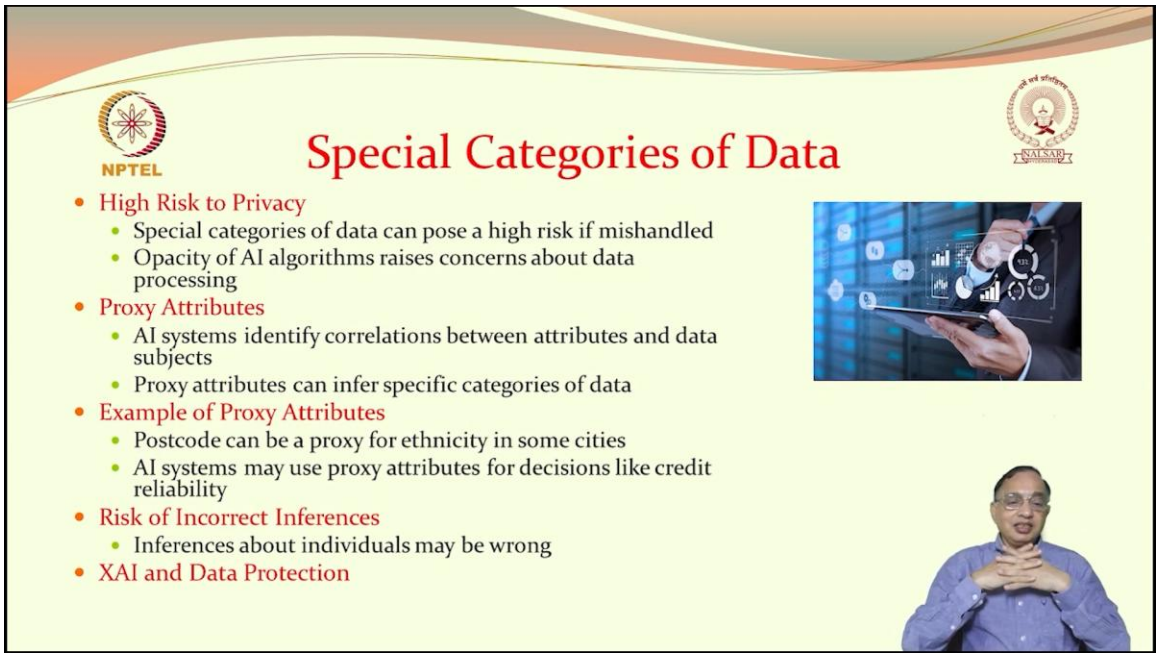
in AI systems; when systems are explainable and fully transparent, people can understand the creditworthiness and reliability of the system from the stakeholders' perspective.



Then, of course, we need to talk about the data controller. This often comes as part of the data processing ethics or data governance. So, there should be a correct mechanism for employing data protection principles, which again differ from country to country and from application to application. The oversight of the data protection mechanism and the data flow should be easily understandable and accountable. Because various acts, including the DPDP Act in India, identify someone as a data controller or someone who is in charge of the entire data processing mechanism, that person's duty and responsibility is to adhere to the way the Act defines certain things, particularly in processing sensitive data. So, the data controller will also need to look at the risk in the sense of whether the data is reliable, whether it is synthetic data, substitute data, or proxy data, what the level of accuracy is, and what the level of usability is.

So, the role of Explainable AI (XAI) is very important here because it will facilitate audits, it will also hold organizations accountable for the decisions, so it will promote responsible AI development, and more importantly, it will make the public understand and then foster public trust. Data controller accountability should be part of Explainable AI (XAI), and when we say Explainable AI (XAI), we are not talking about a very abstract concept; we are discussing a concept that is also a practice. In other words, when we say Explainable AI (XAI), the organization or institution should have a series of steps with proper responsibilities assigned to each individual or division that can take care of the Explainable AI (XAI) steps, including how the system works and who will do what when it comes to Explainable AI (XAI). So Explainable AI (XAI) should be thought of more as a process and something that results in an outcome than a vague concept.

So, data minimization is also important because one of the fundamental principles is that an AI system should not use data that is not relevant for its functioning. In the sense that data minimization says to use only the relevant data, and that is the minimum data. For example, the data for which the credit score has to be assessed and then a loan has to be given may include data that may not be directly relevant to that. It could have the banking details of the spouse. It could have details about the children or grandchildren. It could have some other details that may not be totally relevant there. The data minimization principle says to use the minimum data that is relevant and adhere to that so that you do not use data that is not relevant, and the irrelevant data does not influence decision-making. So, data protection principles have to be adhered to. And then, when we implement Explainable AI (XAI), we can also look at whether the data minimization principle has been fully put into use. So, complying with data protection regulations will also result in meeting the norms for Explainable AI (XAI) in one way or another. So, we can integrate the Explainable AI (XAI) process as part of the data protection mechanism itself and then draw certain insights from the data protection mechanism that would feed into the Explainable AI (XAI) process.



