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 – 02
Lecture – 08

Algorithmic, AI, Decision Making - In this session, we will be discussing Algorithms, Artificial Intelligence, and Decision making, which is a key theme in any discussion on the use of AI in law. So, this is session 8.

And let us do a recap of what we discussed before. In the last session, we discussed AI being applied in the legal sector in India and the role of legal tech in furthering innovation. We also highlighted the emergence of new options and challenges, such as AI-based legal agents, and discussed how the legal tech situation in India is evolving, as well as the potential opportunities and challenges for the legal sector in dealing with AI, particularly the emerging AI agents. In this session, we will revisit some of the topics that we discussed earlier and also look at what we need to have a better understanding of algorithms and then algorithmic decision-making.

Some of the definitions are worth revisiting because we need to understand them in a much better way, and we will be discussing algorithms in the subsequent session as well. So, an algorithm is nothing but a series of instructions for performing calculations or tasks. In the context of AI, this helps the computer learn and perform tasks. And algorithmic decision-making is corollary to the way we understand algorithms. Algorithmic decision-making, or ADM, refers to using outputs produced by algorithms to make decisions, and when we say that algorithms can also make decisions, we mean that we can make decisions based on the outputs given by algorithms.

In the historical context, algorithms have a long history and have been applied differently over time. The innovation in information technology in the late 20th century refined the basic ideas, and in the 21st century, we are seeing them go further ahead.

History

- "Algorithms have been around since the beginning of time and existed well before a special word had been coined to describe them. Algorithms are simply a set of step by step instructions, to be carried out quite mechanically,so as to achieve some desired result [...]. The Babylonians used them for deciding points of law, Latin teachers used them to get the grammar right, and they have been used in all cultures for predicting the future, for deciding medical treatment, or for preparing food. Everybody today uses algorithms of one sort or another, often unconsciously, when following a recipe, using a knitting pattern or operating household gadgets"
- Matteo Pasquinelli From Algorism to Algorithm

So let us take a quick look at the history of algorithms. Algorithms have been around since the beginning of time and existed well before a special word was coined to describe them. Algorithms are simply a set of five-step instructions to be carried out mechanically to achieve some desired results. The Babylonians used them to decide points of law. Latin teachers use them to get the grammar correct. And they have been used in all cultures for predicting the future, deciding medical treatment, or preparing food. Today, everybody uses algorithms of one sort or another, often unconsciously when following a recipe, using a knitting pattern, or operating household gadgets. In other words, algorithms are not very new; they are very old. But the way they are being used now is quite new and novel, and that is also causing some concerns. So, this quotation is from Matteo Pasquinelli's Algorism to Algorithm. The link to the paper will be provided later.

So, algorithms take a set of inputs such as age, residence, marital status, and income, process them through a set of steps, and produce outputs or decisions. We have discussed this before, and they are used in various sectors. That is very obvious.

