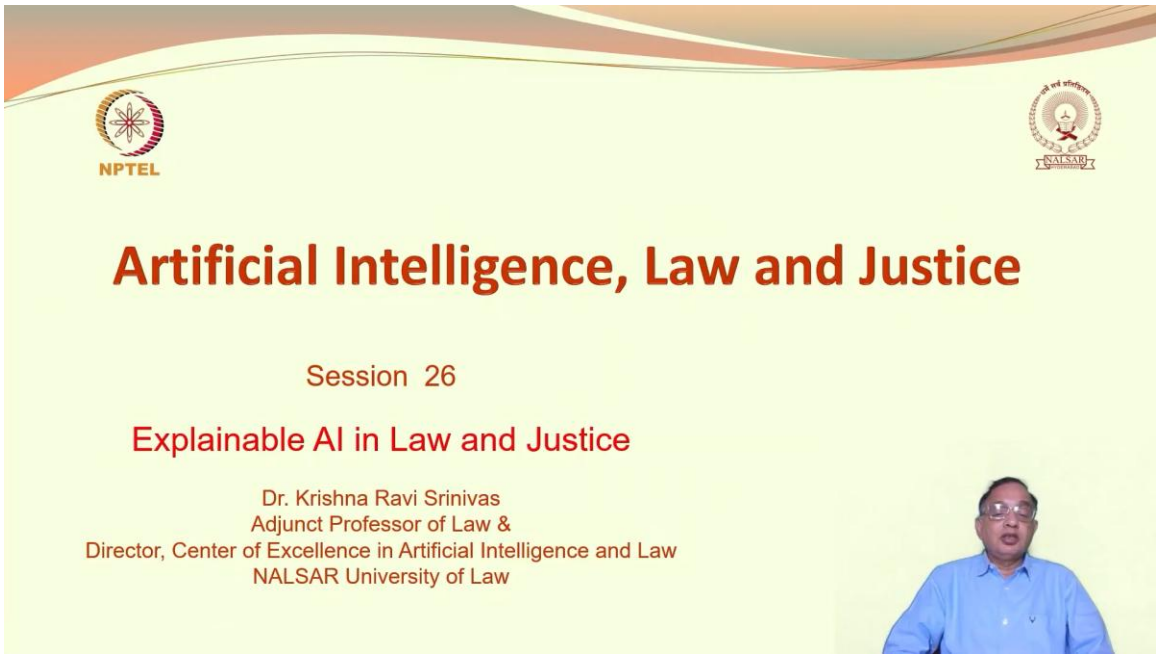
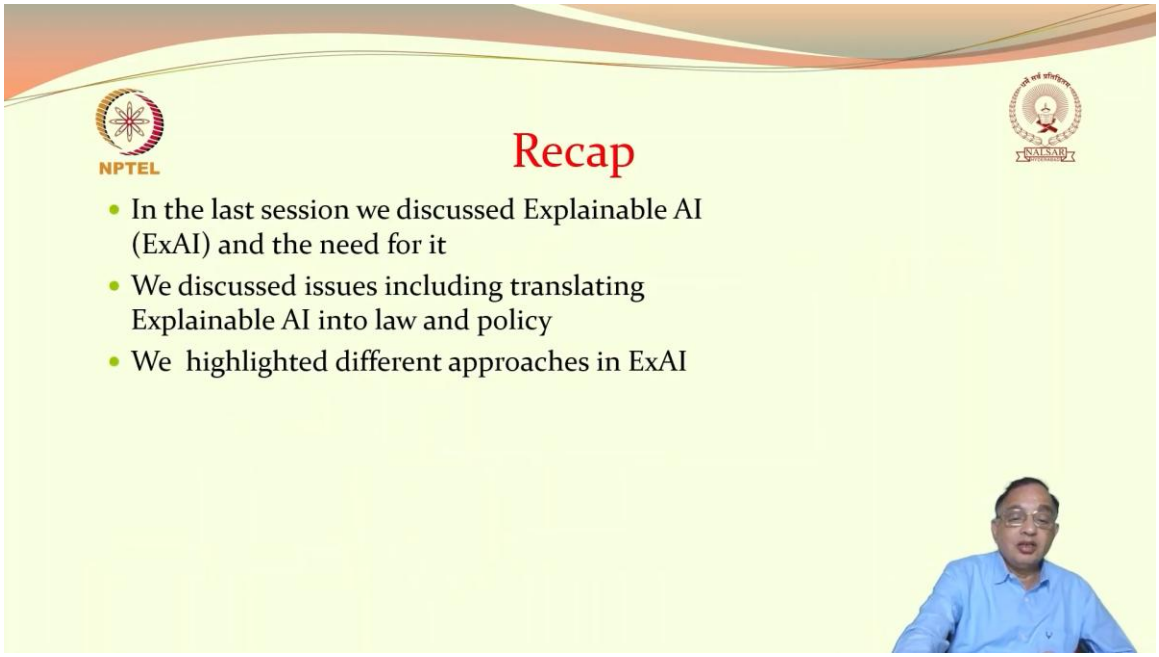
Course Name – Artificial Intelligence, Law and Justice
Professor Name – Dr. Krishna Ravi Srinivas
Department Name – Center of Excellence in Artificial Intelligence and Law
Institute Name – NALSAR University of Law
Week – 06
Lecture – 26

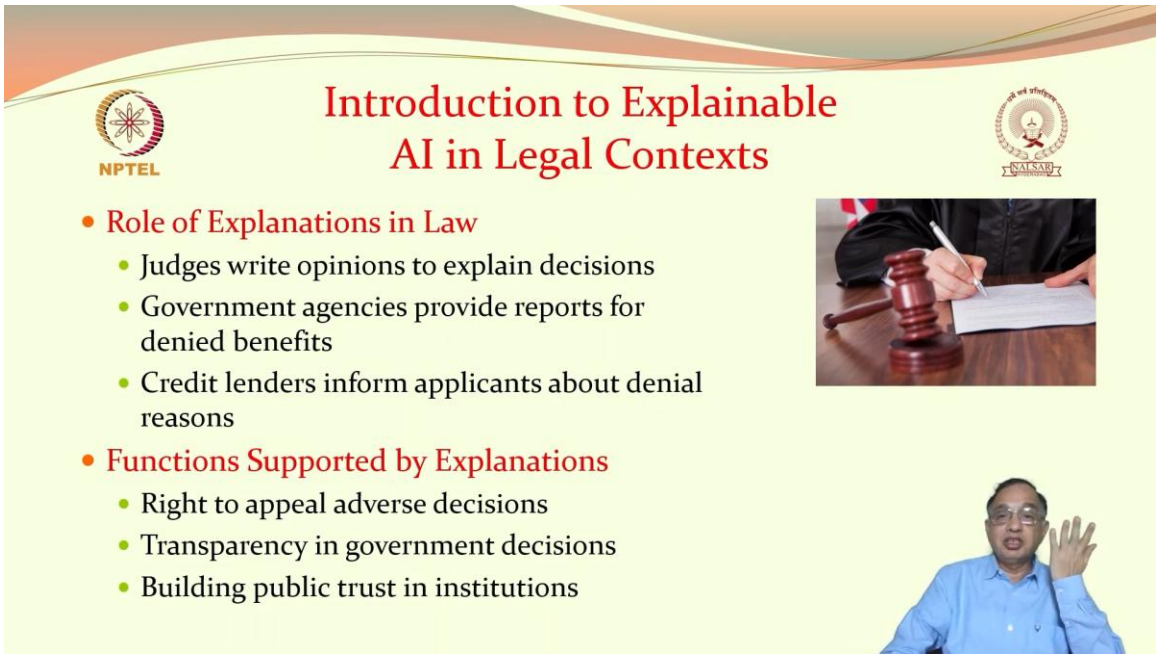


Artificial Intelligence Law and Justice Course Session 26: The topic here is Explainable AI in Law and Justice.



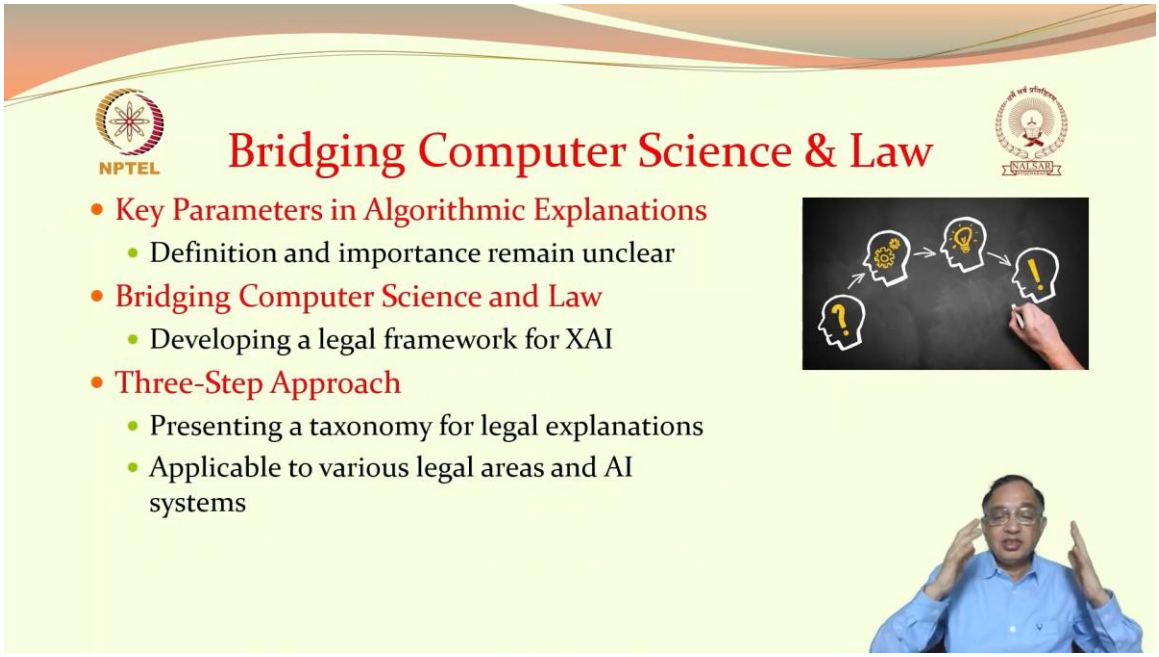
In the last session, we discussed Explainable AI (XAI) and the need for it; we also discussed issues in translating that into law and policy, and we highlighted different approaches to Explainable AI. In this class, we will move from a very general perspective

to very specific issues seen in law and justice, and then discuss why we need XAI in law and justice.



The role of explanation in law is very important because judges write pages of judgments trying to tell us why they decided the matter in such a way, what the rationale is, and more importantly, what the binding precedents are, which case laws they referred to, what the legal principles are, and what the constitutional principles are; and then even if the judge dissents, that dissenting judge also has to record his or her reasons in detail. So, it connects to the right to appeal, and transparency, people only understand things when they are explained to, or only when they know why; or can say that they know why this decision has been made this way. So, that they will be able to either appeal it, accept it, contest it in part, or contest it in full. So, explainability is part and parcel of legal decision-making, and it is inevitable that any use of AI in law and justice has to take this into account.