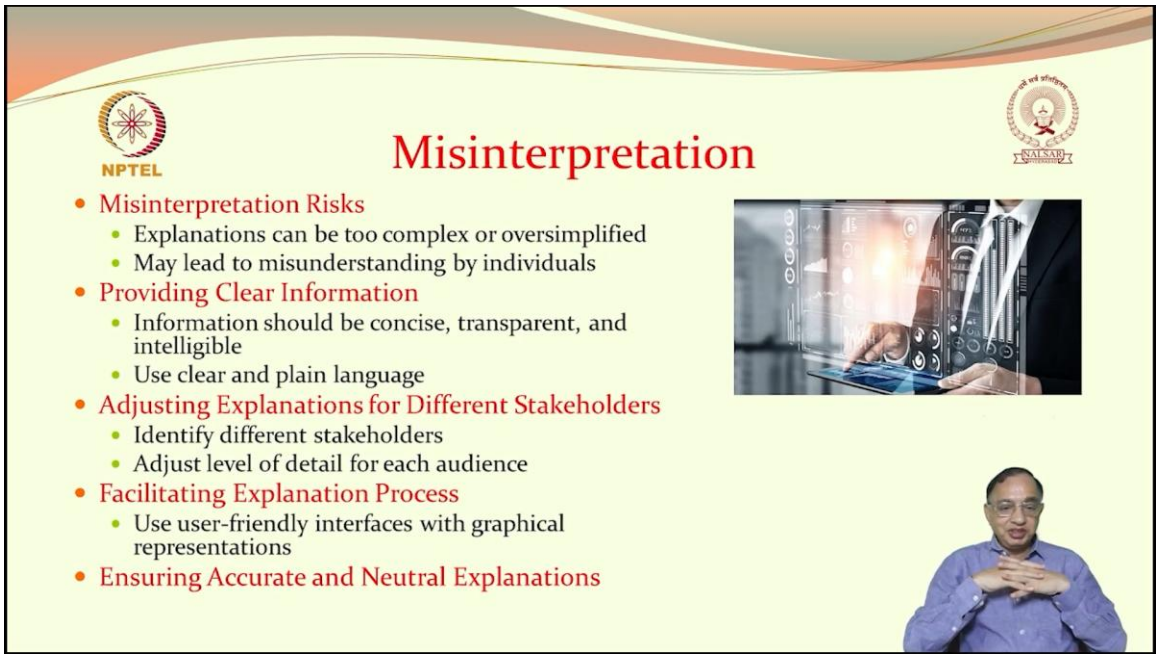
Then some categories of data, like high risk to privacy, are very important. Explainable AI (XAI) is equally important because the decision-making has to be really accounted for in Explainable AI (XAI). So, if high-risk privacy data is present, whether the data minimization principle was used or whether the data that was used was properly acquired under the DPDP Act or the GDPR Act, or whatever norms were invoked should be seen first to ensure they were really put into place and then implemented. But the opacity in algorithms can be a major hindrance here because it may not exactly disclose what data was used, particularly with the high-risk data that affects privacy. Often, what happens is that the systems identify correlations between attributes and data subjects.

In common understanding, we are data subjects or the subjects to whom the data pertains. So, they can identify correlations between attributes in the sense that there is a correlation between a person's income and his level of car ownership, or a correlation between a person's income and savings and the house he or she owns, or the correlation between a person's credit card transactions versus the person's overall financial position. So, a lot of correlations can be built in, in the sense that systems can identify correlations and then build a model that if a person is earning more than 10 lakhs, it is likely that that person will go for an SUV, sport utility vehicle, rather than an ordinary car. The system can find correlations and then say if a person has got an SUV earlier, it is most likely that, given the financial conditions remain the same, he will go to buy another SUV and not a smaller car. Similarly, these attributes and their correlations can be used for decision-making, but the question is whether the proxy attributes can infer specific categories of data.

Here, something needs to be addressed very carefully. We talked about the proxy data, so we need to really keep that in mind. For example, my postcode or pin code cannot be taken as a code for ethnicity, income level, or affluence. Similarly, my street can be

considered a form of proxy data in the sense that the organization I work for, or the bank with which I do my financial transactions, cannot be taken as a proxy to measure my financial worthiness. Therefore, if the systems are going to use proxy attributes as a substitute in decision-making, or when they look at decisions like credit reliability and whether the person is really eligible for an educational loan or housing loan, the data management and data protection aspects should also be taken into account. This should be very clearly mentioned: the decision was made on the basis of proxy data, either because the actual data was not available or it was not of good quality, or the proxy data was resorted to because it was the only one that was more reliable than any other available data.

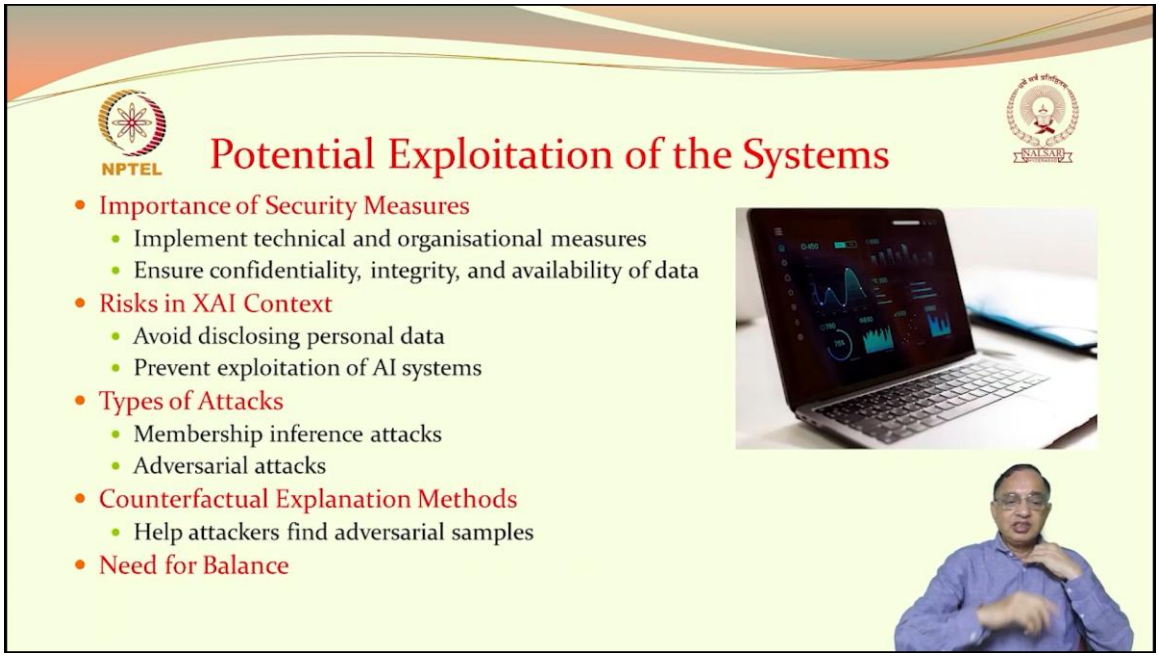
But Explainable AI (XAI) demands that these things be stated upfront when you talk about the explanation for the given decision. In the sense that some of these nice nuances or some of this nuanced analysis, which could be hidden or of which people may not be aware, should not be an excuse to make decisions that ultimately rely on impacting people negatively. So, the inferences that could happen on account of the wrong decision-making also need to be taken into account. So, to put it in other words, Explainable AI (XAI) is also linked with data protection in more than one way, and adhering to data protection norms will, in fact, help in adhering to Explainable AI (XAI) norms as well.