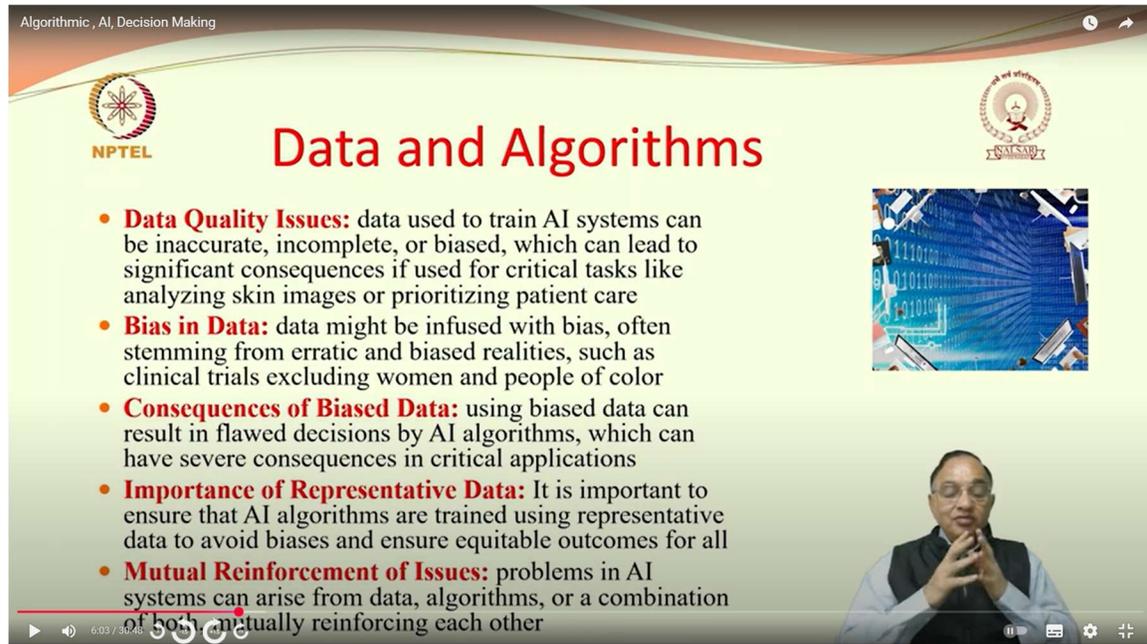Algorithms can learn from training data. AI algorithms are designed to learn from training data, which can be either labelled or unlabelled. This information is used to enhance algorithm capabilities and perform tasks. And then, continuous learning is a plus point for algorithms, particularly in AI, because they can continuously learn and redefine their process by incorporating new data and new evidence, while some of them will need

a programmer's intervention to optimize their performance. And then the task execution - The algorithms use the information from the training data to carry out their tasks effectively. So, task execution can also be done effectively by AI algorithms.

We have seen some of these things before, but it is worth repeating. For example, the AI has quality data issues. Data used to train AI systems can be inaccurate, incomplete, and biased, which can lead to significant consequences when used in critical tasks like analysing skin damage or prioritizing patient care. Again, data might be infused with bias stemming from an erratic reality, such as clinical trials not including women and people of colour, which we have seen earlier as well. Then comes the question of what happens if there are consequences of biased data. Seeing the biased data can result in flawed decisions by AI algorithms. Of course, it can also result in flawed decisions by humans when they are presented with these decisions, or when data is analysed by AI algorithms. And this can have severe consequences in critical applications.

But then, these critical applications can be in any sector. It could be in education, it could be in finance, it could be in housing, it could be in health, and it could also be in jobs and other places. So, the consequence of the biased data due to algorithmic decision-making is something which we need to be critically aware of. Then comes the question of representative data. It is important that the algorithms are trained on data that is sufficiently representative so that the biases and equitable outcomes are addressed.

Biases have to be avoided at any cost, and equitable outcomes for all should be ensured. So, when we talk about representative data that is biased, we cannot ensure that equitable outcomes will be possible from it. On the other hand, some of these issues, like biases and how to ensure equitable outcomes for all, cannot be ensured; it is a question of not

just law and technology but also the way society organises data. Mutual reinforcement of issues - problems in AI systems can have many sources, including data, algorithms, or a combination of both, and they could be mutually reinforcing in the sense that the problem in data can become acute when algorithms take into account the problems of bias and discriminatory data. And then when algorithms make decisions based on that, it could become worse. And then algorithms, again, when they get trained on the same biased and discriminatory data, become mutually reinforcing. So, the data problems and the algorithmic problems can act together or independently, resulting in outcomes that are not good.

Data and representation can occur due to other factors as well. For example, the digital divide in many Global South countries has resulted in data invisibility among historically marginalised people, such as women, caste and tribal communities, religious and linguistic minorities, and migrant labourers. Why? Because these sections of the population, when they are not adequately represented in the data, become invisible to the data. Then, when this invisibility is ignored or the data is used to make decisions or to help enable decision-making through further analysis, they are either excluded or, if not excluded, what they receive is not what they are actually entitled to. So, the data should be representational to the extent that it actually represents or reflects the population, including all combinations of women, caste, tribal communities, religious and linguistic minorities, and migrant labour. Of course, there could be other ways of looking at it, also because even if they do not come under these categories, poor people could get excluded, elderly people could get excluded, and children could get excluded. So, when we talk about that, we also talk about the possibility of so many identities getting excluded for

one reason or another. Biases in AI algorithms may arise due to invisible data or the algorithms themselves being problematic.

