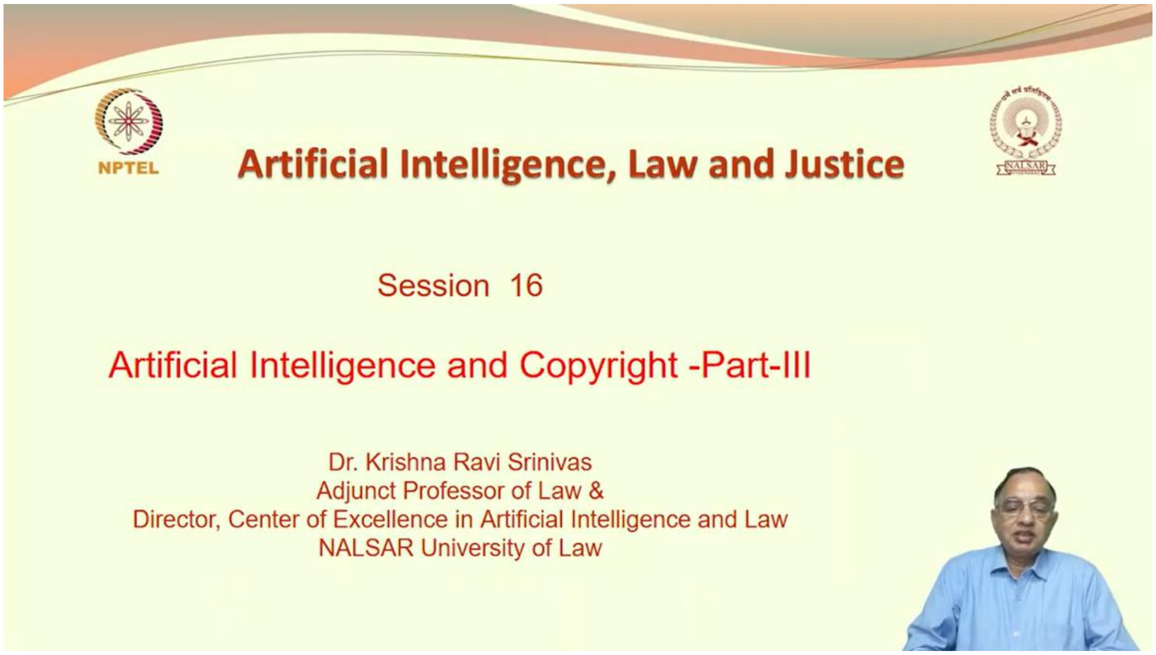
Course Name – Artificial Intelligence, Law and Justice
Professor Name – Dr. Krishna Ravi Srinivas
Department Name – Center of Excellence in Artificial Intelligence and Law
Institute Name – NALSAR University of Law
Week – 04
Lecture – 16



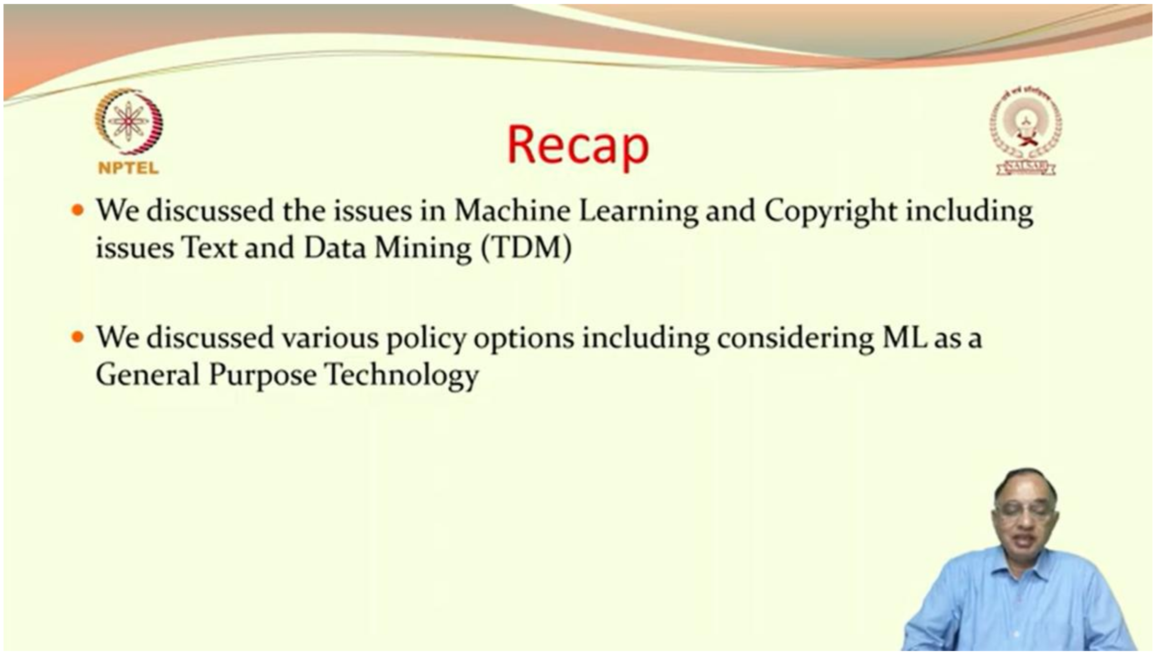
Artificial Intelligence, Law and Justice

Session 16

Artificial Intelligence and Copyright -Part-III

Dr. Krishna Ravi Srinivas
Adjunct Professor of Law &
Director, Center of Excellence in Artificial Intelligence and Law
NALSAR University of Law

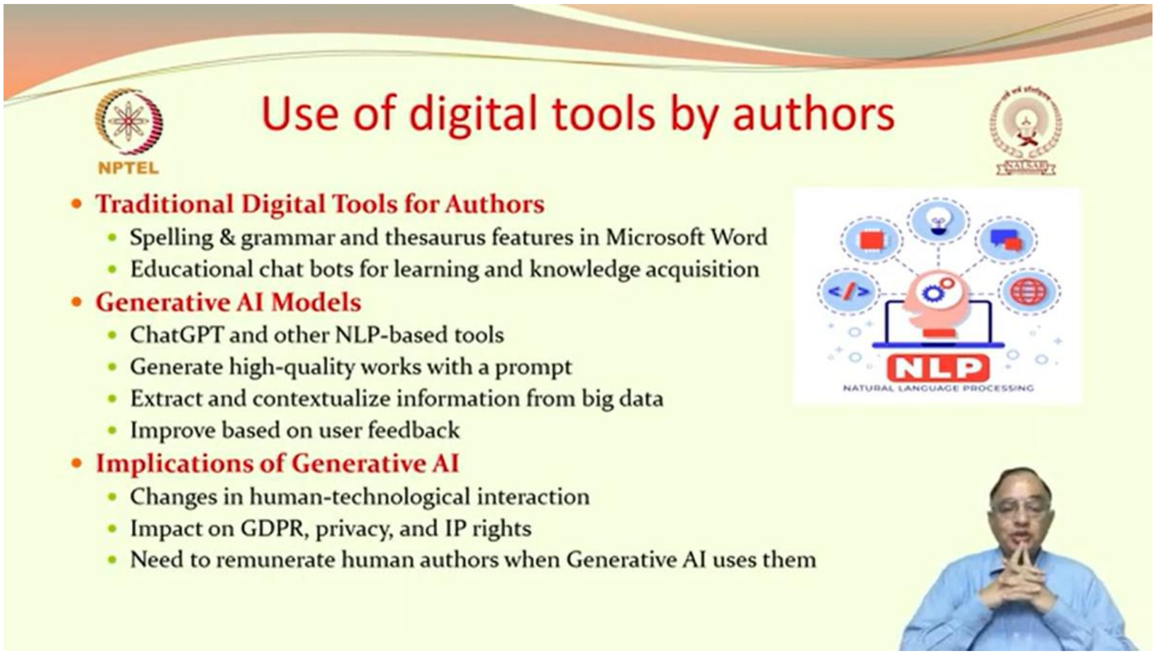
Artificial Intelligence, Law and Justice, Session 16. Part 3 of 4: Generative AI and Copyright Law.