But as we have seen, there are a lot of risks associated with automated decision-making systems. We have seen it again and again, and one of the ways the legislators and policymakers come to grips with that is to let the ADMs, or automated decision-making, or algorithmic decision-making, or a black box; that black box cannot be allowed to function as is; we need to open it up, we need to pierce the veil. So, the California Privacy Protection Agency has come up with draft rules where they say businesses should inform the logic behind the decision-making and provide an explanation of key parameters, and then application's decisions should be made known to the consumers.

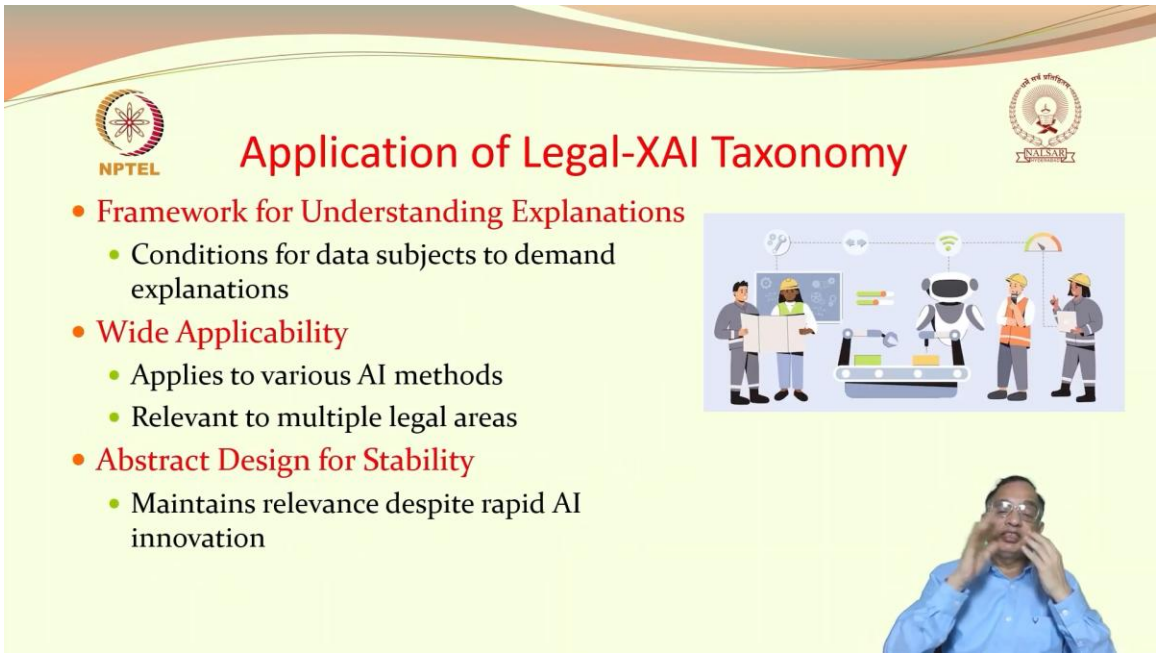


On the other hand, this is not a very simple task because we need to bridge both computer science and law. Computer science is a hardcore scientific technology discipline, while law is a discipline that is more social. So, how do we bring them together? One way to do

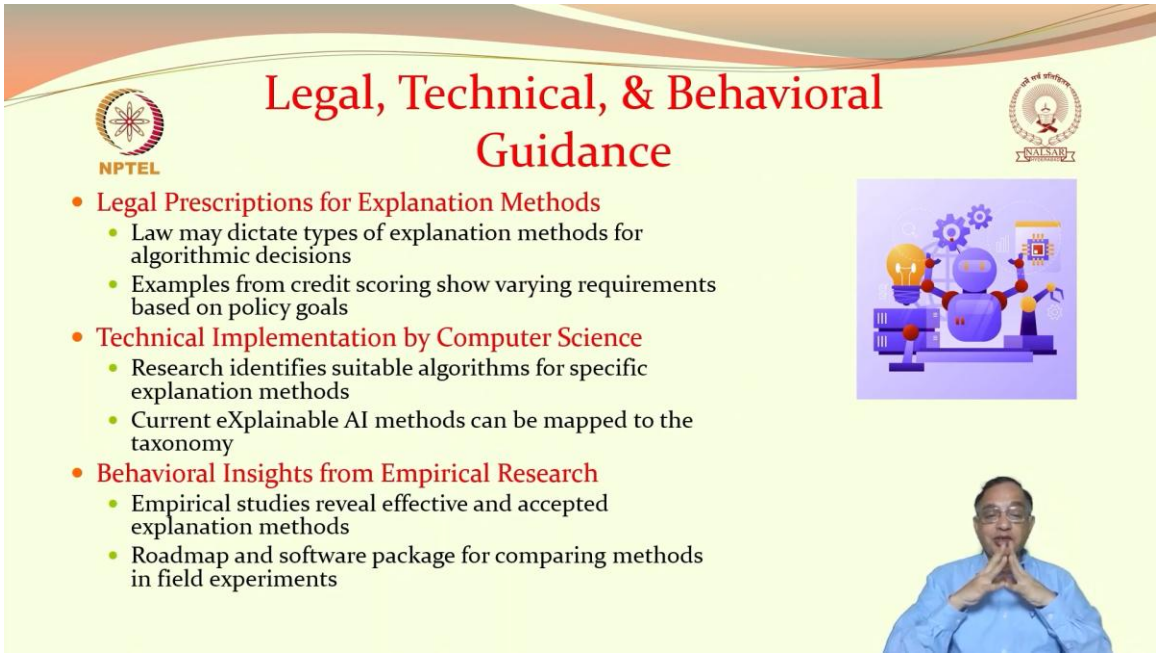
this is to understand the key parameters and algorithmic explanations. However, as of now, we are not very sure of that because the definition's importance is not very clear. Within the whole idea of bridging computer science and law, we can develop a legal framework for XAI or create something to bridge the two. So, presenting a taxonomy for legal explanation is feasible, and we are discussing one such taxonomy in this class. And then this taxonomy can be applied to different areas in law as well as to AI systems.

So, the legal explainable AI taxonomy, which we are discussing, delineates the key factors for the practical implementation of explainable AI. So, it looks into different types of categories in the taxonomy, whether the taxonomy's scope is local or global, which we saw in the last class, and then depth, whether it is non-contrastive or contrastive, and then what alternative explanations are available. Then, whether you want to look at the explanations as conditional or correlational. Conditional is something that occurs if a condition x is met; if this is met, this happens. So that is conditional. Whereas correlation is when you link up certain things, whether 4 or 5 factors, and then correlate the decision with that. So, when you look at correlations, it is different from conditions because correlation, again, we can say these are closely related or that there is a statistically significant relationship; therefore, a decision was arrived at. Then what is the scope and depth of properties we want to understand? And then the factors related to information presentation are also part of the XAI taxonomy.

So the global explanation will look into the general broad model behaviour of the LLM, how it functions in the broader perspective, its broader decision-making capacities, and algorithms. This is important to understand from a larger perspective, particularly for regulatory compliance purposes. In the sense that you want to look at what the insurance system does, what the banking system does, what the payment system does, or what the education system does in a broader way, the global explanation is fine and very important. But when it comes to very specific instances where individuals and groups are affected, we need to go further down the line and then look at where exactly the module behaves, in the sense of how the module reasons, takes decisions, and then is able to explain on what basis. So, to understand the individual predictions as well as decisions, this is important; for example, in healthcare diagnostics, how exactly the diagnostics AI or the tool arrived at determining whether this disease is cancer in the second or third stage, or in the preliminary stage, or if this is exactly the disease or condition. Similarly, the same applies to criminal justice regarding whether bail should be granted, a person should be placed on parole, or what exactly the sentence for such an offense should be. So, the global will give a broader macro perspective, whereas the local will give you the micro perspective, but you need to balance both, or you need to understand both not as contrast but as part of the larger ecosystem.



Having said that, we need to look into how we apply this legal explainable AI taxonomy. Then what are the frameworks for understanding explanations: conditions for data subjects to demand explanations? First of all, the data subjects are people who are using this, or people who are either stakeholders or the ones who are going to be impacted by it. So, what exactly should they know? This should be applicable to different legal areas, different AI methods, and then the design should be abstract, but it should be stable enough because as AI innovation happens at a rapid speed, it should be able to meet subsequent demands and needs as well.



So legal prescription is something that we need to understand. For example, law may prescribe or dictate a certain type of explanation for algorithmic decisions. If you take a typical credit scoring system, it should state what the various requirements are based on

policy goals. For example, if I am going to a public sector bank, public sector banks are governed by the policy goals of the government. Then, if I am going to apply for a loan under a specific government scheme, I should first know whether the bank has accepted and adopted the specific norms when it processes my application. So, the legal perspective's explanation is that it should be aligned with the policy goal. But if the algorithm did not take that into account and then treated mine as a general one, then I would not be able to benefit from the specific rules under which I would have benefitted. So explainable AI should be able to explain to me why this was given or why this was not given. Then the computer scientist and the data scientist should be able to decide what exactly the suitable relevant algorithms are for applying to certain explanation methods. So, they should tell us how we need to map the algorithms, explainable algorithms with the overall taxonomy.

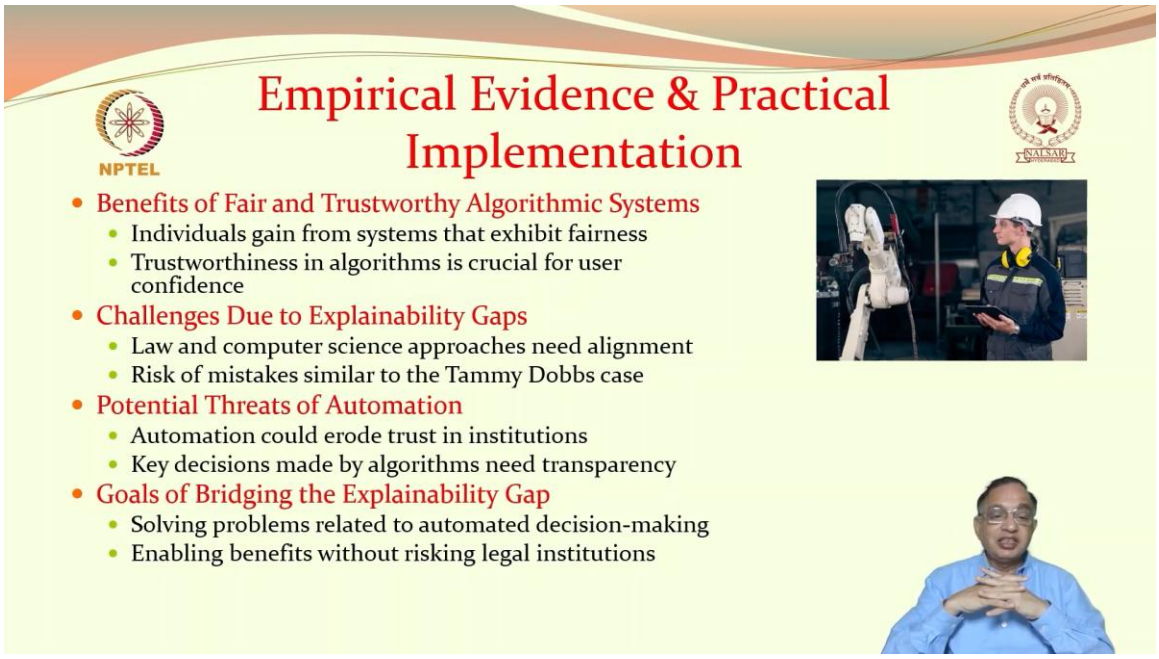
Then when we do a lot of empirical research, we will know which explanation models are more accepted, which are more desired, or what the stakeholders prefer, whether they prefer one explanation model over others and, if so, for what purpose. So, the software package for comparing models from field experiments, where we do a lot of field studies and then ask stakeholders for their perspectives, builds the system with their inputs regarding what they need from explainable AI. We need to look at that behavioural guidance so that when we test the system and later run the system, we know that this could be the explanation often sought, or this is the reasoning, and then this is the alignment with the policy goals or alignment with the overall loan processing method people will be able to relate to. So, it is not merely a question of giving explanations in terms of technical details; it needs to go much further than that.