So, we need algorithmic transparency and accountability, which itself has to be discussed in a separate session, which we will do. Algorithmic transparency audits are done when we need to really look at what the AI systems are doing. So, AI systems can be audited based on various factors, including transparency in the sense that whether the algorithmic decision-making is sensitive to the issues in the data or the algorithms themselves are biased due to this lack of transparency. And then there are ways of achieving algorithmic transparency in the sense that the process will entail going through the step-by-step functioning of the algorithm, whether the data or the algorithms exclude, assign priorities, assign weightages, assign equally, or decide without taking into account the biases or problems in the data. Or to put it in other words, algorithmic transparency audits should ensure that algorithms are transparent; they are not just black boxes, and the way they arrived at decisions is well known or at least explainable.

But algorithmic transparency cannot be merely dismissed as a technical issue because it could also reflect the socio-technical systems that resulted in this. So, to look at algorithmic transparency as a socio-technical issue means that the interface between society and technology is uneven or is not in favour of some sections of society itself, and this could be addressed not only by addressing the technical issues per se but also by looking at how society organises technology, how society organises the data collection and analysis perspective, and then how society interacts with these technologies as such. So, when we talk of data and representation, we look at a much broader picture, which means that there could be issues that are technical, socially related, or a combination of both. Coming to that, we need to bear in mind that all these things can have a significant impact on marginalised groups. For example, addressing data invisibility will ensure that algorithms do not perpetuate biases against them, do not discriminate against them, and also ensure that they do not get disenfranchised or denied any understanding of the problems in the data, in the algorithms, or in both.

Direct and Proxy Data

- Proxy data is often used when direct data is unavailable or insufficient. However, using proxy data requires caution as it can introduce unintended biases .
- Examples of proxy data include using location as a proxy for income level or status. This can lead to biased decision-making even if the bias is not directly evident .
- AI systems may make predictions based on proxy data that resemble restricted categories of data, such as race, even if race is not explicitly included as a parameter .
- It is crucial to ensure that proxy data is used exclusively for legitimate purposes to avoid unintended biases and ensure fairness
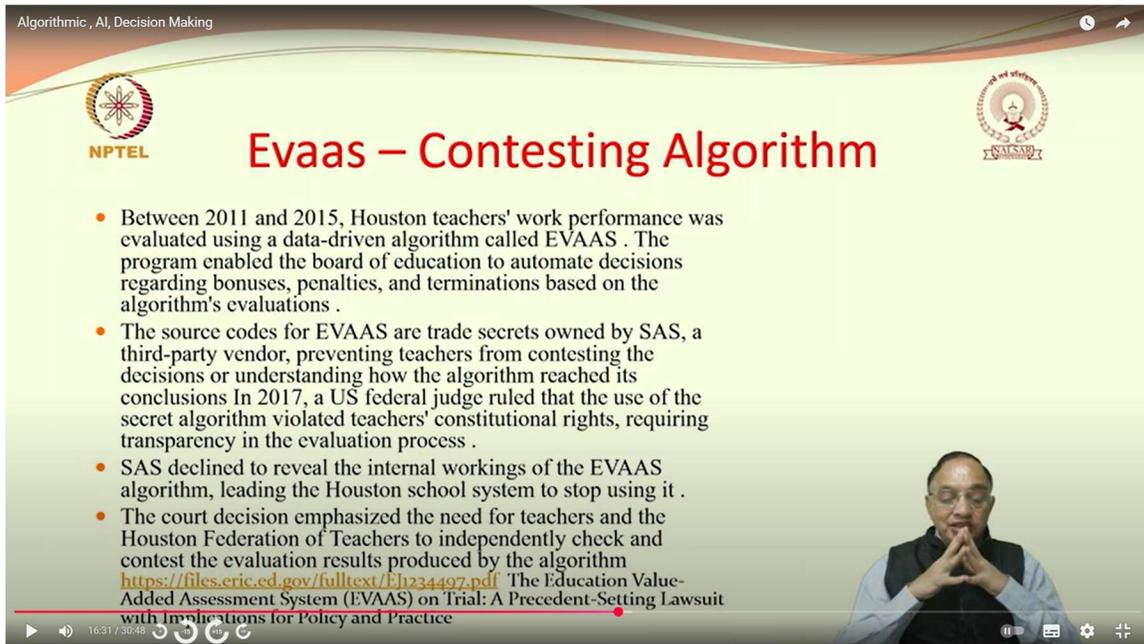
But direct and proxy data have also been taken into account. What is proxy data? Normally, when the data collection is done or when the data set does not fully represent, the practice is to take one set of data or one indicator as an indicator for another. For example, using proxy data requires caution as it can introduce unintended biases. And what is proxy data? Proxy data can be used as an indicator of income level or status, or vice versa. For example, if I take the pin code to decide the status of a person or a family, or to decide the income level, I am using the location as proxy data.