Recap

- We discussed the issues in Machine Learning and Copyright including issues Text and Data Mining (TDM)
- We discussed various policy options including considering ML as a General Purpose Technology

We will do a recap of the earlier session; we discussed issues in machine learning and copyright, including issues in text and data mining. We also discussed various policy options, including ML as a general-purpose technology. In this session, we will look at some things that are very interesting.



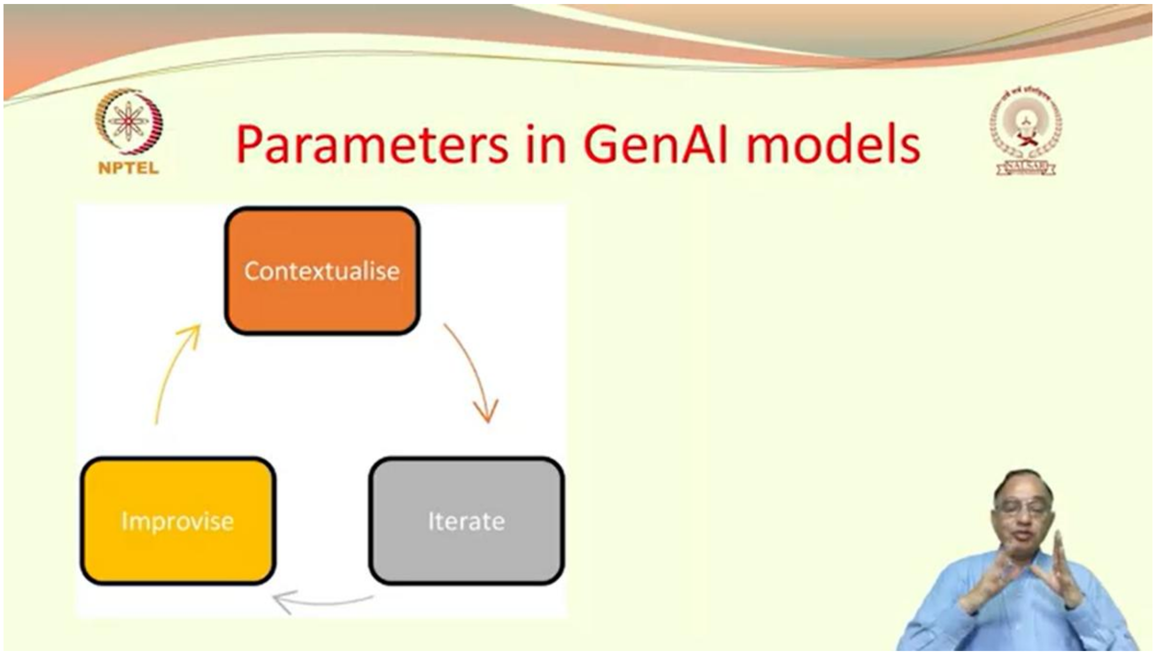
Use of digital tools by authors

- **Traditional Digital Tools for Authors**
  - Spelling & grammar and thesaurus features in Microsoft Word
  - Educational chat bots for learning and knowledge acquisition
- **Generative AI Models**
  - ChatGPT and other NLP-based tools
  - Generate high-quality works with a prompt
  - Extract and contextualize information from big data
  - Improve based on user feedback
- **Implications of Generative AI**
  - Changes in human-technological interaction
  - Impact on GDPR, privacy, and IP rights
  - Need to remunerate human authors when Generative AI uses them

First of all, we should also understand that technology is not something authors have not been using. In fact, authors have been using technology since time immemorial. For example, the spelling and grammar of the thesaurus, which is in Microsoft Word, are something we use traditionally and on a day-to-day basis. And then, this is a traditional digital tool for authors. And there are many educational chatbots for learning and knowledge acquisition. But the generative AI models became a game changer; they resulted in a transformation because they are more versatile than what we have been using as digital tools by creators. ChatGPT and NLP models created a completely new framework for creators as well as for others. Because they create high-quality works with a single prompt, and more importantly, they have the unprecedented capacity not only to create works but also to contextualize them, they then improve based on the user's frameworks. In the sense that if you know how to create good prompts or how to handle prompt engineering, there are immense, endless opportunities for you as a creator to keep creating works. Also, to keep adding text in the sense that it requires creative capacity.

So, Generative AI has, in fact, resulted in a new way of working with computers, a new way of using computers, and it has become such a creative option for creators. So, it has brought in a change in the human technological integration because here users know one thing. That they are dealing with something that is immensely creative and unprecedented in a way, and then the powers and creative capacity are limited more by what you really want to do with it or what you want to derive from it than by the limitations of technology per se; rather, the limitations may come more from your own ability to imagine, adapt, and use it than from anywhere else. So, these are the implications here. But then there are a whole lot of things that we really need to figure out.

For example, what are the implications of the General Data Protection Regulation (GDPR) on data privacy? And then, what are the implications of privacy issues over IP rights? And more importantly, when generative AI uses so much of the material from human authors, how do we remunerate them? How do we honour them by ensuring that they receive the decent remuneration that is due to them? Particularly when the copyrighted works are used extensively by generative AI. So, the use of digital tools by authors is something we need to be very appreciative of and careful about. But what do they derive in turn if they use digital tools that also utilize their own materials?