For example, in Medicaid: Medicaid is the U.S. scheme where people get access to hospitals, something like a government system very similar to what we have in India, Ayushman Bharat, but the categorization and other things are very different. Similarly, when it comes to higher education, there could be a system, or there is a system, where

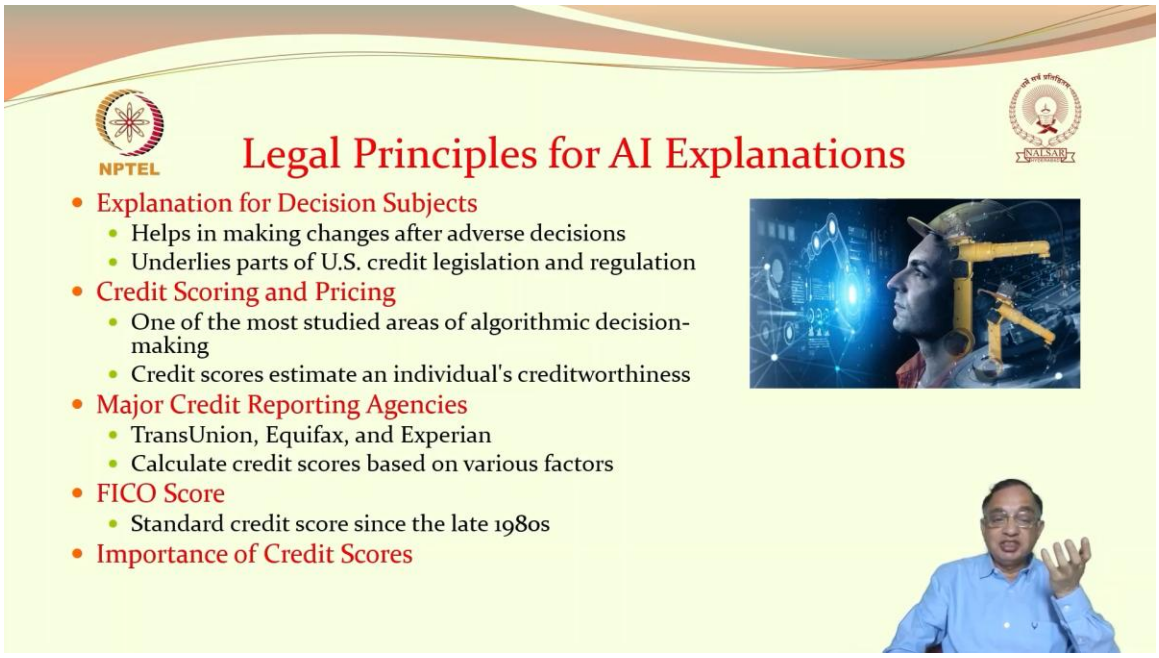
you can automatically apply for a loan and then get benefits, or you can automatically, using your student enrolment ID, be able to get some benefits. So, these are areas where automated decision-making may also be possible, but some Acts specifically mandate explanation or explanatory systems as part of the overall regulation. For example, the EU AI Act seeks to explain AI as part of its overall framework. Similarly, when California tried to regulate AI, the idea was to come up with an explainable AI as part of the regulatory mechanism, but this law was reversed in the sense that it didn't go further. So, implementing the AI taxonomy in the real world is something we really need to take into account, and taxonomy, broadly understood, is something like a structure.

So, we need a lot of interdisciplinary work. We need to look at the current laws and models because the problem is that laws are meant for humans, and translating those models into technical reality will be very difficult. Decision-making models do not focus on data subjects' interests because data subjects' interests are not the idea of digital decision-making models. The decision-making models are driven more by algorithms, statistical patterns, and then statistical consistency with the overall objective of giving a decision, whereas the law does not take those algorithmic decision-making processes into practice; it treats everyone as equals and as persons who are entitled to equal rights and human rights. So, unless the law itself very specifically says that under these circumstances bail cannot be given, or a loan cannot be given, or this condition cannot be weighed, incorporating them into the technical thing should be a factor. But then the algorithmic decision-making, thinking, and rule-making will be very different from what the law normally says. Then there is a potential contradiction, if not conflict. So, we need to really have empirical evidence as to how the algorithms work in practice, and then we need to also assess how the explanations have been really effective. So, we need to do things first. We need to look at what the effective algorithmic explanations are when they are put to practical use and then what the framework is for assessing the explanations' effectiveness in the sense of how effective the explanation given by the algorithm is. So, for policymakers, both are important, and then computer scientists need to look at the

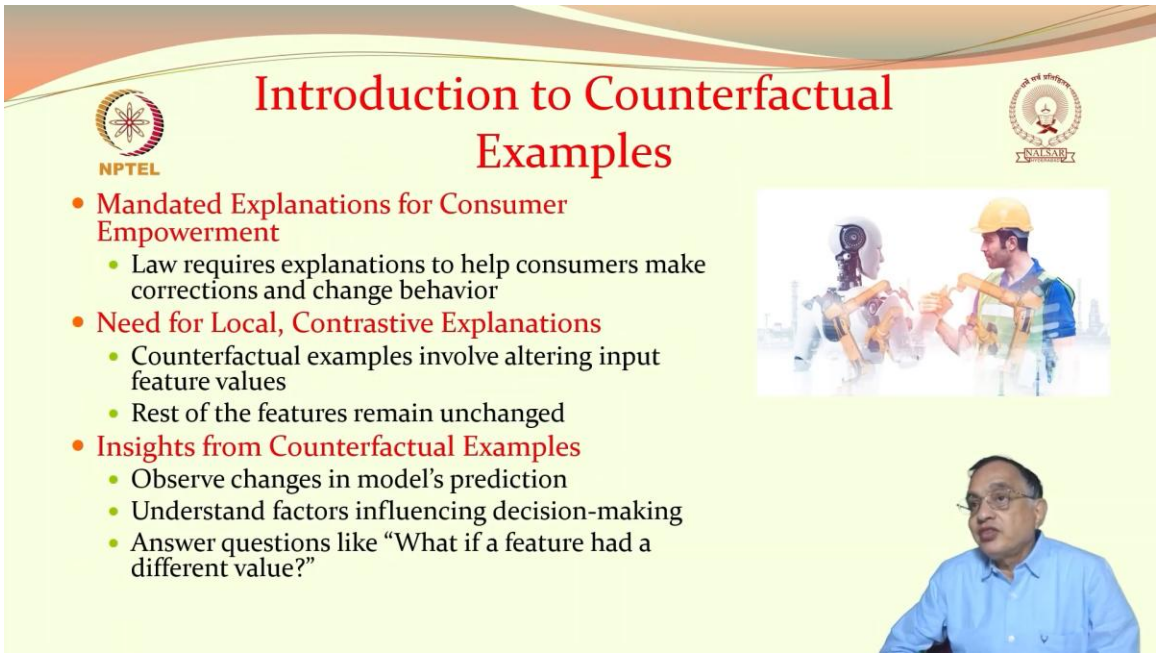
intricate legal issues involved, as well as how the system interprets certain rules and whether that interpretation itself can be a direct or indirect violation of the rules.



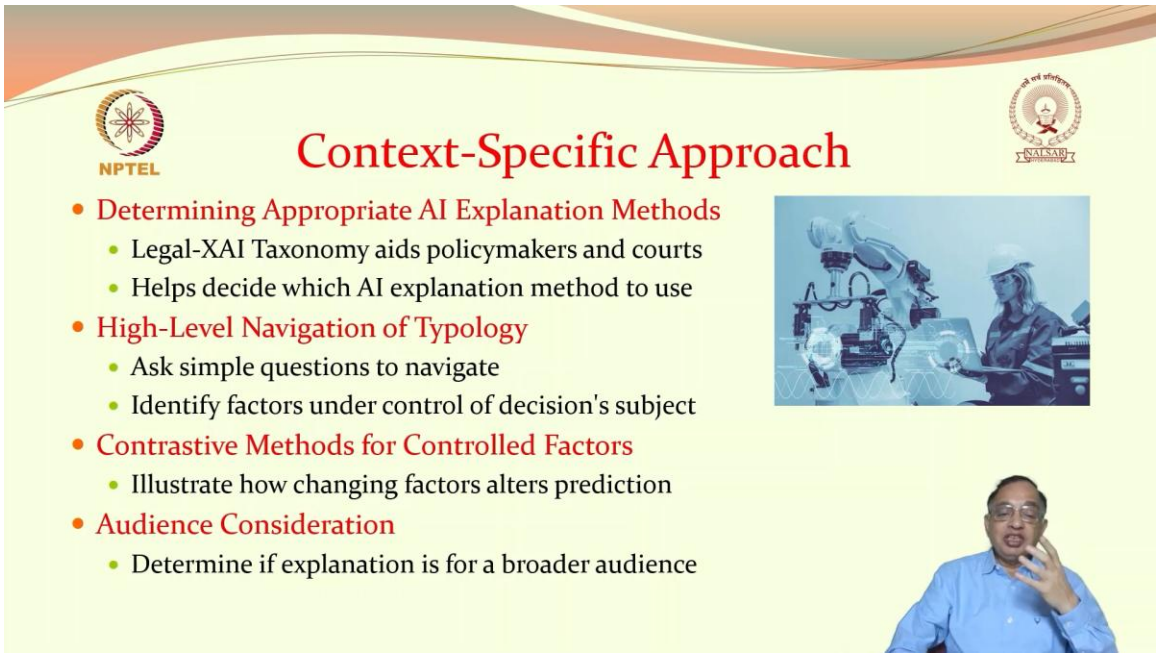
So, we know the empirical evidence shows that a lot of things can come from it. It can result in fair and accurate systems, and then the explanatory gaps need to be really filled in. But when we go for automation and algorithmic decision-making at all levels for all types of decisions, it will erode trust in institutions because people will ultimately know that it is algorithms, then machines, and then some crazy computers that run the organization, not people. So, the explainability gap will always be there when automation continues, and people are either not fully comfortable with that idea or are unable to have explainable AI brought in to give them convincing reasons why these decisions are made this way. So, we need to bridge the explainability gap by making Explainable AI part of any larger AI system, and we should also enable benefits for the people, but it should be done without violating any legal rules or laws.