So, when I take that as a proxy, I should be fully aware that I am using the proxy data and not the original or representative data. Now the use of this proxy data can result in biased decision-making or can result in totally wrong data because in a locality, there could be people with different income levels, and people could have different status levels. But if I assume that this location is a place where the majority, if not all, of the people are poor or are not entitled to anything, or have some criminal precedent or are likely to commit some crimes, then I am using location as a proxy in a very wrong sense. But when this happens, the decision can go wrong, the algorithm biases can go wrong, or the algorithms, even if they have no biases, can produce outputs that are very problematic on account of taking the proxy data as a substitute for real data. So, when predictions are based on proxy data, we have huge issues that could pop up even when we are not aware of them.

So, it is necessary that the proxy data be used exclusively for legitimate purposes to avoid unintended biases and ensure fairness. For example, if I use proxy data for only one occasion, I should not try to use the same proxy data as a substitute for something else on another occasion. Or I should not try to use the same proxy data again for the same

purpose; taking location as proxy data for income on one occasion and then using it for the same purpose on another occasion would be to perpetuate the same mistake again and again. So, the direct and proxy data segregation analysis has to be done with a lot of care.
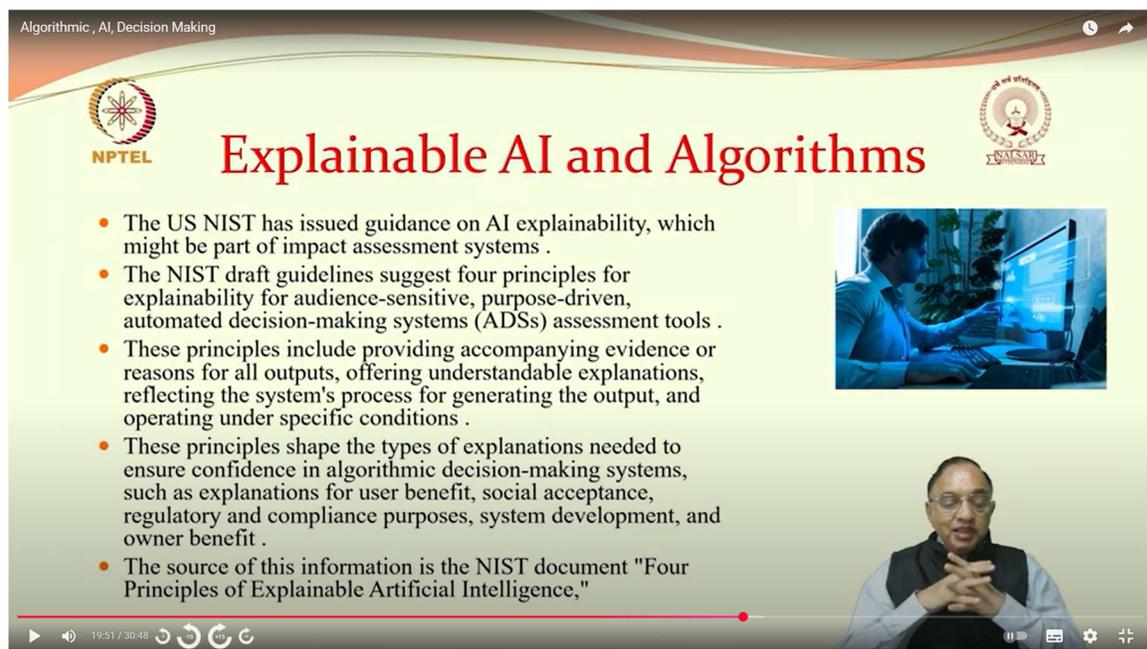