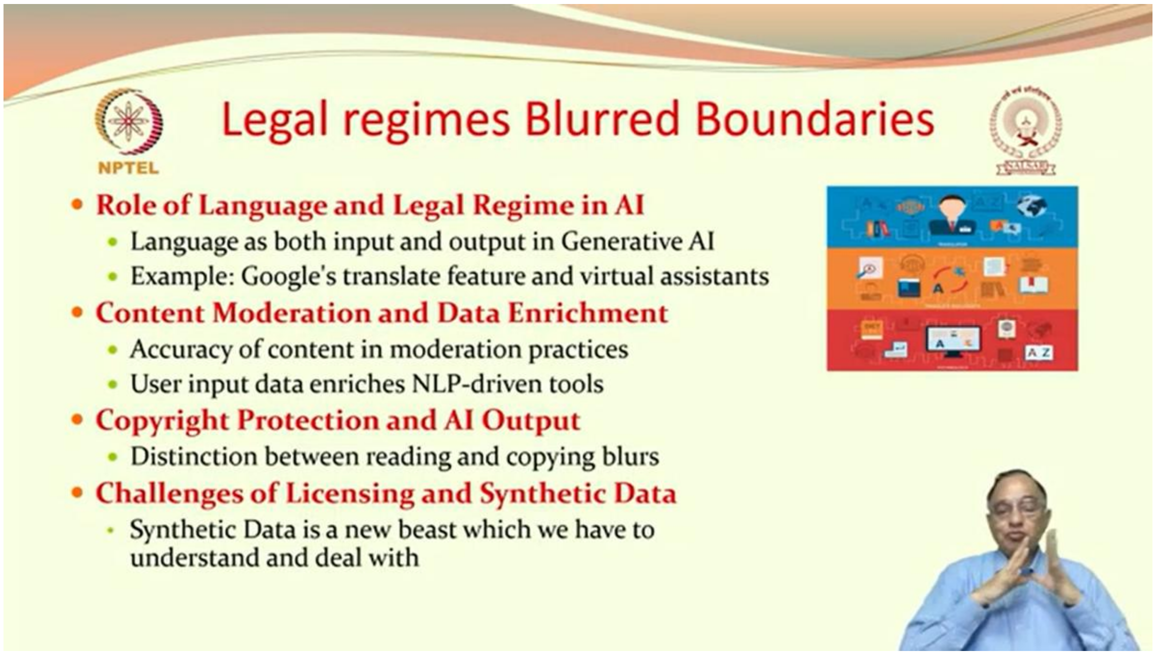
Parameters in GenAI models

Contextualise

Iterate

Improvise

So, the parameters here are that you contextualize, you iterate, you improvise, and then this is a virtual circle that goes on; so, when these things happen in generative AI models, it is like a learning cycle that loops on and on endlessly. So, when this happens, the use of materials also gets enhanced and becomes something like a virtual cycle. Not a vicious cycle.

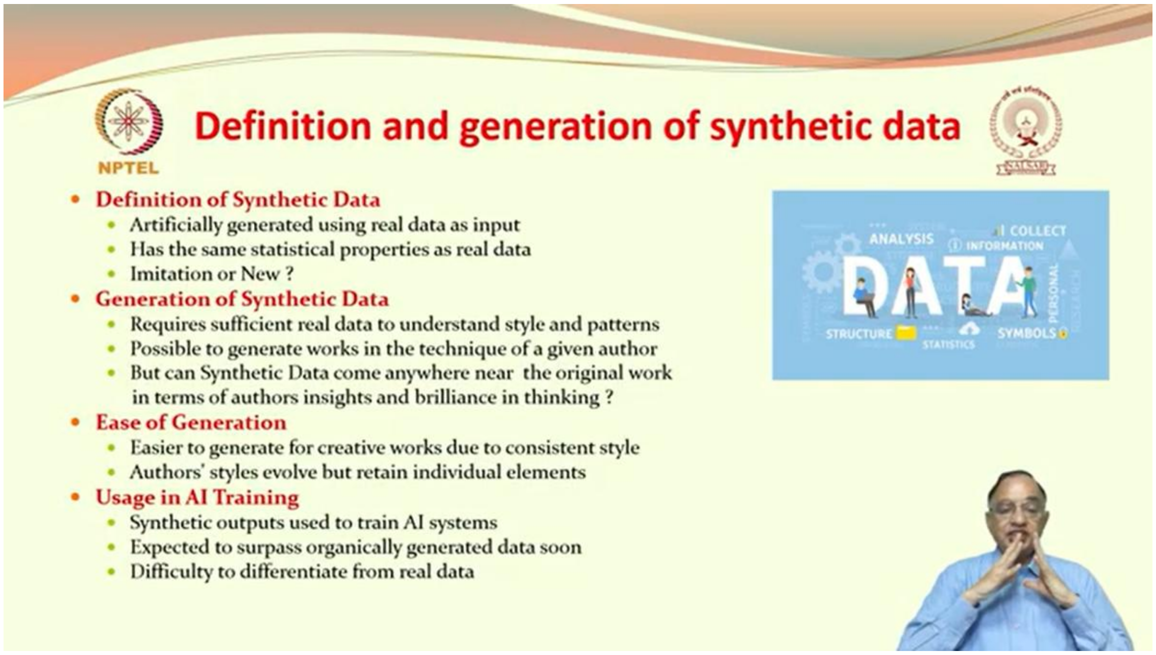


Legal regimes Blurred Boundaries

- **Role of Language and Legal Regime in AI**
  - Language as both input and output in Generative AI
  - Example: Google's translate feature and virtual assistants
- **Content Moderation and Data Enrichment**
  - Accuracy of content in moderation practices
  - User input data enriches NLP-driven tools
- **Copyright Protection and AI Output**
  - Distinction between reading and copying blurs
- **Challenges of Licensing and Synthetic Data**
  - Synthetic Data is a new beast which we have to understand and deal with

But language here is more of an input as well as an output. So, Google's translation feature and virtual assistants use language both as input and output when they translate or when they transform our queries into different formats. But the accuracy of content in the moderation process is equally important because user input data enriches NLP-driven tools. So, content moderation and data enrichment often go hand in hand. But the problem comes in a very unexpectedly interesting manner. There is a distinction between reading and copying. In the sense that when my eyes look at a text, when I read a book, I am merely reading it. I am not copying it. But when an AI system uses it, the very act of reading also takes into account copying, or it becomes copy. Because when I read a book or see a painting, I am only experiencing it.

I may be mentally imbibing it, I may be mentally absorbing it, or it could enter my brain as a visual image. But then I am not copying it in the sense that I am not storing it or using it further in my own mind. But AI systems use it in a specific way, so even when they scan or read, "they copy." The problem here is that today AI models are also using a lot of synthetic data, and then the synthetic data is again generated by AI models. And then the synthetic data could be based on and derived from the available data, which again could be protected by copyright, so this becomes a huge issue because synthetic data is really a new beast that we have to understand and deal with. Although synthetic data could be very much derived from the available data, that data itself would be protected by copyright. But boundaries here—who exactly owns the copyright or who exactly owns the rights to the materials that synthetic data keeps producing and reproducing—derive problems in many other ways as well.