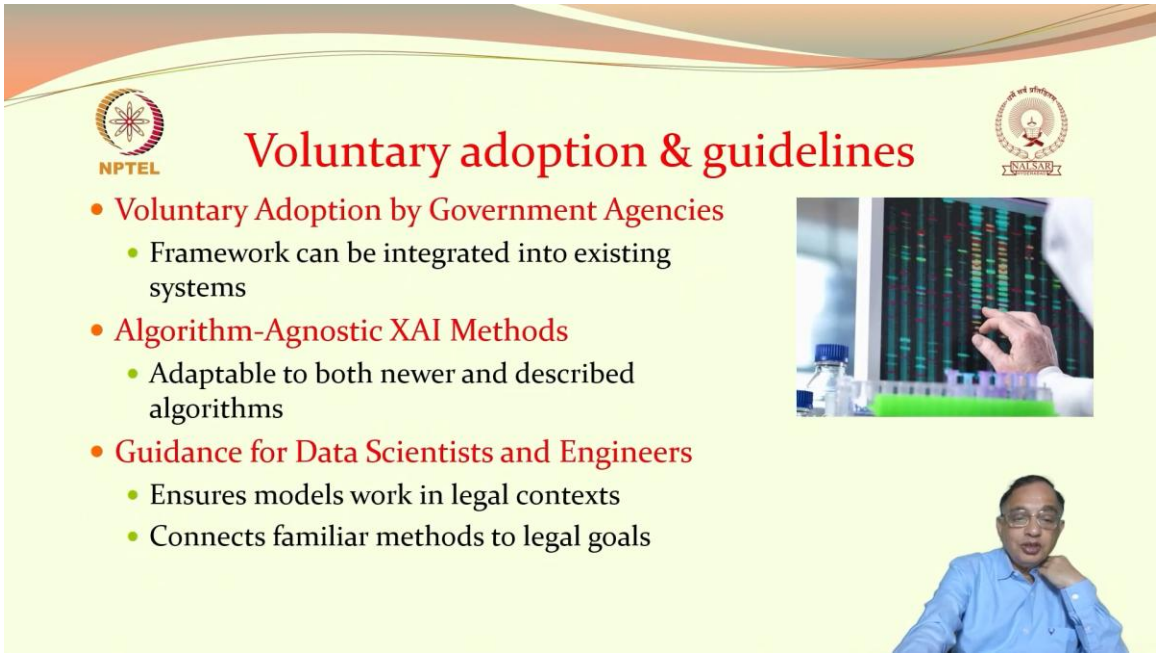
Then the legal principles for AI explanations are very important. People should be able to understand changes after adverse decisions. So, whether it's credit legislation or anything else, people should understand why the decision that was averse to them was taken, particularly because credit scoring and pricing are among the most studied topics in the literature. Algorithmic decision-making often scores people's credit history, ranking them in terms of their eligibility, ineligibility, or eligibility with a higher interest rate. Additionally, pricing in the sense of whether you're going to go for an insurance policy, health insurance, or car insurance is very important. So, when you assess creditworthiness based on algorithms, we also need to look into whether the algorithmic decision-making makes it difficult for some people to buy on credit, obtain certain insurance premiums, or access some insurance benefits due to the higher price. This is an example taken from the USA, but there are many credit reporting agencies. There are also a couple of agencies in India. So, they share the data with the bank, and then the banks also process a lot of information that they internally generate. And then these things are going to the decision-making mechanism. And then there are agencies in the U.S. and elsewhere where they look at the creditworthiness and credit history of individuals, so the credit scores ultimately can make or break the decision-making. How do we assess that the credit scores are being used without any bias and discrimination, and whether the credit score system itself is being rightly interpreted and applied is something that Explainable AI will be able to help us understand.



And then counterfactuals: earlier we also had a quick look at the counterfactuals. Law will need that consumers understand all the issues in the sense that if the counterfactual says, "had it been this way" or "had it been that way," the counterfactual example I give to you can help you identify why you didn't get it. So, I can change my behaviour, I can change my policy, or I can change my needs, or I can go for a lower-level loan and then try to get it again. So, the counterfactual example should be part of Explainable AI not to give a wrong impression to the consumer but to give the consumer an idea that to understand this, they should look at the counterfactual perspective and then get a better view of why this decision is reliable and justifiable. So the counterfactual examples often result in the observer or the person who is impacted trying to get a better sense or trying to understand things more, and then they also understand which factors were more influential in the decision-making, whether his previous history, his overall credit score, or whether it was that his previous history was not all that satisfactory, or changes in his financial status over the last 18 months, in the sense that that person lost a job and was not able to get employment for some time. So, he had a lot of credit card spending dues but couldn't clear them, so the person would get an idea as to why he or she was not granted or was not given that loan. So, if the factors that influence decision-making were available and known, people could change their behaviour; people could be careful in the future.

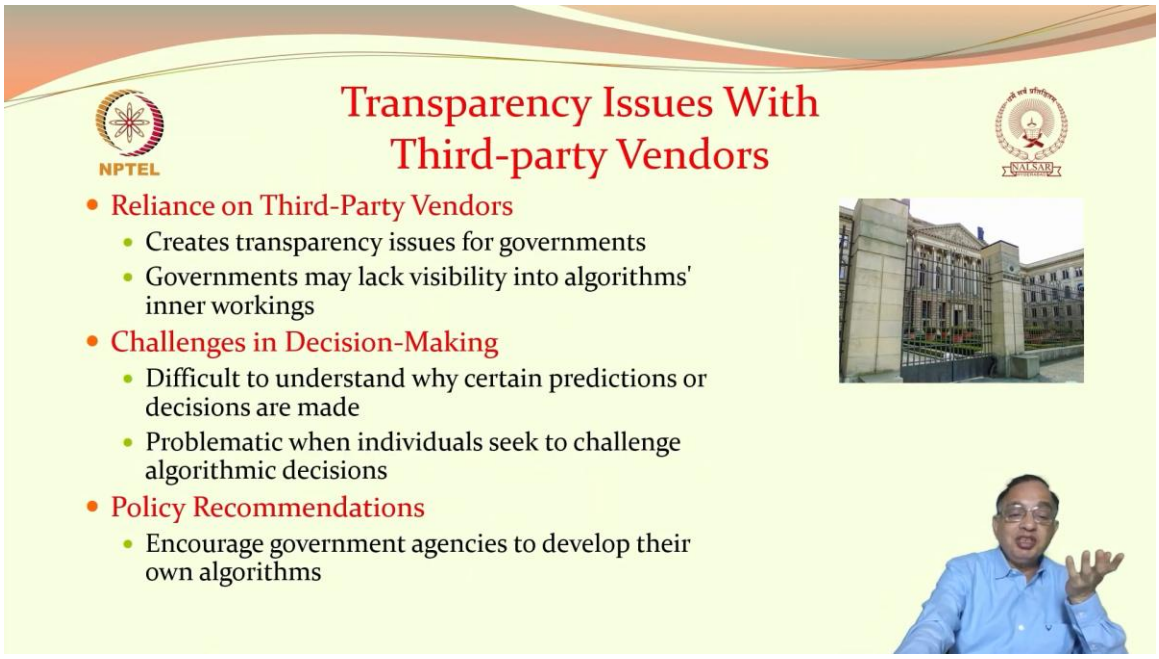


But explainable AI should also look into very context-specific applications. It should also look into what sort of audience it is trying to convince. So, the legal XAI taxonomy that we are talking about should be able to guide policymakers and courts in determining which Explainable AI model or method they should use.

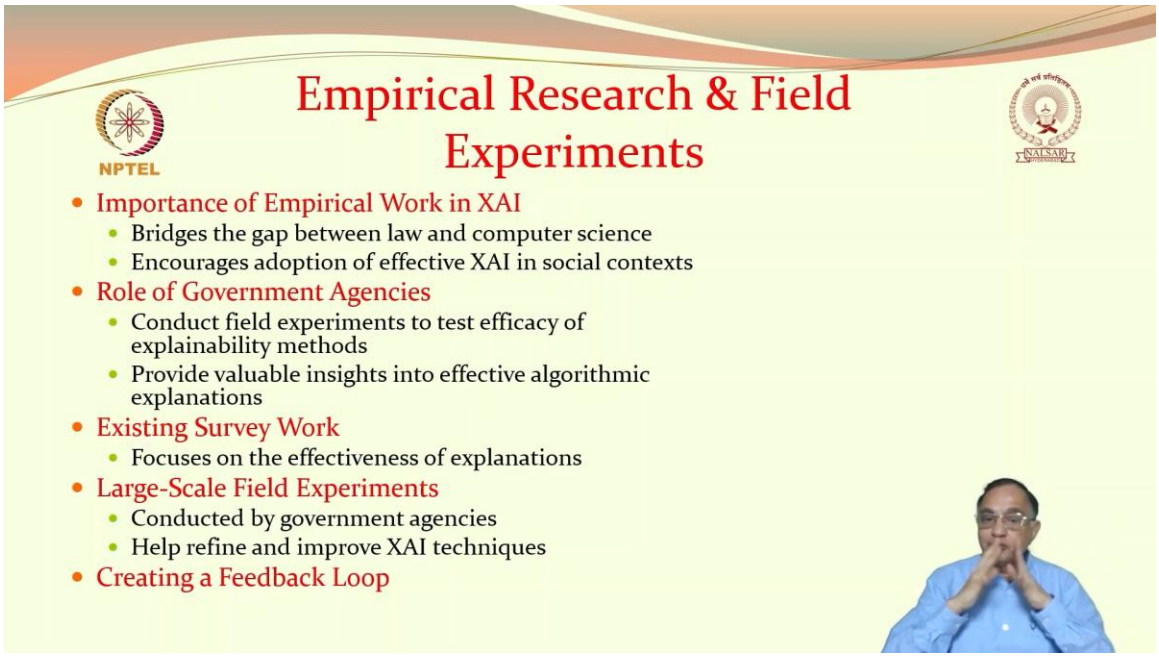


And voluntary adoption: Voluntary adoption of this model by government agencies will help them to understand. But it is also important that some of the Explainable AI models should not be algorithm-dependent in the sense that they should not be tailored to certain patterns of algorithms or certain algorithms. Having an Explainable AI model that can be applied across algorithms is better, and such models can be called algorithmically agnostic in the sense that they will work with all algorithms, but they are not crafted for any specific algorithm or type of algorithm. So, it is here that data scientists and

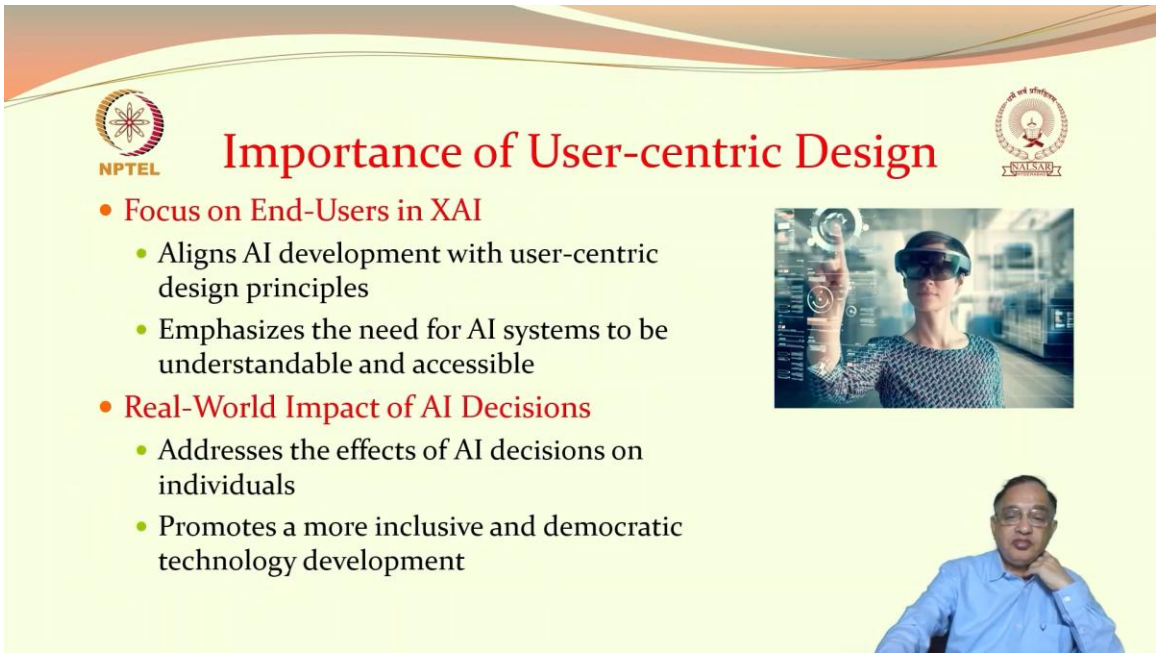
engineers, when they develop a system, particularly one that comes with Explainable AI, should be aware of these issues.