We will give an example where algorithms resulted in wrong decisions, but then we even contested that it could become problematic. Between 2011 and 2015, the work performances of Houston teachers were evaluated using a data-driven algorithm program called EVAAS. The program enabled the Board of Education of that county to automate decisions related to bonuses, penalties, and terminations based on the algorithm's evaluation. Or in other words, the board of education wanted to use the algorithmic evaluation to assess whether the teachers were entitled to a bonus or whether penalties could be levied or whether their services could be terminated. When this happened, this program was owned by a company called SAS.

It's a third-party vendor. So, the company said these are trade secrets owned by them. So, we won't share them. And then this prevented the teachers from contesting the decisions or even understanding how they were assessed and how the algorithms arrived at these decisions. So, the teachers went to court, and in 2017, a federal judge ruled that the use of the algorithmic secret violated teachers' constitutional rights. Remember, we talked about constitutional rights; we discussed the rule of law in an earlier session.

Requiring transparency in the evaluation process was complicated by trade secrets, which made the issue all the more contested. But SAS decided that they would not reveal the internal workings of the EVAAS algorithms as they are covered by a trade secret. And then trade secrets are IP rights on which the owner can say that they will not reveal them unless it is warranted by a court or unless it is made mandatory to ensure that the trade

secret does not result in something that is not a desirable outcome. But the Houston school system decided not to use it instead of contesting the whole idea that the trade secrets could be used to deny accountability and also to deny the fact that they acted in a biased manner. So, what exactly happened is a very peculiar decision.

The court decision emphasised the need for teachers and the Houston Federation of Teachers to independently check and contest. So, the court said you have the right to check and contest, the teachers had the opportunity, but then the company declined to provide transparency or the source code, which they could understand, which was covered by trade secret. As a result, the teachers could not succeed in proving to the satisfaction of others that they were discriminated against and biased against, and therefore the evaluation per se was wrong. But the Houston authorities decided that they would not use it any further. So, this is a case where algorithms have to be contested and have to be proven wrong. Although the people who were affected did not have any other opportunity to prove that they were the ones who did the work, they were discriminated against.

That is why we would need explainable algorithms and explainable AIs. We will deal with this explainable AI much later, but let us keep in mind that this is a key concept. The U.S. National Institute of Standards and Technology issued guidance on AI explainability, which might be part of the impact assessment system. So, if AI explainability is integrated into part of the impact assessment system, which means the impact assessment of the algorithmic decision-making, it could result in some sense of transparency. So, the draft guidelines came up with four principles of explainability. It should be audience-sensitive, purpose-driven, have automated decision-making, and have

automated decision/assessment tools. So, when you are engaged in sensitive, purpose-driven automatic decision-making, the guidelines should be followed.

The principles they came up with are that you should provide accompanying evidence or reasons for all the outputs. The system should be made to explain why this decision was arrived at. And then the explanation should be made understandable, and that should reflect the system's process for generating the output and operating under specific conditions. So, these four principles should be made available to people who are either affected by the algorithms or should be part of the evaluation process for the algorithms. Having said that, we should also understand that when these principles shape the explanation needed to ensure confidence, decision-making systems—such as explanations for user benefit, social acceptance, regulatory and compliance purposes, system development, and owner benefit—are involved.

So, the benchmark set by the USNIST is a benchmark that should be applicable where algorithms are used to enable decision-making or themselves make decisions, and this assessment should be made part of AI explainable tools. Or, in other words, explainable AI should ensure that algorithmic decision-making is assessed and evaluated based on some parameters. Now the source of this document is based on a NIST document, named "Four Principles of Explainable Artificial Intelligence", which can be downloaded from the internet.