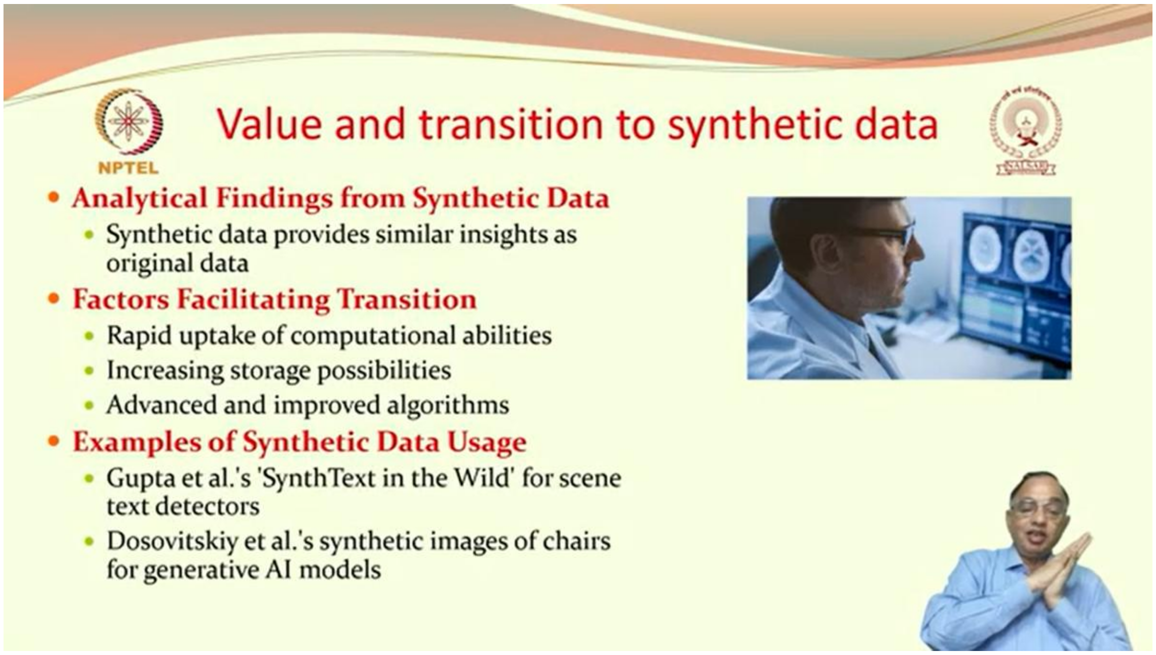


Definition and generation of synthetic data

- **Definition of Synthetic Data**
  - Artificially generated using real data as input
  - Has the same statistical properties as real data
  - Imitation or New ?
- **Generation of Synthetic Data**
  - Requires sufficient real data to understand style and patterns
  - Possible to generate works in the technique of a given author
  - But can Synthetic Data come anywhere near the original work in terms of authors insights and brilliance in thinking ?
- **Ease of Generation**
  - Easier to generate for creative works due to consistent style
  - Authors' styles evolve but retain individual elements
- **Usage in AI Training**
  - Synthetic outputs used to train AI systems
  - Expected to surpass organically generated data soon
  - Difficulty to differentiate from real data

First of all, synthetic data, if we define it as artificially generated data using real data as input, has the same statistical proportions as real data. So, is it really an imitation, or is it something totally new? That's the first question: even the synthetic data, "Is it really synthetic?" If it is not, then it is not something that actually needs sufficient quantities of real data. In the sense that it is not something that is 10% synthetic and 90% original, it is not like that. Now, I can create synthetic data for an author by using the author's works and imitate the style in that synthetic data. I can ensure that the writing style is imitated, including the same phrases, catchphrases, and the phrases for which the author is well known; the writing style and the author's typical phrases can then be part of the synthetic data by understanding, imitating, and reproducing it in the same way. So, whether synthetic data can come anywhere near the original work in terms of the author's insight and brilliance in thinking is a debatable point. For example, if I create synthetic data out of the work of either Dostoevsky, Shakespeare, or Rabindranath Tagore, will that synthetic data come anywhere near the original work in terms of sheer brilliance? I may think it will not, but then when I read it, I may be able to see the traces; I may even be fooled into thinking this was written by Dostoevsky, although it may lack originality and brilliance in terms of content. The real problem arises when synthetic data can be produced left, right, and centre in unimaginable quantities.

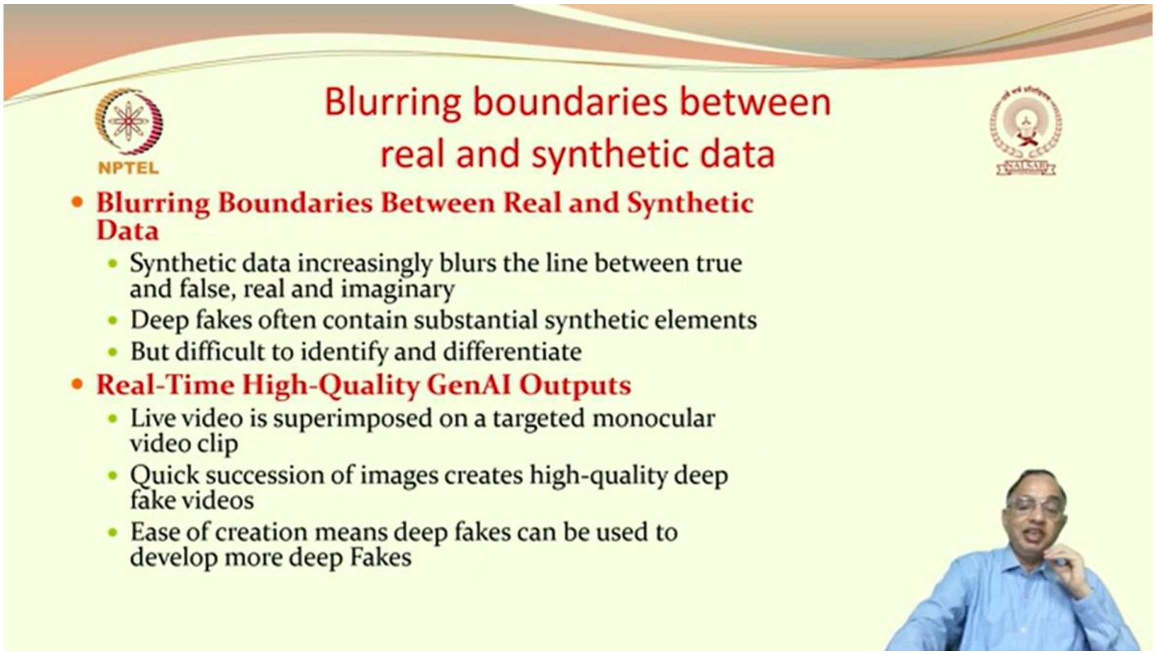
If the author took a decade, months, or weeks to write something, my synthetic data can create quantities unimaginably large. In the sense that my synthetic data can create 4,000-page and 5,000-page novels from the works of an author, Although the authors' works that they themselves put together may not be more than 1,000 pages or 2,000 pages. So, the enormous quantity itself creates a huge problem for copyright in terms of copyright holders' rights regarding the very idea of copyright as an IP right. So, authors' styles may evolve, but they may retain the individual elements. When I use synthetic outputs to train my AI system, they are going to generate surpassing data very soon. As it happens beyond a certain level, it is really difficult to differentiate between real data and synthetic data, as it multiplies and becomes huge in terms of quantity. The situation is like this. If we conceptualize synthetic data as counterfeit notes, when the circulation of counterfeit notes exceeds the notes printed or put into circulation by the authorized authority in India, namely the Reserve Bank of India, it is really difficult to differentiate between them. The amount of synthetic data being too large makes it difficult to differentiate it from the real data.