Then transparency is a major issue because when governments want to engage in algorithmic decision-making, they should also understand that they have the right to ask for transparency, and they should have the right to understand how the decision-making process is being interpreted and implemented. More importantly, they should also understand what to do if the decisions are challenged in a court of law, in a consumer forum, or in some other forum. So, when government organizations use court, administrative law, education departments, banks, or consumer-facing departments, they should also be aware of the transparency issues in third-party vendors that have significant legal implications. Government agencies, if possible, should develop their own algorithms or at least have a strong understanding of the algorithmic decision-making process, and then they should also try to rework certain aspects for their own benefit and strive to understand them better.

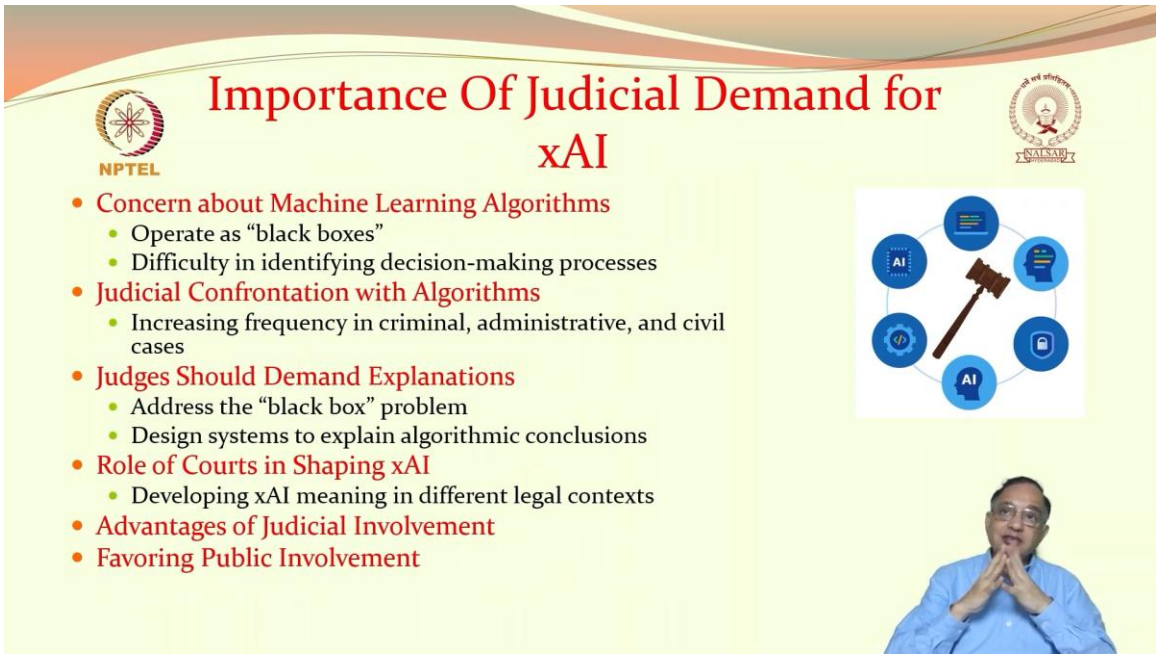


So, we need to do a lot of things here; first of all, the government agencies cannot take Explainable AI for granted; they need to do certain field-level testing. They should test whatever models were given to them to assess their accuracy, reliability, and acceptability with a wide range of stakeholders. And then they should also look at whether people are convinced by these explanations or whether the algorithms, when they make a decision, are able to come up with explanations that sound reasonable and legally acceptable. So, they need to develop some methodology so that they can conduct empirical research and field experiments with these Explainable AI models, algorithms, and so on, so that they can improve them. They also gain a better understanding of how they can deal with Explainable AI-related queries, demands, and even notices from courts.



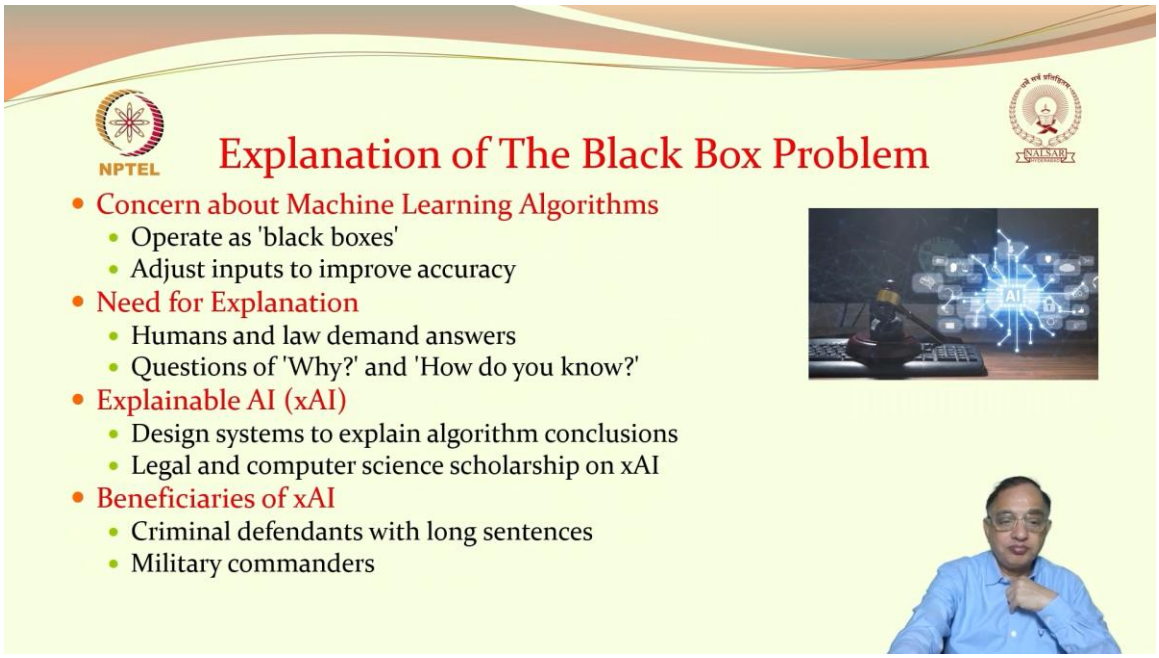
One way to address many of these issues is to ensure that the AI system is not built on some abstract model developed for some abstract purpose; rather, it should be built with a user-centric approach aligned with the user-centric perspective, where the user is not a simple piece of data or a statistical correlation but a person with rights and expectations. Then, it will be easy to incorporate Explainable AI as part of that. For example, the user-centric design in AI systems will give importance to providing the right explanations to the user; it will also take into account what sort of potential biases and discrimination can creep in when the system runs, so it should have already been tested, or algorithms should have been tested for that. And more importantly, when we build a user-centric system, the design will also mean that we are democratizing technology development. We are making technology more understandable, and then users can also contribute to it. So, making a user-centric design, particularly in AI development and in the context of Explainable AI, is a good policy. But then it is not all that simple to do so. For the simple reason that the systems are not meant for users' prima facie. The systems are used or meant for organizations that want to implement them. So, when the organization that wants to implement them, they may prefer technical efficiency, cost efficiency, and economic efficiency over user-centric design. So, this is an issue that the developers and those who want their systems to be developed and installed have to discuss and solve.

So, lots of field experiments may be needed for test robustness, applicable framework and then relevance and effectiveness across the legal domains. So, we need to develop what I would say is not just a huge case law but a library of all these experiments done, all the outcomes that were achieved with the different methodologies that were used, and the different data sampling methods that were used to provide an understanding of where exactly we can make Explainable AI fit well with the overall development and deployment of AI in law and justice.

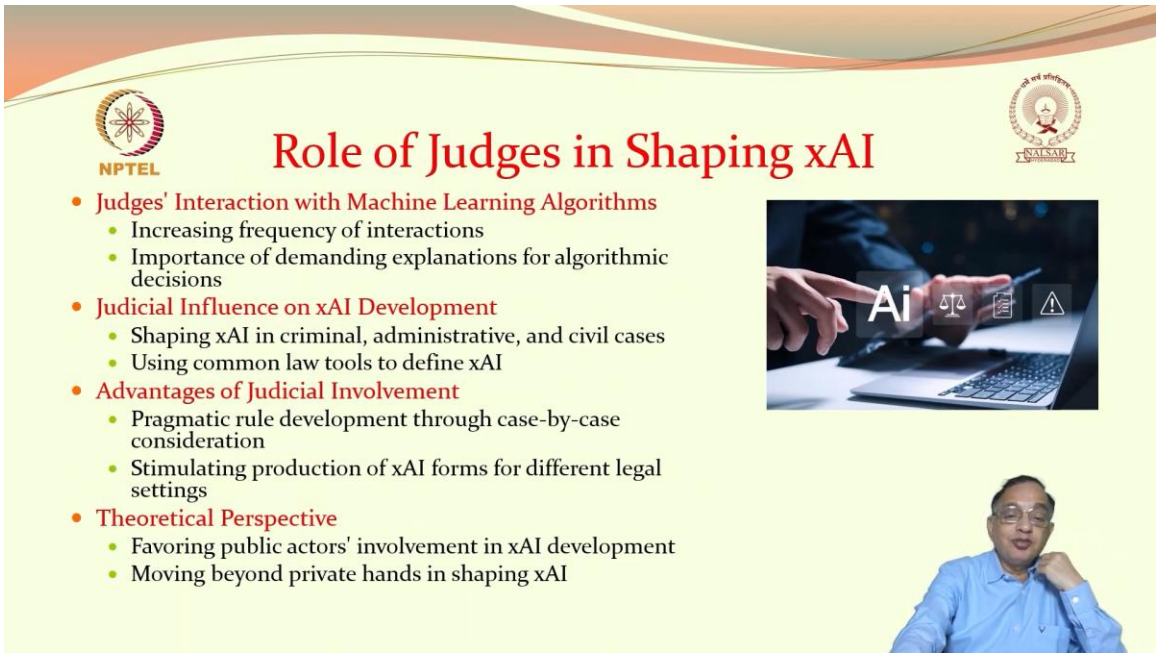


Now, the demand for Explainable AI should naturally arise from the judicial system, but it is not happening that way, or often the judiciary may not feel that it needs Explainable AI as a key component of the AI systems. It could be that they may not be able to really understand, and then they may think that this is something beyond our comprehension.