So, we are entering a domain called Rights where, contrasting with algorithmic logic, a defendant faces an uphill task here because they need to challenge the logic of an algorithm, the source code, training data, or the required data sets, to which they have no access to. Again, the task is all the more complicated because the defendants who are

impacted by the algorithms may not even be aware of what information they should ask for or what information they should be able to use to challenge the algorithmic decision.

So, what information should be provided to the defendant to contest the logic of an algorithm? Mind you, you are not contesting the logic of a person; you are contesting the logic of a system that is opaque, totally lacks transparency, and is a black box. And these things could affect the rights and claims of the people who are going to be affected by them. So, the defendant should have access not only to the inputs and outputs but also to the underlying logic. Further, what is the margin of error? The margin of error for algorithms, when stated to be 10%, raises the question of whether the 10% margin is acceptable in all cases and in all sectors. A margin of error of 10% can be a matter of life and death when applied in health.

It could again be a question of severe financial consequences if applied in banking or in credit-related decisions. It could again be something that could be very substantial if applied in education. For example, if the margin of error in admitting a student is 10% and the algorithm is biased against some categories of students, then the margin of error can have very severe consequences. So, the margin of error cannot be allowed to be even 1% in some cases. But if the authorities say that the margin of error can be 5% to 10% in some cases, then we will review the decision, if it is proven that it is going to have very severe and serious consequences, for those who have been impacted. So, when we challenge algorithms in terms of rights, we should know where to start and then what sort of data we should contest and what sort of decision we should make.

But the tricky question here is what the due process of law is that the court should enforce in such circumstances. Particularly when algorithms deploy machine learning and

no one, including the developer, understands how the machine learning does analysis, comes to a complete understanding and then comes up with a decision. So, if we talk about due process of law, it demands that there should be transparency, fairness, and, more importantly, that both the defendant and the other party have access not only to some data but also to the whole due process. So, it is much more than data; it is much more than transparency.

The due process examines whether the machine system itself can be considered as due process of law; if not, how do we ensure that the system imbibes the principles of due process of law, particularly when decision-making is undertaken by algorithms or algorithmic AI systems that are enabled by algorithms? To put it in other words, in a technical matter like that, can courts really assess the accuracy of algorithms, particularly if they forecast future human behaviour? This is of critical importance in the criminal justice system. Coming to that, can we say that algorithms should have social and legal responsibilities instead of hiding behind statistically derived impartiality? No. Algorithms should be made accountable and responsible, but who again are algorithms? Are the developers the algorithms, or the ones who deployed them, or both, and if not both, who else? So, when it comes to matters that impact rights, a lot of things need to be paid attention to. And finally, who is really liable if the algorithmic decision-making is wrong, as in the case of students impacted by EVAAS?

And then comes the question of the civil justice algorithm. In fact, algorithms have been deployed in various areas of the civil justice system, including family courts, housing courts, debt, employment, and consumer litigation. So, in all these consequences, in all these cases, algorithms play a major role in decision making. But when civil courts are

increasingly collecting data about administration, pleadings, litigant behaviour, and decisions, offering opportunities for automatic judicial functions, we need to be all the more careful. In the sense that when fully AI-driven civil courts are going to function, we need to be very sensitive about what data they collect, how they are processed, how they are fed into the decision-making behaviour, and then what sort of analysis is being done. For example, if they are going to assess the litigant's behaviour or if they are going to give opportunities for semi-quasi-judicial functions to be fully automated or not, who else is going to get the opportunities? Who else will get the opportunity to vet them, to understand them, and then to point out defects? So, using the algorithms for criminal justice system has got huge consequences. But then it is equally the same when they are done for civil justice system.