Value and transition to synthetic data

- **Analytical Findings from Synthetic Data**
  - Synthetic data provides similar insights as original data
- **Factors Facilitating Transition**
  - Rapid uptake of computational abilities
  - Increasing storage possibilities
  - Advanced and improved algorithms
- **Examples of Synthetic Data Usage**
  - Gupta et al.'s 'SynthText in the Wild' for scene text detectors
  - Dosovitskiy et al.'s synthetic images of chairs for generative AI models

Coming to the next point, the analytical findings from the synthetic data often give the same insights as the original data, and synthetic data becomes easy to uptake for computational abilities; storage possibilities make it all the easier to store and analyse, and it can also result in better, more advanced, improved algorithms. There are some uses of synthetic data that we can mention. "SynthText in the Wild" is for scene text detectors. Dostoevsky's synthetic images of chairs are for generative AI models. So, for synthetic data usage, there are many good examples available. The value and transition to synthetic data as it happens also creates huge problems in terms of enforcing copyrights, particularly for the rights holders.



Blurring boundaries between real and synthetic data

- **Blurring Boundaries Between Real and Synthetic Data**
  - Synthetic data increasingly blurs the line between true and false, real and imaginary
  - Deep fakes often contain substantial synthetic elements
  - But difficult to identify and differentiate
- **Real-Time High-Quality GenAI Outputs**
  - Live video is superimposed on a targeted monocular video clip
  - Quick succession of images creates high-quality deep fake videos
  - Ease of creation means deep fakes can be used to develop more deep Fakes

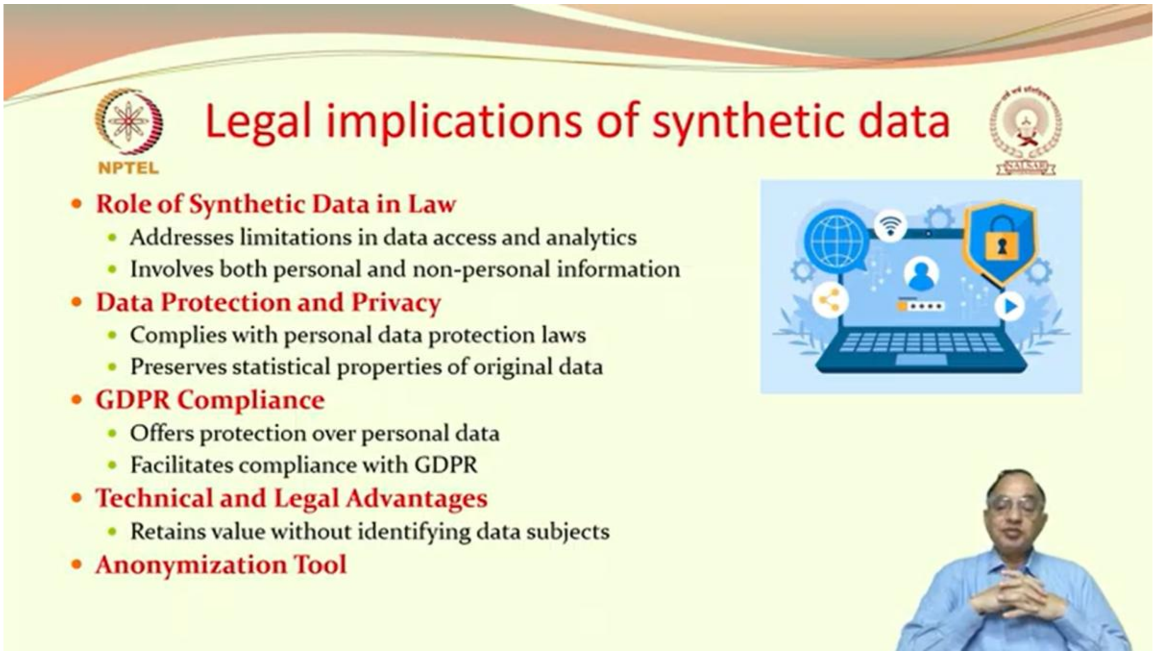
Coming to that, the boundaries between the real and synthetic data are becoming more and more blurred; what is true, what is false, what is real, what is imaginary, and what is something that floats in between is a problem. Deepfakes often contain substantial synthetic elements derived from the original works—synthetic elements that are built upon the original works—and identifying and differentiating them has become a huge issue now. To make things more complicated, today it is feasible to create real-time, very high-quality generative AI outputs, whether they are deepfakes or not; that is a different question. For example, live video can be superimposed on a targeted video clip and then can be generated endlessly, and a quick succession of images captured from different sources fed to a single source and then multiplied becomes a huge problem because they occur in rapid succession, resulting in high-quality deep fake videos that look so authentic that it is simply difficult to say that they are deep fakes. They are so easily manipulable that we tend to consume them; we tend to see them as the original works or the works that are really true. The ease of creation also means that you can create deep fakes from deep fakes. Again, further circulation of deep fakes becomes easily possible, and then the endless cycle of deep fakes creating deep fakes is a nightmare. But then, technically, it is feasible, and it is happening.

NPTEL

Example of synthetic data in AI training

- **Synthetic Data Diversity**
  - Includes datasets, images, and audio-visual works
- **AlphaGo by Google's DeepMind**
  - Trained on real-world inputs like game rules and strategic moves
  - Used repeat simulation to learn and improvise
  - Identified a strategic blind spot overlooked by humans
  - Defeated the best AlphaGo player effortlessly

When synthetic data is used in AI training, it can use datasets, images, and audio-visual works. For example, AlphaGo from Google's DeepMind was used with real-world inputs such as game rules and strategic moves. It was using them repeatedly to simulate, learn, and improvise. But it identified a strategic blind spot that had been overlooked by humans. And that made the huge difference that Google DeepMind's AlphaGo, which again was not used to play the game Go, defeated the best AlphaGo player effortlessly. When it identified a strategic blind spot, it was against the best player. So, the use of synthetic data can have or can result in unanticipated consequences.