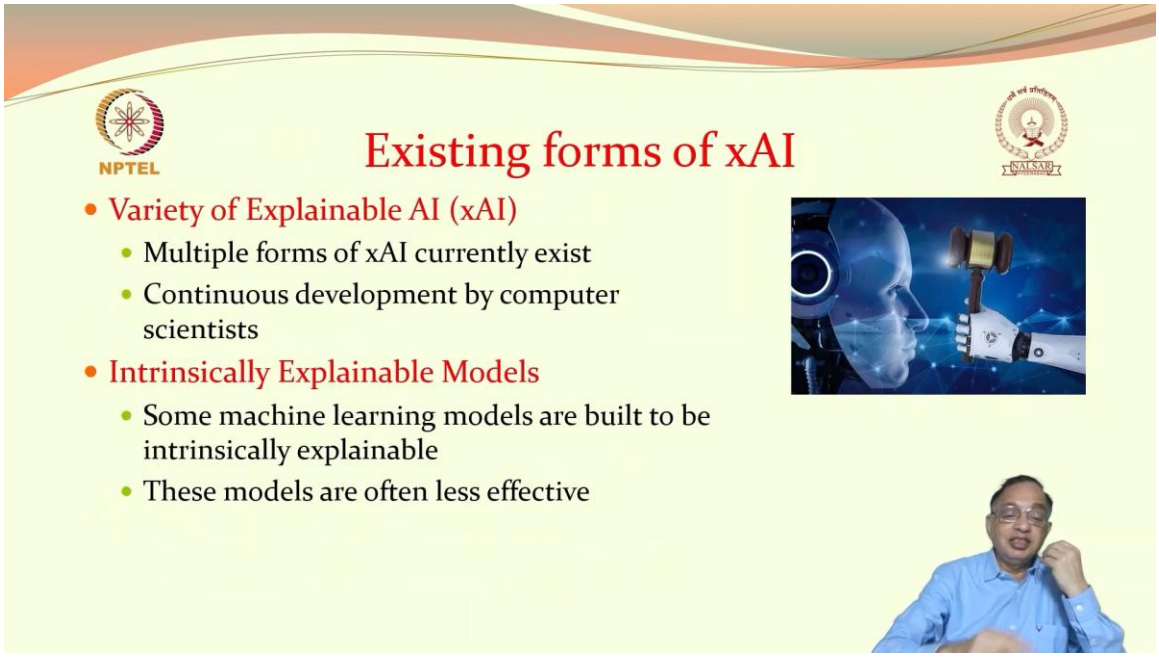
We should not break our heads, and Explainable AI itself may not be something that we will be able to really comprehend; it could be too technical. Then, when judges confront algorithms, they may think, "I know intuitively that something is wrong," but then how should I address it? How should I convey that my intuition is correct? How can I prove my intuition is right? There should be a mechanism where judges can really ask for Explainable AI solutions as part of the black boxes. And then they should also know how the system can explain to them the conclusions and then go in the reverse direction to tell them why step-by-step decision-making was done and then where exactly the decision tree finally took this decision in contrast to the other decision. So, the role of the courts here is very important. They can play a huge role in developing effective Explainable AI, provided judges are also keenly interested, work with the developers and the deployers, and, more importantly, are empowered to deal with them effectively instead of making judges mere lame users of the systems. So, this, again, as we said earlier, is part of the public involvement, and it also enhances judicial involvement.



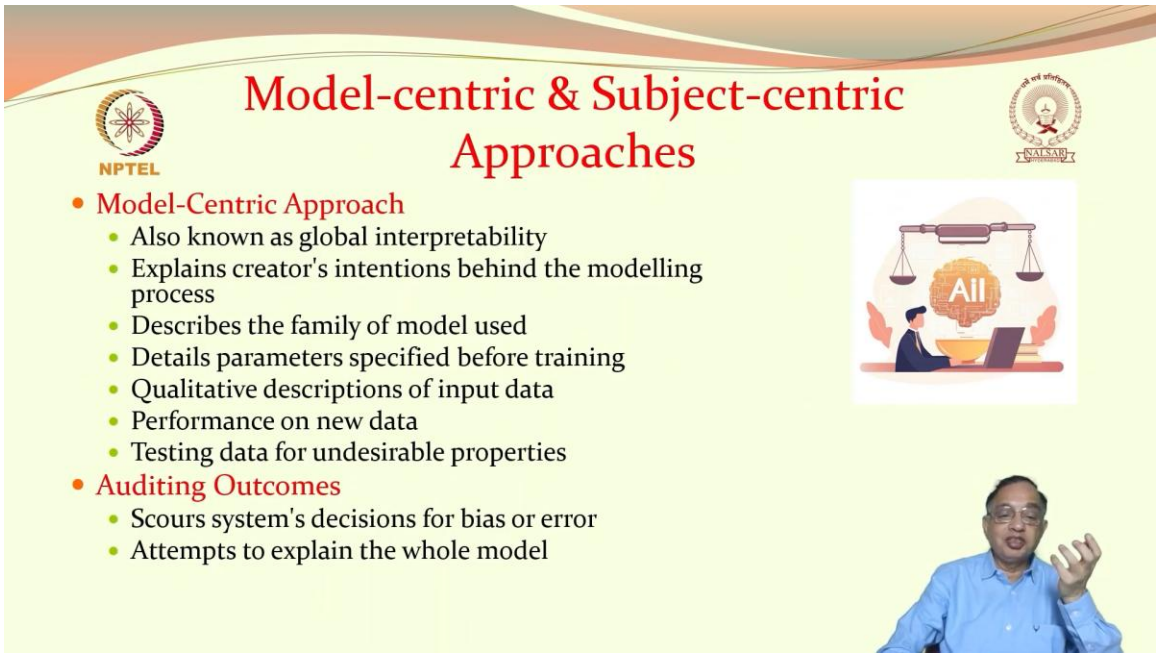
But in some other legal contexts, like the deployment of autonomous weapons, which is again a huge controversial area right now, there is no global convention; however, informal talks, or what you would call not exactly among diplomats, track two diplomacy efforts are going on among various stakeholders in that. More importantly, explainable AI will be required in a field called health, where doctors can be sued. We will see that in another session, or doctors may have doubts about the diagnostics, the classification, or the treatment prescribed. So, the legal consequences for them are one part, but even for their own better understanding and for their own professional requirements, including the ethical requirement, doctors should also see that Explainable AI systems are available to them. Then we have a huge upcoming and developing literature on algorithmic decision-making, which decisions really need explanation, which form of explanation they should take, and how algorithmic decision-making and Explainable AI should be co-developed.



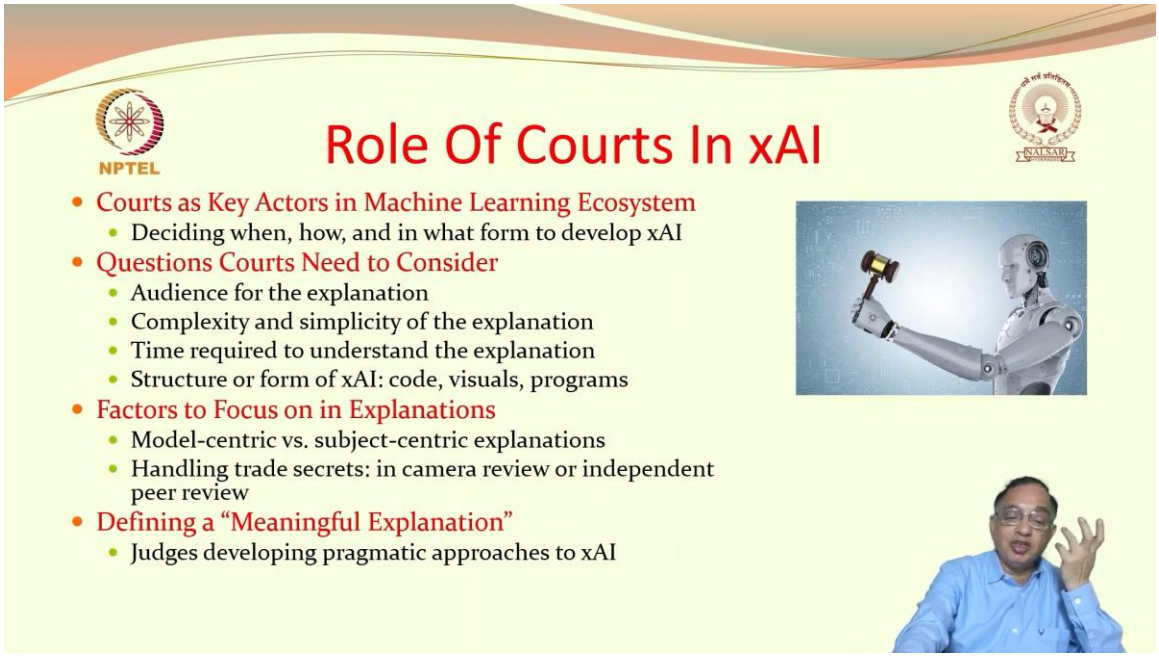
So, as we said, judges have a lot of roles in that.



And more importantly, judges should also look into what sort of Explainable AI models are preferable for the work they do. So, there are multiple forms of currently available Explainable AI models and they are being increasingly developed. So, some models that are intrinsically inbuilt with the capacity to be Explainable may be used, but some studies show that they are less effective in the sense that they are not technically very efficient. So, the intrinsic Explainable AI models, although they seem to be prima facie preferable, should not be at the cost of technical efficiency.

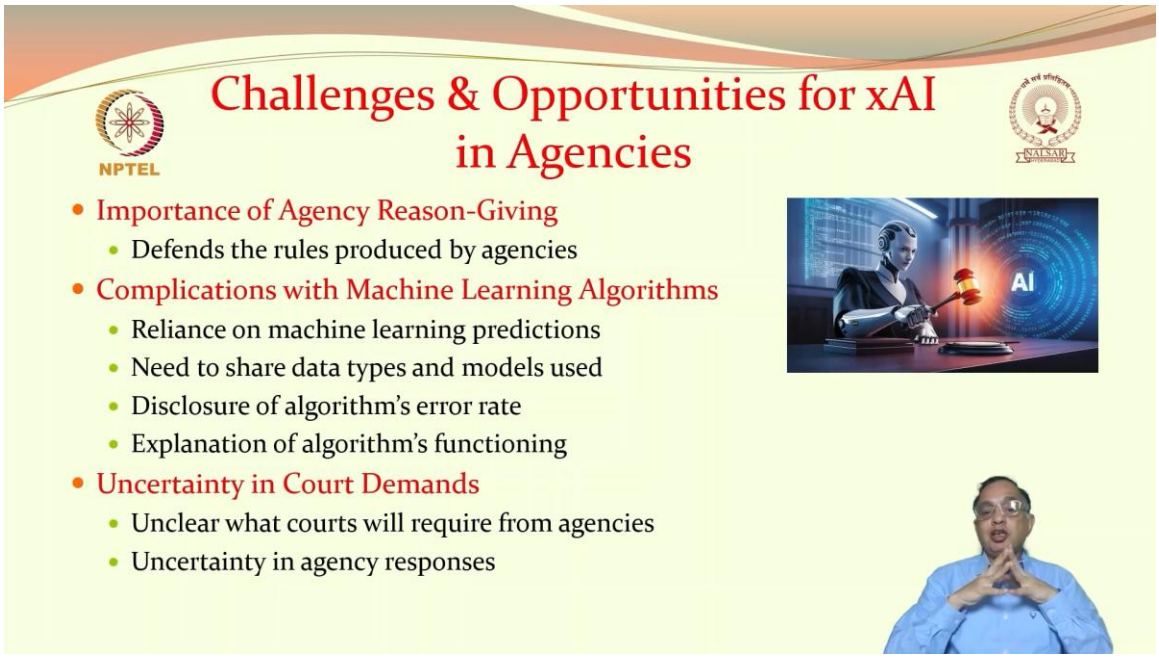


So, we discussed global interpretability and then what exactly the intention of the developer was when he or she developed it, as well as the parameters, the qualitative input data that was fed into it, and how the system deals with new data that it has to process or has used for learning and decision-making purposes. These are some of the things we need to consider when we build and examine Explainable AI itself being developed. So, auditing outcomes could be one way, and then attempting to explain the whole system or whole model in terms of an Explainable AI type of explanation is another way to look at it.