Then comes the question of misinterpretation. Explanations can be too complex or oversimplified. They can be misunderstood. They could be very ambiguous. So, we need to provide unambiguous, clear-cut, concise, transparent information that should be explainable in plain language. Then the explanation needs to be adjusted for different stakeholders depending on their level of understanding and need.

So, Explainable AI (XAI) is not a process that gives the same output for everyone, in the sense that certain stakeholders might need very detailed explanations given their level of

financial literacy and digital literacy. Certain stakeholders may not need such a detailed level; they will be able to understand, but their concerns will also be addressed. So, the explanation process itself should have user-friendly interfaces with graphical representations that are more like an interactive mode, and then the user's input should be gathered so that their answer can be tailored to the way the user wanted more information or indicated their input. So, the need for an accurate and neutral explanation is very important to avoid misinterpretation and also to ensure that Explainable AI (XAI) does not result in another black box in terms of explanations that are vague or ambiguous, making it difficult for a person to make sense of them.

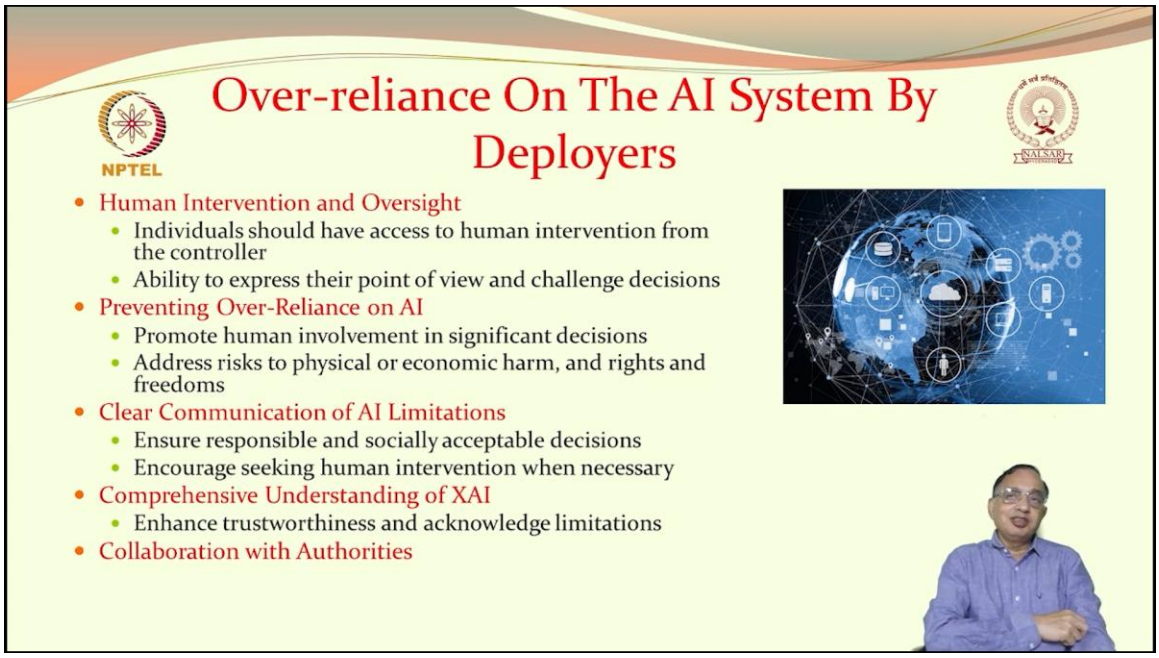


But we should also understand two things. When we make Explainable AI (XAI) a criterion and then want it to be built into the systems, the potential for the systems becoming vulnerable on account of that is also present because Explainable AI (XAI), when it is inbuilt, could be hacked; it could be technically insecure, and then someone could really play with the Explainable AI (XAI) system, tamper with the system, and make the system do certain things that it was not originally meant to do or corrupt the Explainable AI (XAI) system itself in such a way that what you get from the Explainable AI (XAI) system is something that is not understandable or totally wrong. So, the problem is that Explainable AI (XAI) content in the context should be understood, and then, in the name of Explainable AI (XAI), only relevant data that is important should be given; not all personal data should be shared or put in the public domain. In other words, Explainable AI (XAI) explanations and the information given should also adhere to data protection norms. More importantly, in the name of Explainable AI (XAI), the systems should not be exploited or used for data mining purposes by seeking more and more information in the name of Explainable AI (XAI). And there could be different

adversarial attacks; there could be membership inference attacks; so many attacks could be possible.