When we use these systems to draft or pre-draft judgment templates for judges so that they can review them and then see whether they will go by them, junk them, or use them in part but not in full, we need to know what logic was used by these systems and what algorithms they used. So, when the sentencing recommendations pertain to bail, sentencing, and financial calculations, all the more care has to be taken because the judge cannot go with the pre-drafted judgments. The judge has to do the homework again. But if the judges begin to rely on algorithmic decision-making, thinking that they will be impartial and that they can be sure algorithms have taken into account all the factors and parameters and assessed them without any bias and discrimination, sometimes this can go wrong. And then who will be responsible for it? So, when we discuss all these things, are we creating a scare? No, we are not creating a scare; rather, all we are talking about is that whether it is civil justice or criminal justice, the basic tenets of the due legal process—the basic tenets of due process—being made available to the people who seek justice cannot be compromised on account of technical matters or problems with algorithms, however much they may appear to be impartial and only data-driven and not driven by any biases. Coming to the point, it is possible that AI tools can significantly reduce the research time in the preparation of decisions, but that itself cannot be an alibi or an excuse to rely extensively on algorithms.

Predictions

1. An AI algorithm developed by researchers from Université Catholique of Leuven, the University of Sheffield, and the University of Pennsylvania is used to predict judicial rulings of the European Court of Human Rights.
2. The algorithm, led by Dr. Nikolaos Aletras, has an accuracy rate of 79.70% in predicting these rulings.
3. AI is not intended to replace judges or lawyers but to assist in identifying patterns in case outcomes.
4. It helps highlight cases that are most likely to be violations of the European Convention on Human Right

As we discussed predictive algorithms earlier, we will give an example here. An AI algorithm was developed by researchers in the University of Catholique of Leuven, the University of Sheffield and the University of Pennsylvania. They wanted to predict the judicial rulings of the European Court of Human Rights.

So, the algorithm devised by Dr. Nikolaos Aletras had an accuracy rate of 79.70% in predicting these rulings in the sense that the rulings were predicted with this much level of accuracy. But AI was not intended to replace judges or lawyers but to assist them in identifying patterns in case outcomes, which is a good thing. It helps to highlight cases where there are most likely violations of the European Convention on Human Rights. So, predictions can easily indicate that the violation of human rights in European countries can be predicted accurately when we analyse these patterns. In these cases, the predictive analysis can flag that these violations are likely to occur. In other words, predictive analytics could flag certain things that we might have otherwise missed.

Algorithm as Authority

- When Algorithms are embedded in decision making they become 'authorities' by de facto and as part of AI black boxes
- But as we will discuss in subsequent sessions use of algorithms can impact human rights, rule of law and adversely affect access to services
- Integrating them into AI systems raises questions about ethics, accountability and responsibility.
- AI tools using algorithms are of concern when they are used in law and justice
- Some of these will be elaborated in the subsequent sessions.
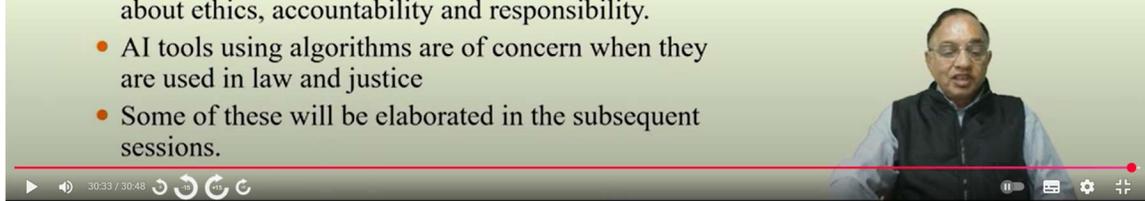
But when algorithms are used as authorities and when they are embedded in decision-making, they become authorities de facto, if not in practice, and then they become part of the AI black boxes. So, algorithms as authority are a serious matter because we may not even be aware of them. Coming to the second and most important point, we will discuss in these subsequent sessions the use of algorithms that can impact human rights, the rule of law, and adversely affect access to services.

So, in the next class, we will discuss some of the key concerns and then see how they could be addressed. Integrating them into AI systems raises questions about ethics, accountability, and responsibility. So algorithmic decision-making and algorithms as authority have serious consequences for the legal system if they are used indiscriminately, carelessly, and without regard for the potential negative impact on human rights. So, we need to be very careful about what sort of AI tools are being used and what sort of algorithms are being used. We will elaborate on them in the subsequent discussion. Thank you.