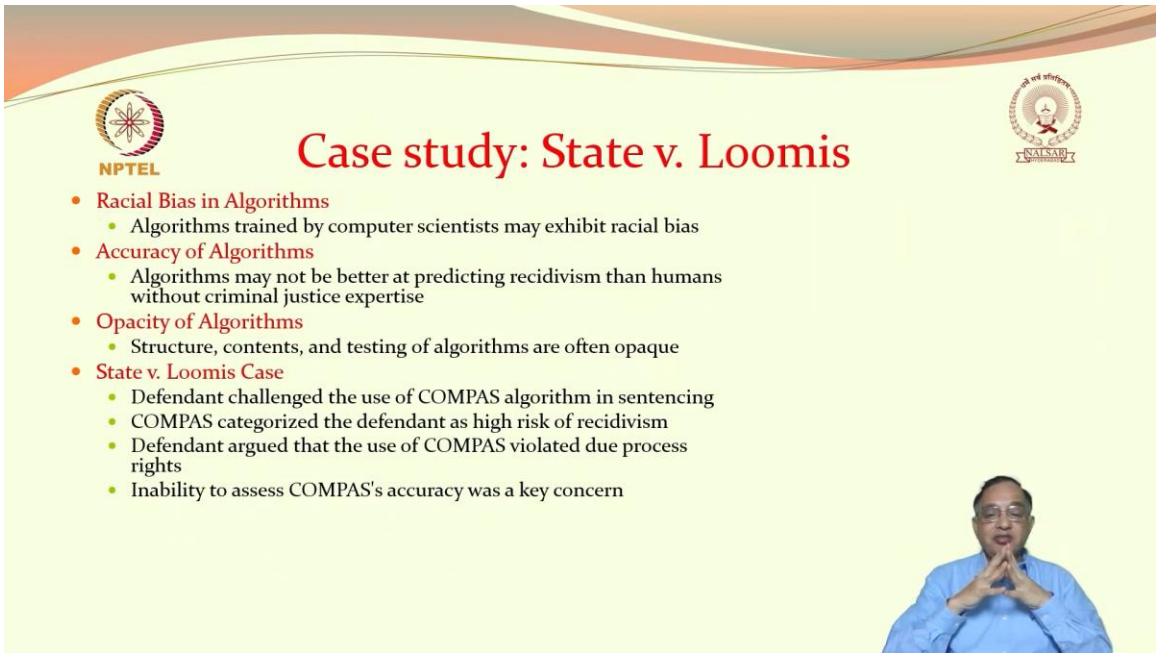


So, as we said, courts have a huge role, and when judges ask for meaningful explanations, they should also know what sort of explanations will be available and what sort of things they should primarily focus on. Should they focus on the core explanation

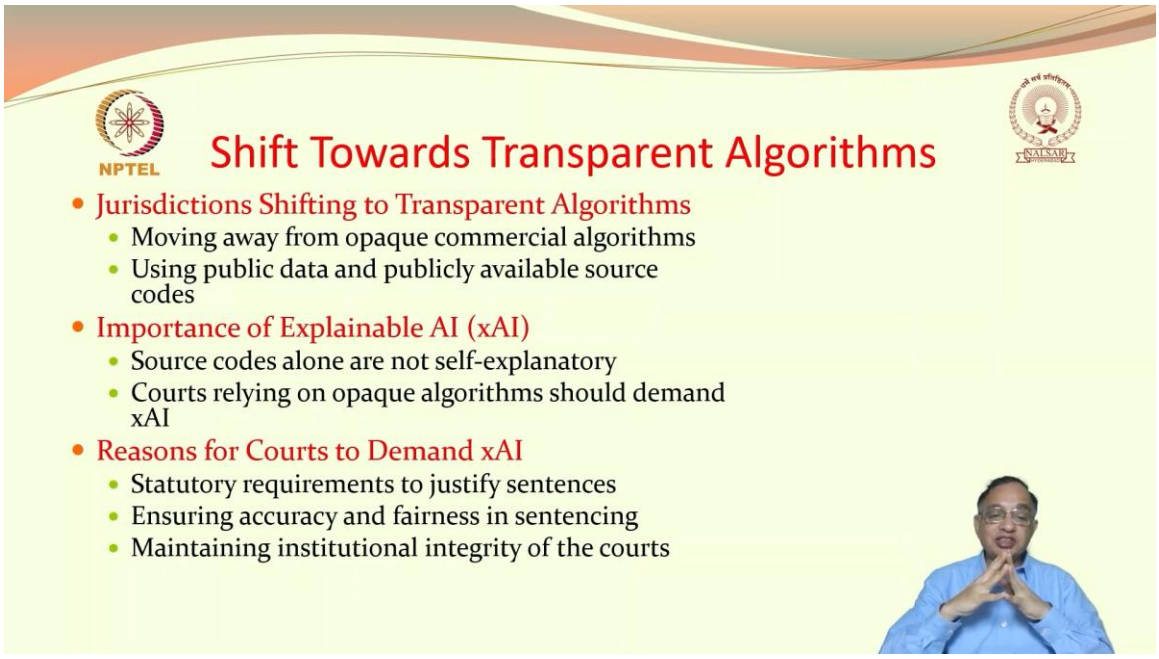
that really matters, or should they go for a detailed explanation that describes each and every step?



So, there are a whole lot of challenges and opportunities for Explainable AI in government agencies and departments, including courts. But there is a lack of what one would call clarity and confidence when it comes to Explainable AI for three reasons. First, the subject itself is very new in the sense that it has not been widely deployed across AI applications in all sectors. In law and justice, the application of AI models is relatively less compared to some other sectors. So, what exactly Explainable AI means here is something that not many people understand, or even if they understand, they are not very clear about it. Then the last and most important point is that there are a lot of things that are vague or unclear at the moment when it comes to applying AI in law and justice because the AI governance or the AI regulation itself is not fully in place in many countries. So, many countries are adopting a piecemeal approach, a sectoral approach, or an innovation-focused approach toward AI regulation. So, when that happens, it is possible that those in judicial decision-making or those who are in the law and justice system may think that the regulation itself is not geared toward that; so when the regulation moves in that direction, that could be the right time to go for Explainable AI, because right now our concern is not just on Explainable AI, but we have other broader concerns, including whether the systems are really compatible with AI and the rule of law, and then what sort of discrimination or bias is actually happening when the systems are deployed. So Explainable AI is part of a solution, but then there are a lot of other things that need to be addressed before we come to this part of the solution.

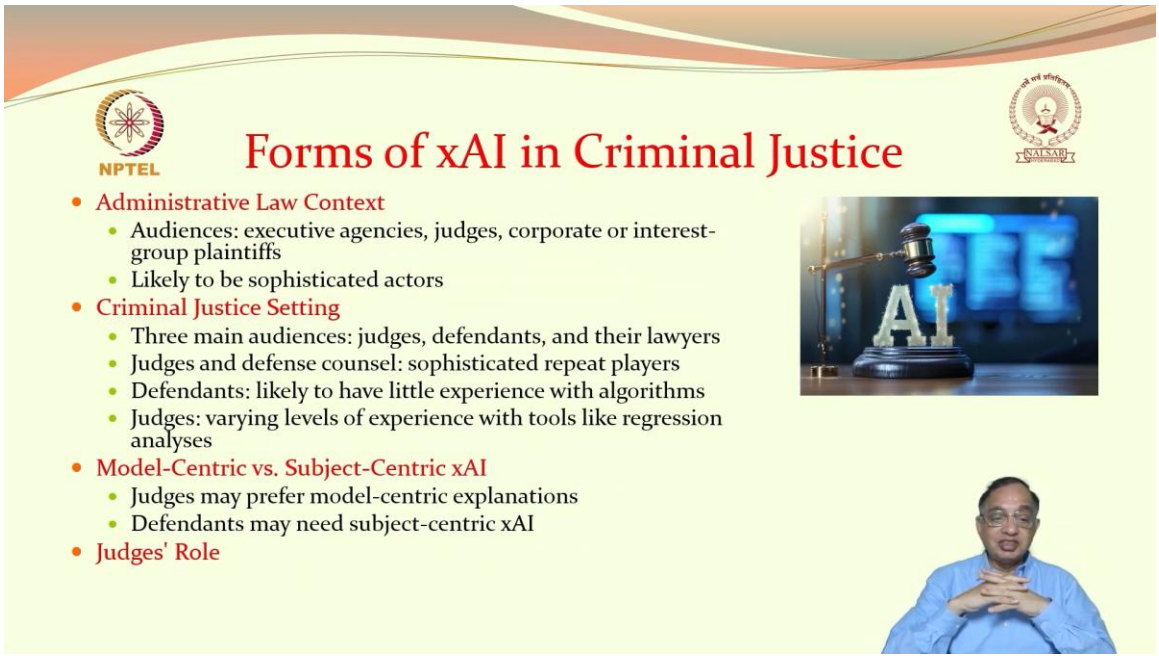


So, this is the case State v. Loomis where the algorithm COMPAS, trained by computer scientists, was biased. COMPAS is a system used in algorithms. It categorized the defendant as a high risk for recidivism. Recidivism is the potential to indulge in or commit a crime again. So, what happened here was that he was not given bail. The reason given was that he said he was protecting against recidivism, so he was not able to grant bail. He first challenged that the use of COMPAS itself violated due process rights, but this was not accepted by the court. But then he could not access the accuracy of COMPAS, which was a key concern for him. So ultimately what happened in this case while his concern was not addressed fully it exposed thoroughly that the system was biased.