But the counterfactual explanation method is also equally important because once the counterfactuals are understood, the hackers can find adversarial samples and then try to inject certain things and hack the system using them. So, when we need Explainable AI (XAI), we also need cybersecurity. Or to put it in other words, an Explainable AI (XAI) system cannot come at the cost of cybersecurity. In fact, when we bring in Explainable AI (XAI) as part and parcel of the AI system, we need to doubly safeguard that cybersecurity is also in place so that the Explainable AI (XAI) system cannot be hacked and cannot be tampered with directly or indirectly so that it does not result in meddling with the overall system by a third party.

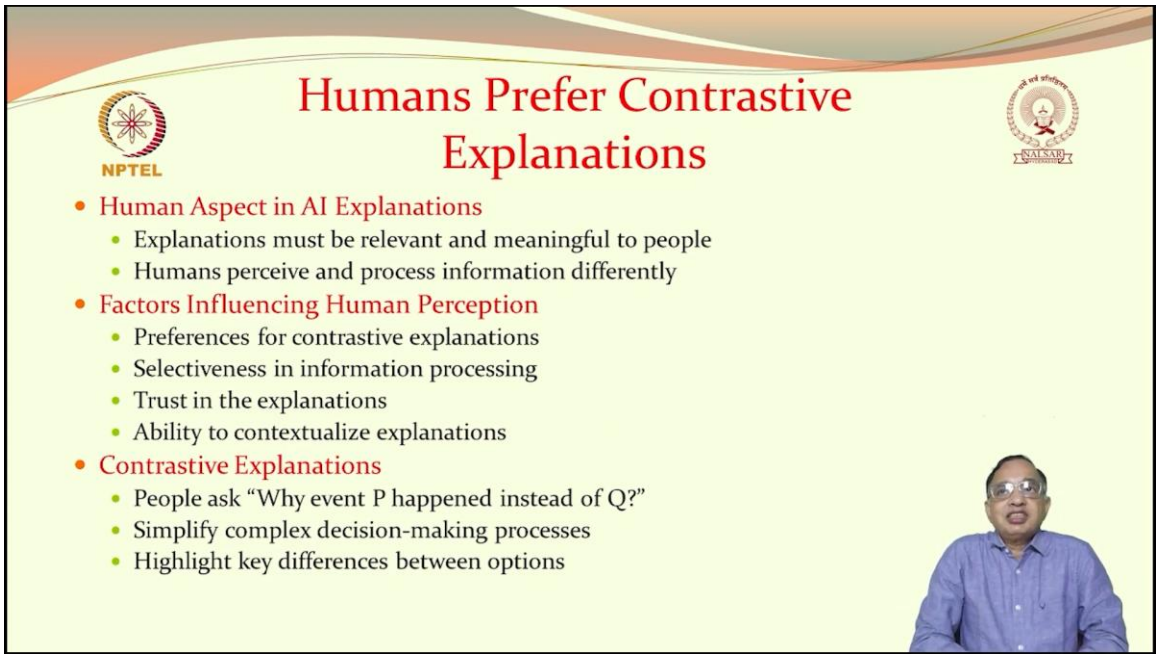
But as we saw in one of the earlier cases, the company refused to disclose it because they thought of it as a trade secret. So, what do we do in such a case? In such a case, we need to find a balance. We need to know exactly what is relevant and then what is important from the point of view of Explainable AI (XAI). If it is a trade secret and if the company feels that disclosing even a relevant part of it would really expose the trade secret system or third-party understanding and then really infringe upon it, then the court should provide direction on what part of the trade secret aspect can be made understandable to the developer so that the developer can work back and extract the real algorithm, which again relies on trade secrets to come up with Explainable AI (XAI) solutions and explanations. So, we should also take into account that in the name of Explainable AI (XAI), we should not build systems that could be used to infringe IP rights, tamper with systems, or meddle with systems in any way. So, whether it is the distortion of trade secrets, infringing IP rights, or anything else, Explainable AI (XAI) should be thought of as a good measure, but the potential for its misuse should also be taken into account.



But when there are too many systems in place, over-reliance on these systems also raises a lot of issues. First of all, for example, when we say to keep humans in the loop, the system can be designed in such a way that before the final decision is taken, for example, before the final decision of the housing loan is processed, the user should be able to intervene, understand the decision, and then challenge it. If the system states that it arrived at this particular facet or particular data, the user should be able to question it and, if necessary, provide more data or more explanation so that the system can rework it. Challenging the system should be part of that because Explainable AI (XAI) is also something that follows transparency and accountability. On the other hand, we should not take that AI system for granted and rely on it too much. So, we should prevent humans from relying too much on AI systems taken for granted and thinking that AI systems will always be right.

So, we need to do some sort of over-reliance prevention, which means that some form of sampling should be done, such that every 5th decision or every 7th decision in certain cases will be put to the test again, and then the whole process will be examined to understand why the decision was made. And then, depending upon the context, for example, if it is a loan granting system or lending system, the financial institution may say that for any such decision of more than 10 crores, the decision has to be reviewed by humans, and then the process has to be run by humans to cross-check everything and to ensure that nothing has been missed by the system. So clear communication is also equally important, but we need to bring a comprehensive understanding of Explainable AI (XAI). But Explainable AI (XAI) is not a panacea for all the ills or for all the problems with AI systems. Explainable AI (XAI) is something that can help them become more transparent, more accountable, and then more user-friendly.

But that is not the panacea; that is not the solution for anything and everything the system has in terms of problems. So we should know that Explainable AI (XAI) cannot be a cover to underestimate the system's problems and then give some explanations and say that everything is fine because we have explained it; rather, Explainable AI (XAI) itself can be something that the user should be able to challenge and then arrive at his or her own understanding and question why what was given to them in the name of Explainable AI (XAI) itself is not sufficient or is faulty.