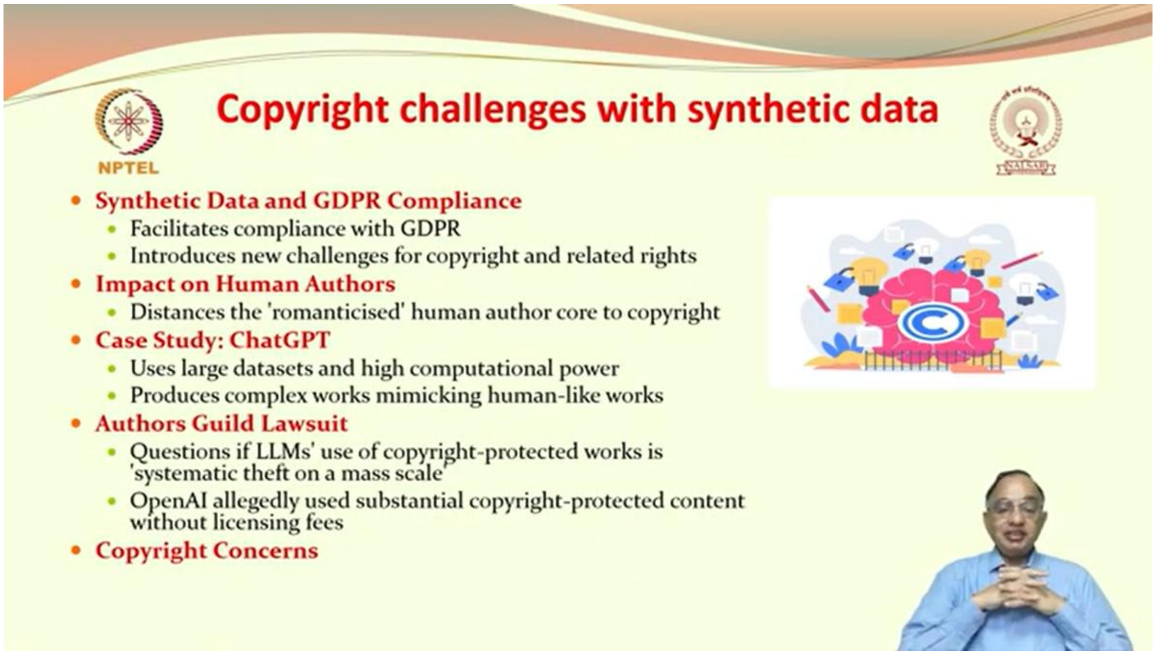
NPTEL

Legal implications of synthetic data

- **Role of Synthetic Data in Law**
  - Addresses limitations in data access and analytics
  - Involves both personal and non-personal information
- **Data Protection and Privacy**
  - Complies with personal data protection laws
  - Preserves statistical properties of original data
- **GDPR Compliance**
  - Offers protection over personal data
  - Facilitates compliance with GDPR
- **Technical and Legal Advantages**
  - Retains value without identifying data subjects
- **Anonymization Tool**

But there are legal implications of synthetic data. First, the role of synthetic data in law is very ambiguous. It addresses the limitations of data access and analytics because it involves both personal and non-personal information. So, where does personal data end?

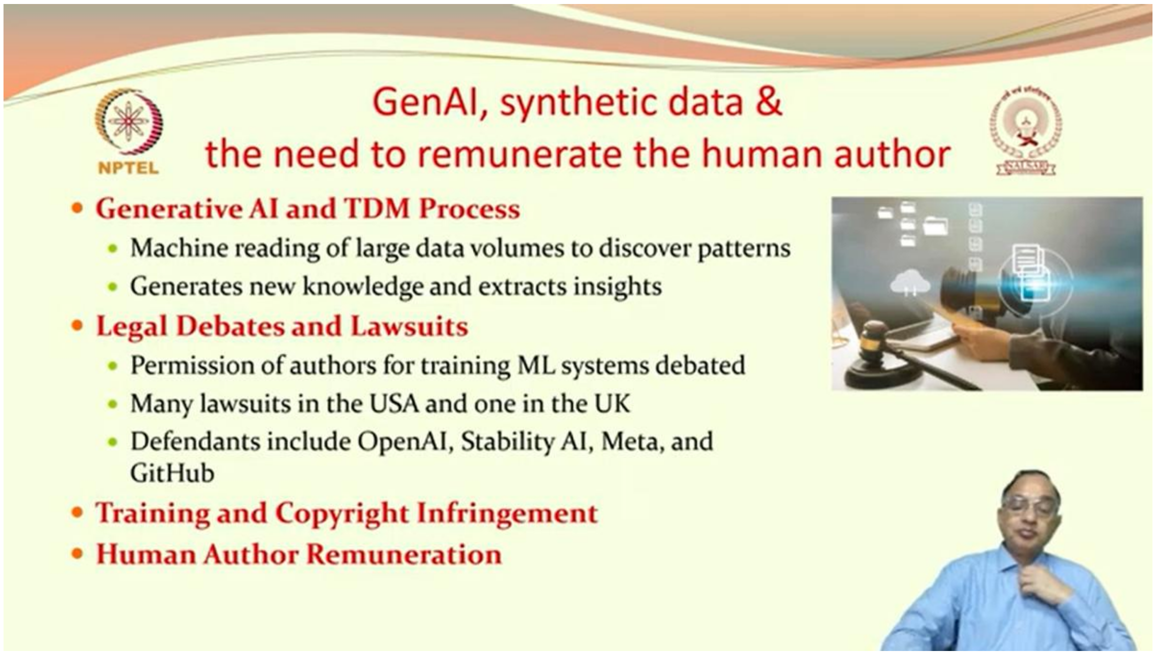
Where does the non-personal data begin and become a problem? But then, data protection and privacy compliance are huge issues. Because the synthetic data retains and preserves the statistical properties of the original data, is it really GDPR compliant? Is protection available for personal data? Again, we are not sure. So, if synthetic data can facilitate compliance with GDPR, then some issues can be addressed and resolved. But the problem here is that it retains value without identifying the data subjects. So, the synthetic data will still be valuable, but the identifiable data subjects may not be known. We are talking about this in the context of the GDPR here. So, it is also possible to perform some anonymization there. So, the legal implications of synthetic data are huge, but which law should be applicable under what context is, again, a grey area.



Copyright challenges with synthetic data

- **Synthetic Data and GDPR Compliance**
  - Facilitates compliance with GDPR
  - Introduces new challenges for copyright and related rights
- **Impact on Human Authors**
  - Distances the 'romanticised' human author core to copyright
- **Case Study: ChatGPT**
  - Uses large datasets and high computational power
  - Produces complex works mimicking human-like works
- **Authors Guild Lawsuit**
  - Questions if LLMs' use of copyright-protected works is 'systematic theft on a mass scale'
  - OpenAI allegedly used substantial copyright-protected content without licensing fees
- **Copyright Concerns**