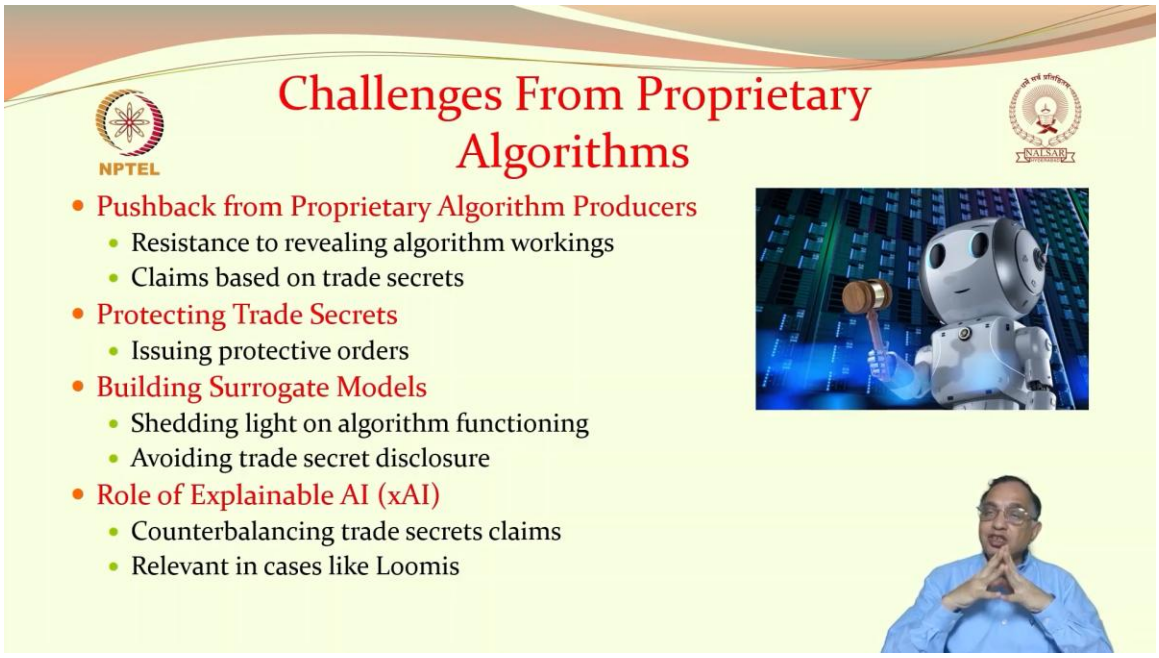


So, the need to move toward transparent algorithms as part of Explainable AI is very

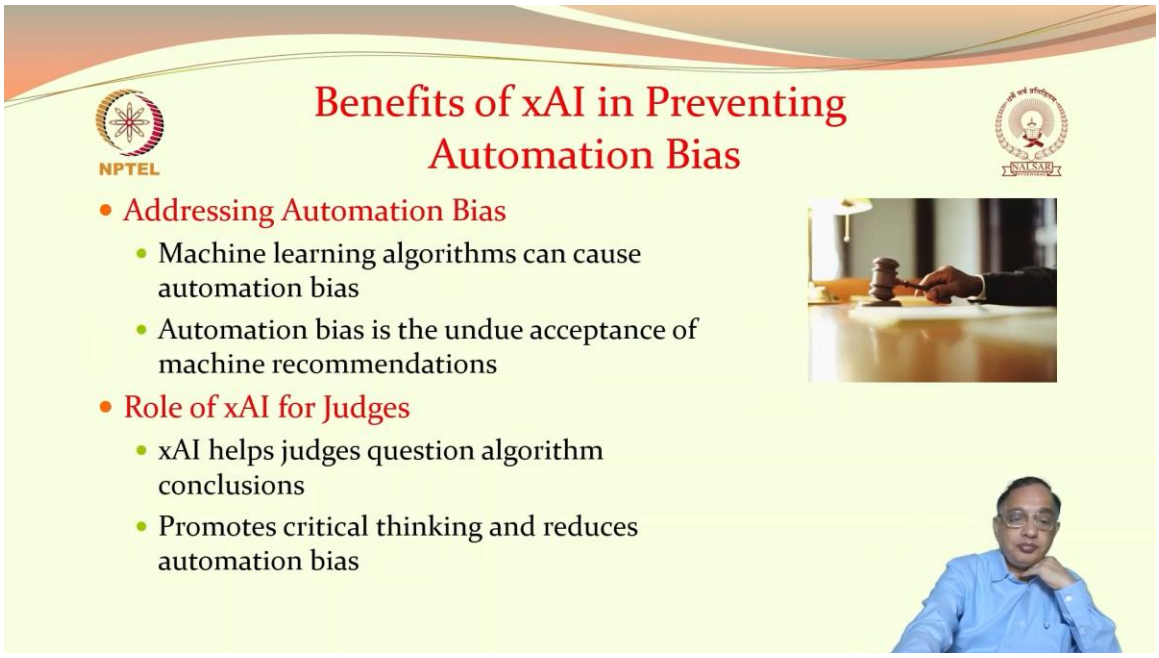
obvious here. And then courts should demand Explainable AI, particularly in critical applications like criminal sentencing.



So, there could be different ways of doing explainable AI in criminal justice. In the sense that in administrative law, it is the executive agencies, judges, corporations, and stakeholders. These actors know well, in the sense that they are well equipped, powerful enough, and resourceful enough. So, they may not hold a lot of capacity building or hand-holding in it, whereas in criminal justice, often criminal justice interferes or has an interface with the downtrodden and the poor, who often end up as victims or who often get unduly punished disproportionately to the crimes they committed. So here, the stakeholders on the other side, or the persons who are facing the judiciary and the law, should be empowered through Explainable AI models, and they should be the ones who can really challenge. Meanwhile, the judges should first understand the rationale behind the decision made by the system before trying to justify it. It is then the duty of the state to ensure that Explainable AI becomes part of the criminal justice system when AI systems are deployed. So again, we need to look into very specific examples where, before being deployed, judges have an understanding, have done some field testing, or have done some random testing with that and are convinced that the system is able to give reasonable judgments, or at least the decision-making is something which they see the potential to go wrong as much less. So, using Explainable AI in criminal justice is not only necessary but should also be the essential component of using AI in criminal justice.

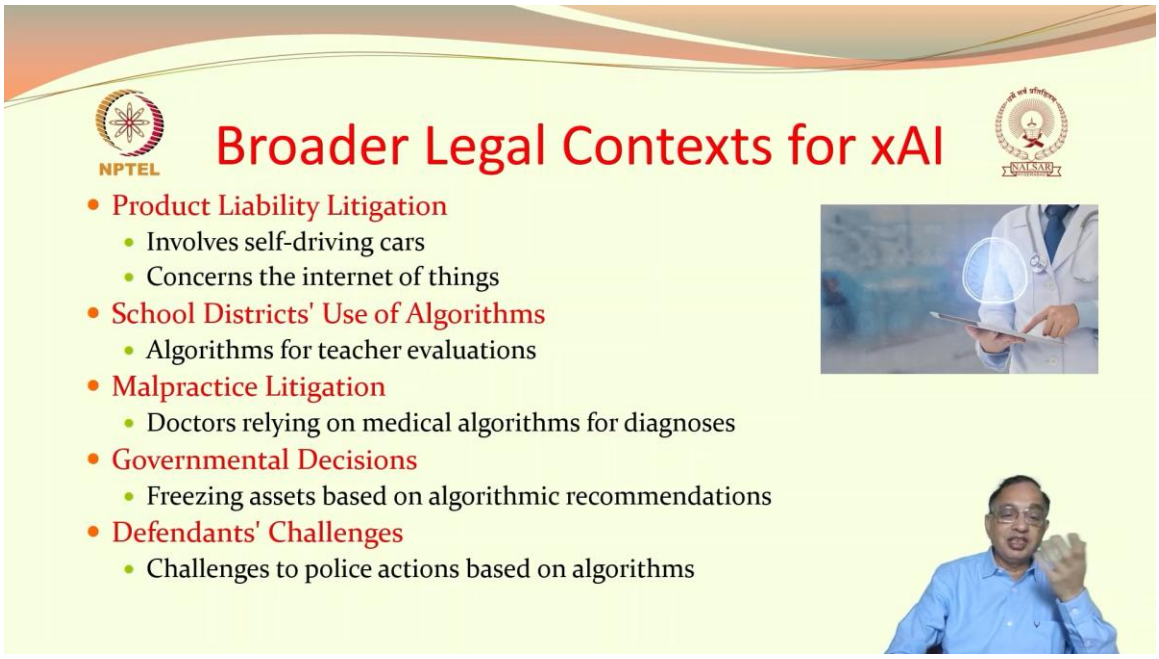


But often the systems are black boxes in one sense; in another sense, they are covered by IP rights and proprietary knowledge, making it very difficult to understand how the system works. So, the companies or the developers will not be willing to disclose the algorithms that are used. And then they can also use the trade secret as one of the IP rights. And then they may build, they may be asked to build surrogate models like synthetic data; we can have surrogate models that are not identical but are very similar or have a similar reasoning algorithm built in, so that this trade secret issue can be partially overcome. Then we can say, "Fine, we will use your surrogate model, run it on this data, run it on this algorithm, and then see you explain to us how this decision-making was done." So, this is a partial solution to address some of the issues related to trade secrets over algorithms and algorithmic decision-making. In a case like Loomis, had such a thing been done, it would be easier for Loomis to decipher, contest, and challenge rather than for the system to become too opaque, or when using AI itself as part of the solution was not considered a violation of due process. So, we need to look at this from the perspective of stakeholders who are on the receiving end; we need to see how algorithmic decision-making can be challenged through Explainable AI, even if there are certain constraints like proprietary algorithms, trade secrets, or any other IP rights that constrain building such Explainable AI systems.



So, this is all we know, but courts are not often very demanding or very forthright in asking for Explainable AI because, as we said, it is totally new, and the trade secret hurdle makes the algorithms very difficult to understand. Courts may often not have full confidence in using AI itself, in the sense that even if an AI system is used, they may not be the ones who have 100% confidence. Therefore, they may not think in terms of asking for Explainable AI because they know that if they go and ask for Explainable AI, a solution will come, but it would also imply that they are fully convinced about the use of AI in court law and justice, particularly in court decision-making. So, making Explainable AI part of the real-world court systems, refining it for better use, and incorporating more and more Explainable AI into the criminal justice and administrative law create a good scope for everyone, including lawyers, academics, developers, software

developers, and people who develop specific systems for legal applications to learn more, understand more, and increase the transparency of the AI systems.



So, there are a lot of other broader legal contexts like product liability, malpractice in medicine, and government decisions. Sometimes the government freezes an account or says that on account of this violation, we are taking over property or your access to your property is limited. So, if that decision is being made on account of an algorithm, Explainable AI should be part of that. So basically, Explainable AI should be available to the defendants as part of their legal rights, but translating this into practice is very difficult because right now defendants don't have that right; they have the right to appeal, they have the right to question, but only when algorithmic decision-making is very widely used. Then, bringing algorithmic decision-making as part of the solution and Explainable AI as one of the tools available to defendants will make sense. On the other hand, when the defendants do not even know that it was the algorithms that had a major role or that it was the algorithms that ultimately decided, they will have no recourse even to ask for Explainable AI. So, prior to thinking in terms of Explainable AI in court and justice, we should also look at the transparency level of judicial decision-making when it comes to the use of algorithmic decision-making in judicial decision-making.

Role of Law Makers in xAI Legislation

- Role in xAI Regulation
  - May demand and shape xAI use across industries and within government
  - Could require xAI in briefings by executive agencies, including intelligence community
- Challenges in Crafting xAI Legislation
  - Statutes must be general to capture basic values
  - Legislation may struggle to keep up with rapid changes in xAI
- Promise of xAI
  - Courts can address xAI issues at the edges
  - Can draw on xAI developments in different legal areas
  - Creators and users of algorithms may respond to court actions

Next

- AI and Labour Law

So, we need to go beyond this, and then in the next class, we will look at AI and labour law. Thank you.