But we should understand some psychology here. One, what sort of explanation do we prefer or really want? Do we want very contrasting explanations? Explanation must be meaningful and relevant to people. In the sense that when I talk about giving a person an explanation as to why he or she did not get a housing loan, I should not extend that and then say, given the economic conditions, and bring in irrelevant factors to justify that or to confuse the person.

And then different people have different capacities, different perceptions to perceive information, and then to understand and process further. So that should also be taken into account. In the sense that if it is not relevant and meaningful, or if the Explainable AI (XAI) process itself distorts certain things and the explanation is meaningless or misleading, then it can't be treated as Explainable AI (XAI). But human perception can be manipulated in different ways. For example, humans, on account of cognitive biases, may wish not to see an elephant when it is there.

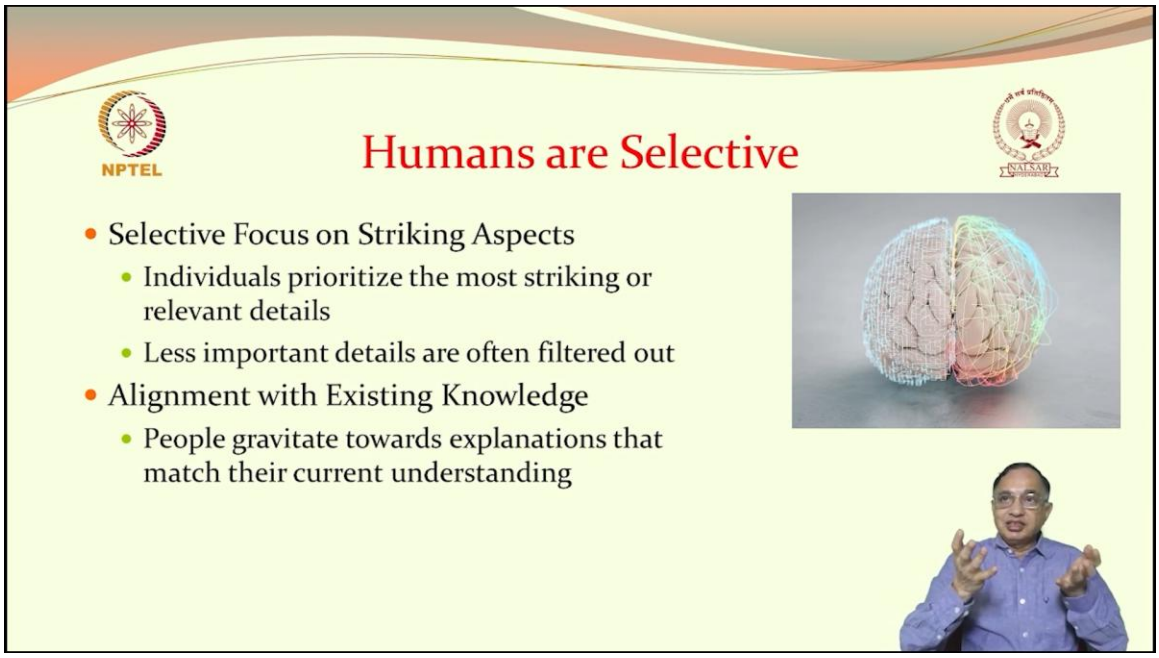
So, our cognition is again conditioned by our biases and prejudices. A person whom I do not like may well be sitting before me in the chair, but then my perception would be that I should ignore him, and I will not even look at his face. Similarly, people will have selectiveness in understanding certain data or looking for certain data that they think is favourable to them rather than examining the whole data, where certain things may not be

favourable to them. So, the trust in the explanation again depends on how trustworthy it is, how much of that is understandable, and how much of that is credible. And more importantly, the explanation should not be vague; it should be contextualized.

Mr. X did not get the loan because of the 1, 2, 3, 4, 5 factors that are related to him. It should not go to the level of 'we are processing 10 crore applications every day; of those, we only give loans for 5%, and of that, we provide the final disbursement to only 2% after fulfilling all the conditions. Since you did not come in the first 5%, we cannot even consider your application further.' That sort of explanation should be avoided; it should be very contextual, aligned to the person's needs and specificities, rather than giving a broad explanation that 95 percent of the people getting their loans rejected is something that is understandable and should be explainable. I mean, that level of logic cannot be used when we talk about Explainable AI (XAI) in decision-making when it comes to individuals. So, the person should be able to contextualize that decision in their life conditions and then either accept it or challenge it.