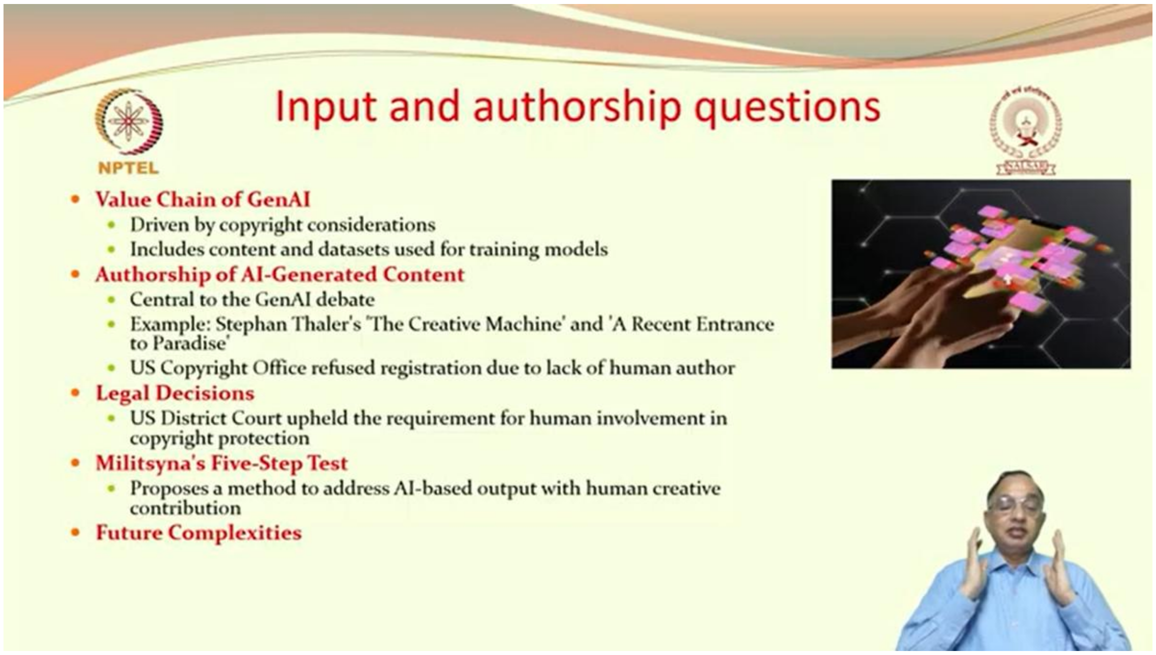
Then the copyright challenges in the European context are that it facilitates compliance with GDPR; this is possible. But then it creates new challenges for copyright-related issues. On the impact of human authors - it distances the romanticized human author when it comes to the core of copyright. For example, ChatGPT uses large datasets and high computational power, employing complex words and mimicking human-like language. Where is the human author? He or she is being mimicked. So, the Authors Guild's lawsuit questioned whether LLMs' use of copyright-protected work is systematic theft on a massive scale, which is true. So, they said OpenAI allegedly uses substantial copyrighted content without giving a licensing fee, so the copyright concerns are very significant, particularly when synthetic data is being created without obtaining any proper licensing.



GenAI, synthetic data & the need to remunerate the human author

- **Generative AI and TDM Process**
  - Machine reading of large data volumes to discover patterns
  - Generates new knowledge and extracts insights
- **Legal Debates and Lawsuits**
  - Permission of authors for training ML systems debated
  - Many lawsuits in the USA and one in the UK
  - Defendants include OpenAI, Stability AI, Meta, and GitHub
- **Training and Copyright Infringement**
- **Human Author Remuneration**

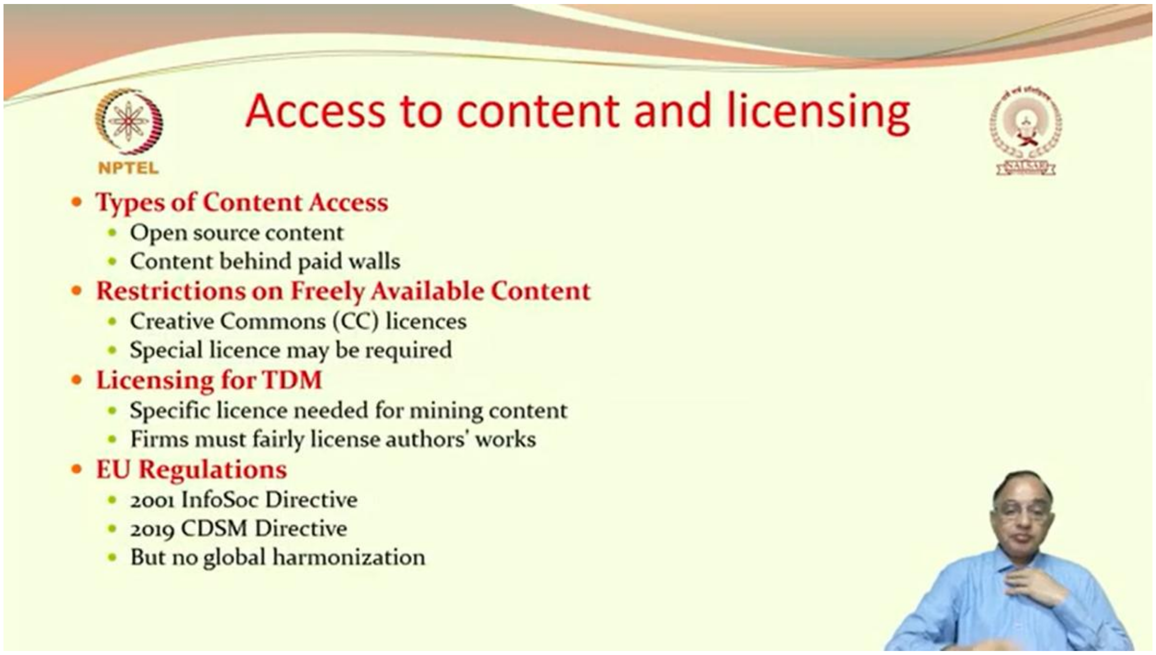
So, we need to remunerate the human author. Machine learning does result in large volumes of data for discovering patents. It creates new knowledge. It also generates a lot of insights. But permission from authors to train LLM models is becoming a topic of debate. Lawsuits are being filed one after another in different jurisdictions. And in the different jurisdictions, the same party is almost always the defendant, whether it's OpenAI, Stability AI, Meta, or GitHub. And then, training and copyright infringement are huge issues. How do we remunerate the real human author when so much data, so much synthetic data floats around?