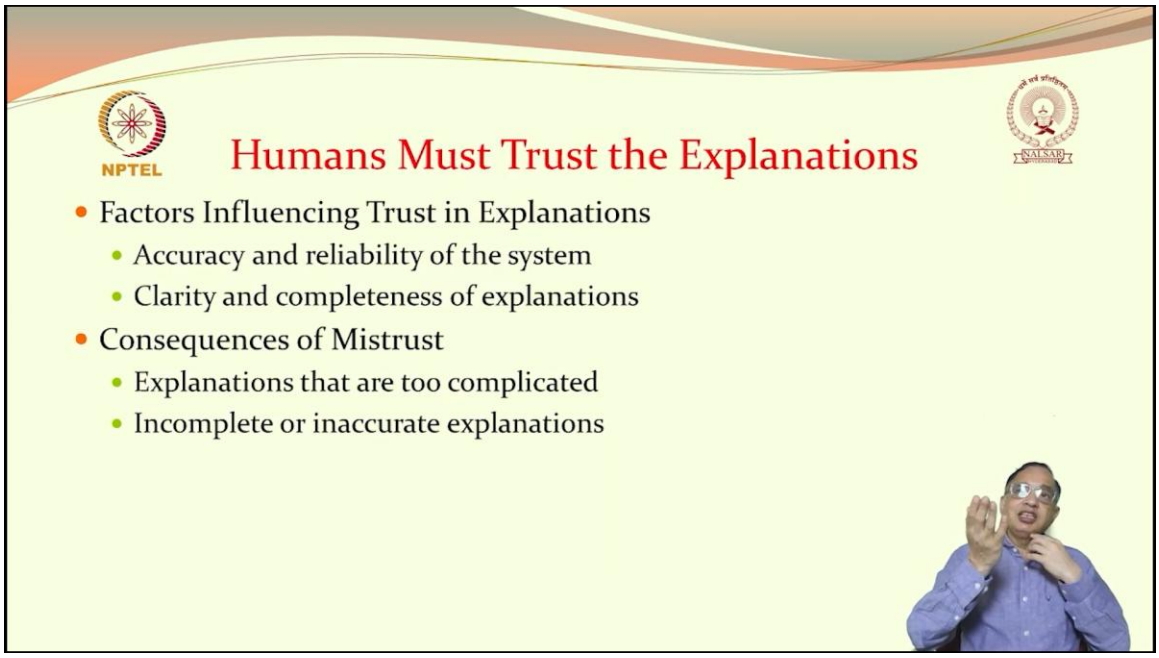
Often the question is whether it is about sequencing, when things happen in parallel, when things go in a flow, or when things happen as a part of a process. The question is why this happened instead of that; for example, why didn't the driver apply the brake instead of not applying the brake, or why did the driver take the wrong turn during the U-turn, or why didn't the driver use the horn, or why didn't the driver alert the passengers that the vehicle was going in the wrong direction. Or why that driver, knowing fully well, did not stop at the signal although the signal colour was red. So, people would ask why event X happened instead of event Y because event Y is the one that would have been the normal response of any average intelligent person.

So, some complex decision-making processes have to be simplified. Now the driver would not have stopped when the signal was red because he would have seen a lorry behind him that was coming all the way to crash or to create an accident. Similarly, some decisions taken in the spur of the moment could be a violation, but then they would be able to justify them on account of the specific facts and the specific context. So why this happened instead of that again needs to be brought into the Explainable AI (XAI), particularly when the systems make decisions.

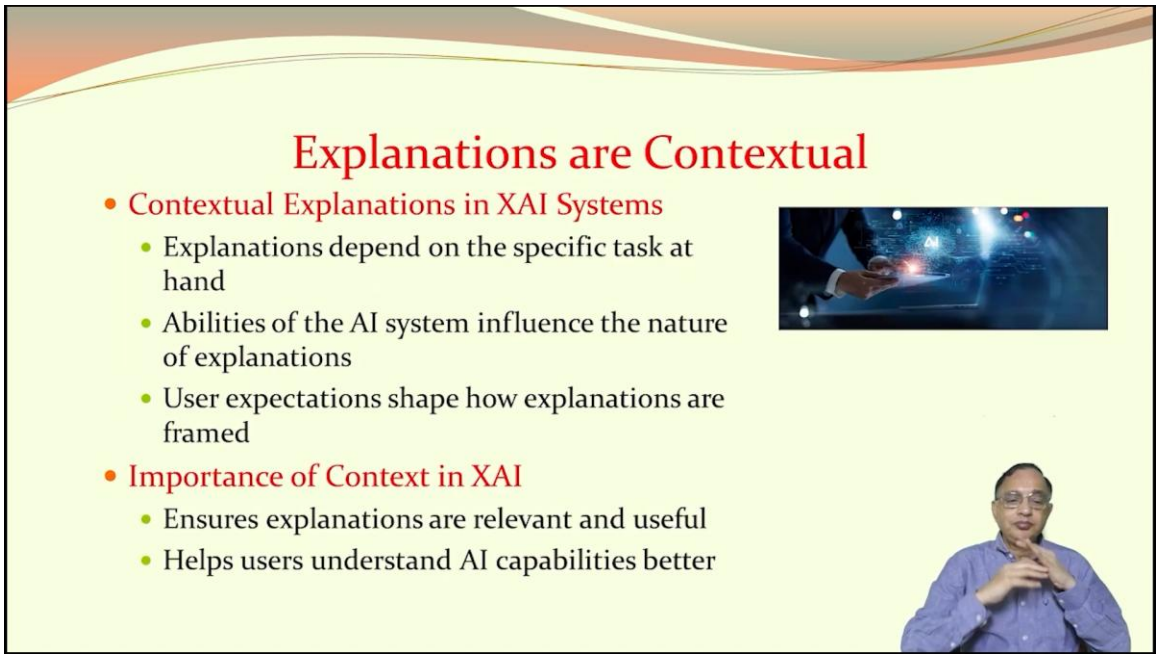


So ultimately, what sort of explanations does humans prefer and what sort of understanding we are dealing with is equally important. If you tell me 10 things, I will not pick up all 10 things; I will pick up only what I think is interesting and relevant to me. In fact, I will not even bother to listen to six of the things you said; I will pick up only the three that I find relevant or the seven that may be something I am not comfortable with. So, people selectively process information, people selectively understand information, and then arrive at decisions and inferences based on them. Often, when so much information is being fed to us, we tend to look at what is more relevant or important for us rather than trying to grasp every piece of information. So, people have the tendency to filter information based on their priorities, needs, and relevance. So, the human being selectively aligns with whatever they try to understand or learn, with what they have already known, experienced, or expected; those explanations would make more sense to people when put in the right context of their understanding and expectations.

For example, in a case of a housing loan where the person has not been given the loan, I should move towards an explanation that justifies it based on some financial parameters and then based on certain things that the person is acutely aware of, rather than trying to justify it in broader economic terms and then saying that banks are not giving housing loans because the housing sector itself is in doldrums. Additionally, many banks are not giving house loans today as the non-performing assets in the housing sector have increased from 25% to 45%, and I should not try to present irrelevant information in the name of Explainable AI (XAI) just to convince a person that this decision was right. So, we should look at what decisions people will understand and what their selective processing capacity is, and then what we need to tell them based on what exactly they are asking for.

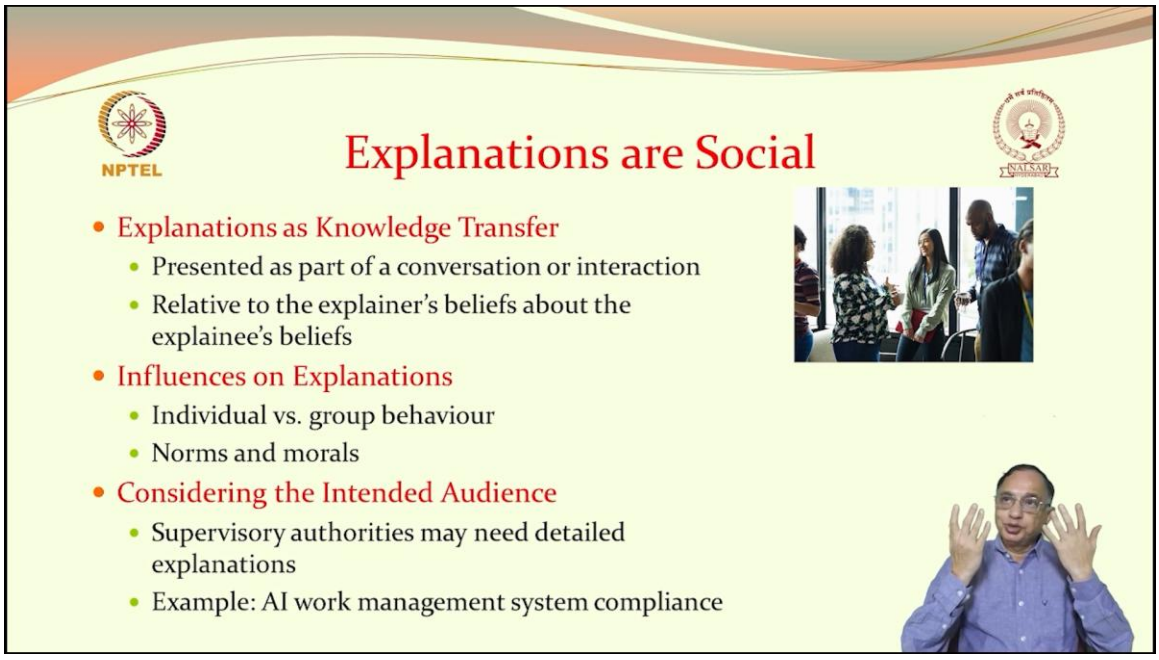


But the trustworthiness should also be built in, in the sense that if I can make the system explainable to a great extent, people can understand it, appreciate it, or may not agree with it, but then they can say that it is reasonable, fair, and justifiable; then trust will be enhanced. But when that does not happen and when Explainable AI (XAI) gives an explanation that does not foster trust or that is not trustworthy or credible, then the trustworthiness of the whole system could become something problematic for people to deal with in the future.



As we said, explanations need to be contextualized, but a specific task at hand also matters, so if the user's expectations are there, we need to take that into account and then understand and tailor our explanation based on the user's needs and aspirations. So, if we

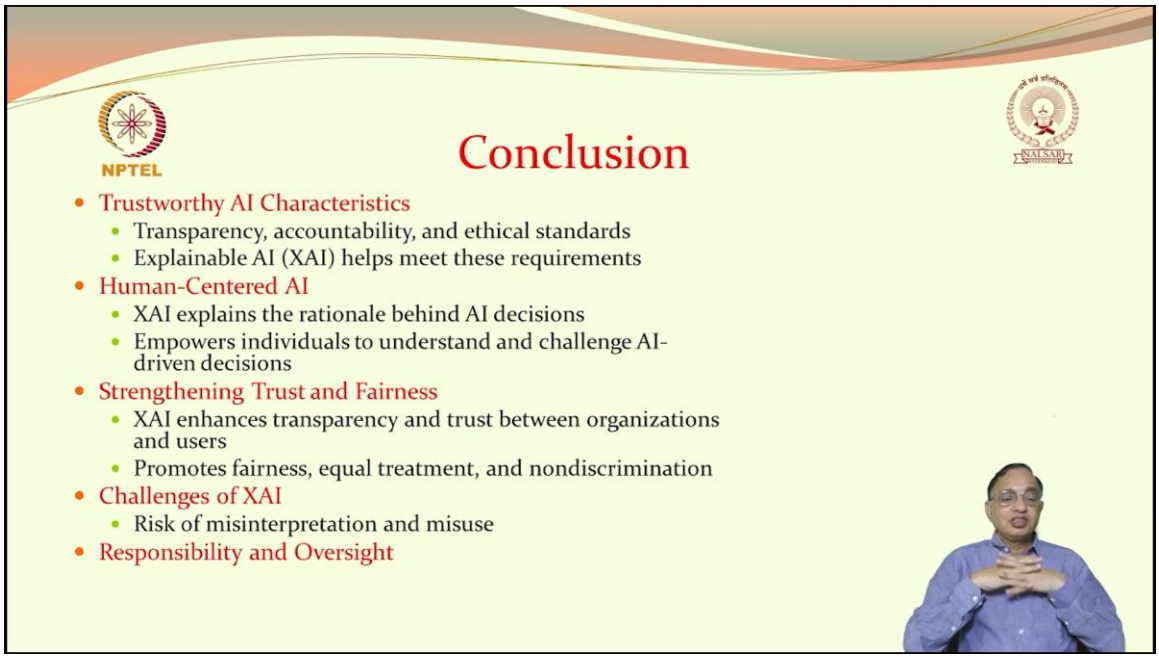
can build a contextualized Explainable AI (XAI) that can meet the needs of the user, that would be much better than having a bland, generic Explainable AI (XAI) that doesn't address the individual needs and concerns or that is not context-specific. So moving towards Explainable AI (XAI) that is context-specific is a challenge for the simple reason that an Explainable AI (XAI) system has to really factor in a lot of things; it needs to consider not just the explanation alone, but also how to make it easily understandable, contextualize it, and present it in such a format that the person who is asking for it is able to understand it within his or her context.



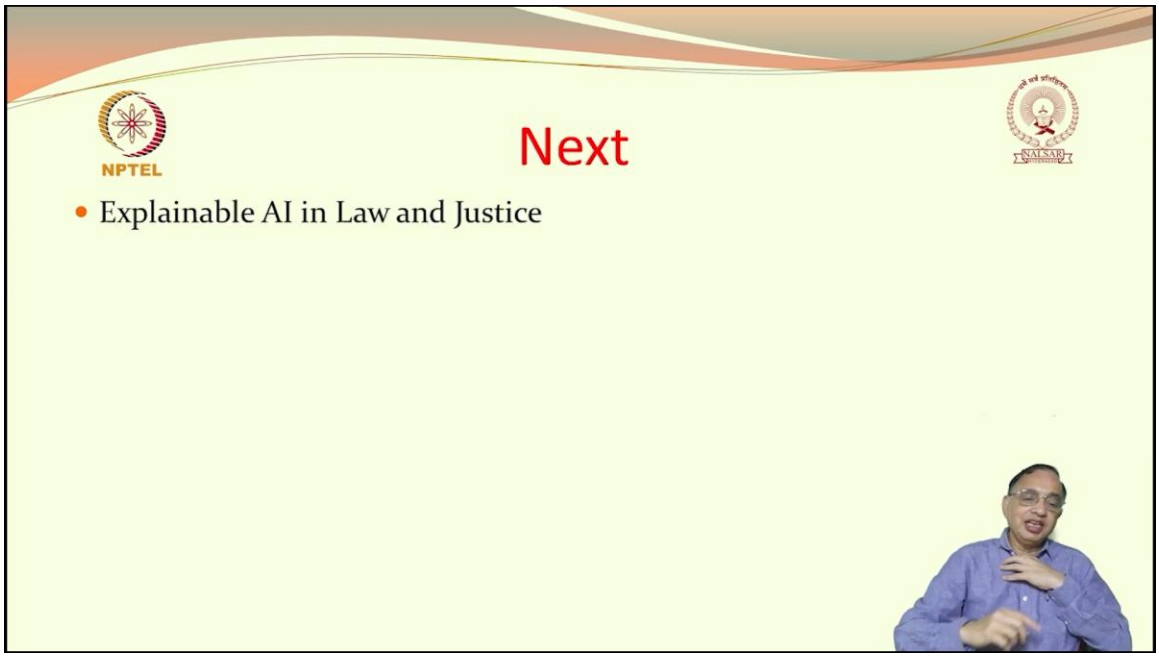
Then the explanations are social. We are social animals; we are social beings, and then we look at others, we look at ourselves in the image of others, we look at how much loan they have got, how much loan I have got. So, we compare, we contrast, we try to imitate, and we measure. So, the Explainable AI (XAI) should take this into account and then say whether the Explainable AI (XAI) makes sense in the broader social context of that.

In a system that is going to process 100 applications from a group that applied for a loan, if the system is giving the loan only to 12 people and then rejecting 88 people, then why did these 12 people alone get the loan out of the 100? Why were the 88 rejected? This is something that would bother the group. As a group, we need to look into their expectations, their needs, and understand why the 88 could not get the loan. This needs to be explained to them in such a way that it does not appear that the system is biased towards the 12 or that the system wantonly disregarded the other 88 and decided not to give them loans. The real issue, or the real thing that could be acceptable here, is that only the 12 could meet the requirements, and they were picked as eligible. For the 88, the explanation should clarify that these were the real difficulties and defects; if they could be addressed, then another round of discussion or decision-making can happen. So, we need to look into the audience, the group, or the broader context in which the Explainable

AI (XAI) system should work, and then the broader group to which we are talking. So, we need to really look at Explainable AI (XAI) not as a technical solution but more as a communicative experience or as a process of communication, with a dialogue also built in for a wide variety of users.



To conclude, we need to build Explainable AI (XAI) systems because they are very important to enhance credibility. But more than credibility, we need to build Explainable AI (XAI) systems because they also help us decipher what could go wrong or what has gone wrong with the systems. So, they will be useful in multiple ways to strengthen the trust and fairness of the system. And then they also enhance the human-centred framework in AI in the sense that humans will feel that they are not something that the system simply considers as statistical information or something the system reduces them to—some sort of indicators put together, as in this ranking or in this system of classification or a criteria put everything together; for example your rank is 0.912345, whereas the normal, which I expect, is 2.453, so that sort of thing can be overcome with Explainable AI (XAI) systems, which can bring in more transparent, easily understandable explanations. But as we said, there are also challenges in them in the sense that an Explainable AI (XAI) system can also be a potential threat. It could be something that could also be a vulnerability. So, this needs to be factored in. And more importantly, bringing in explainable AI should not be thought of as a one-time effort, but rather as part of broader responsibility and oversight over AI systems. Or in other words, the responsibility and oversight of the systems should have an Explainable AI (XAI) component irrespective of whether the systems are functioning or not.



So, in the next class, we will go into what exactly we mean by Explainable AI (XAI) in law and justice. Thank you.