Input and authorship questions

- **Value Chain of GenAI**
  - Driven by copyright considerations
  - Includes content and datasets used for training models
- **Authorship of AI-Generated Content**
  - Central to the GenAI debate
  - Example: Stephan Thaler's 'The Creative Machine' and 'A Recent Entrance to Paradise'
  - US Copyright Office refused registration due to lack of human author
- **Legal Decisions**
  - US District Court upheld the requirement for human involvement in copyright protection
- **Militsyna's Five-Step Test**
  - Proposes a method to address AI-based output with human creative contribution
- **Future Complexities**

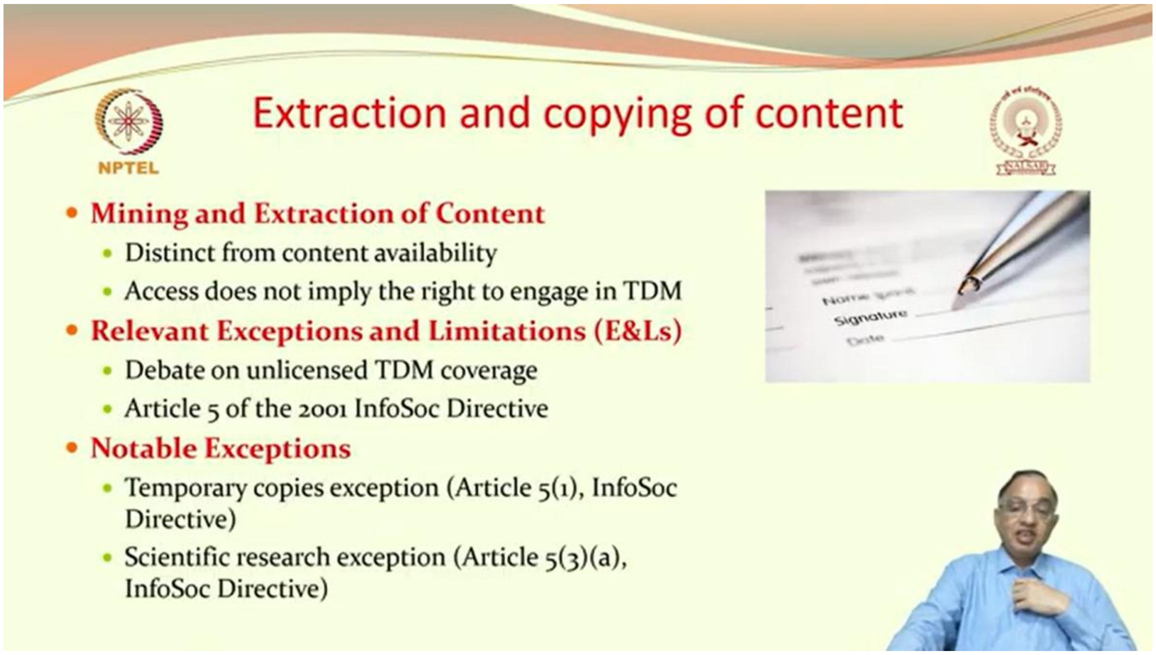
The question of the value chain of general AI is again derived from copyright considerations. It includes content and data sets used for training. So where exactly can the author figure out the value chain and then say, "Look, my material is there in full form, but here my material is there in synthetic data form"? And then the authorship is again a question because we will see later in Stephan Thaler's "The Creative Mission" and the recent "Entrance to Paradise" that U.S. copyright refused registration due to the lack of a human author, as they are AI-generated content whose specific authorship the U.S. Copyright Office refused to register. Then one legal district that upholds the requirement for human involvement in copyright protection. Then Militsyna's Five-Step Test also comes in to address the AI-based output with human creative contribution. In the future, things are going to be all the more complicated.



Access to content and licensing

- **Types of Content Access**
  - Open source content
  - Content behind paid walls
- **Restrictions on Freely Available Content**
  - Creative Commons (CC) licences
  - Special licence may be required
- **Licensing for TDM**
  - Specific licence needed for mining content
  - Firms must fairly license authors' works
- **EU Regulations**
  - 2001 InfoSoc Directive
  - 2019 CDSM Directive
  - But no global harmonization

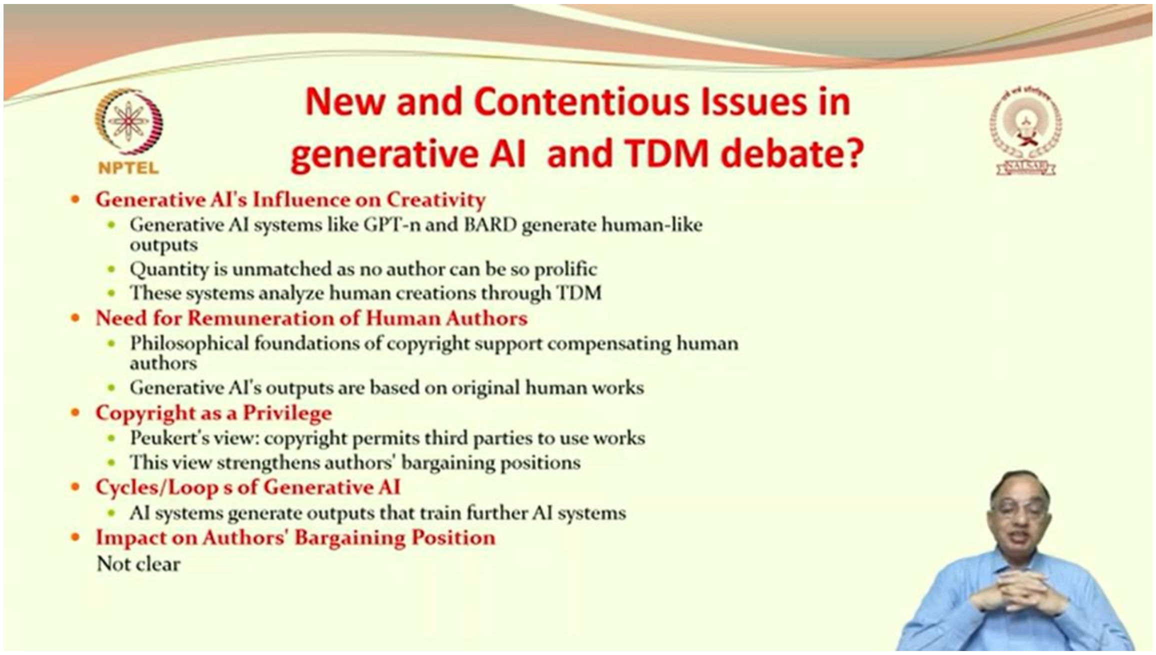
So, access to content and licensing becomes somewhat difficult and somewhat easy if open source content is made available, but then the lack of content becomes payable paywalls, and then you can use Creative Commons licenses; some special licensing may be required for TDM, which is for text and data mining purposes, and then a specific license will be needed for mining, which is for text and data mining license purposes. Firms must fairly license authors' works, but then who will do it? The European Union's regulations are in place, but there is no global harmonization when it comes to text and data mining.



Extraction and copying of content

- **Mining and Extraction of Content**
  - Distinct from content availability
  - Access does not imply the right to engage in TDM
- **Relevant Exceptions and Limitations (E&Ls)**
  - Debate on unlicensed TDM coverage
  - Article 5 of the 2001 InfoSoc Directive
- **Notable Exceptions**
  - Temporary copies exception (Article 5(1), InfoSoc Directive)
  - Scientific research exception (Article 5(3)(a), InfoSoc Directive)

An extraction copy of content is again distinct from content availability; access does not give you the right to engage in text and data mining. In the sense that when I say I give access for you to go through my website, that does not by default extend the license to engage in text and data mining to you. Of course, we discussed that there are relevant exemptions available, and then there are relevant exceptions available for scientific and other research purposes.



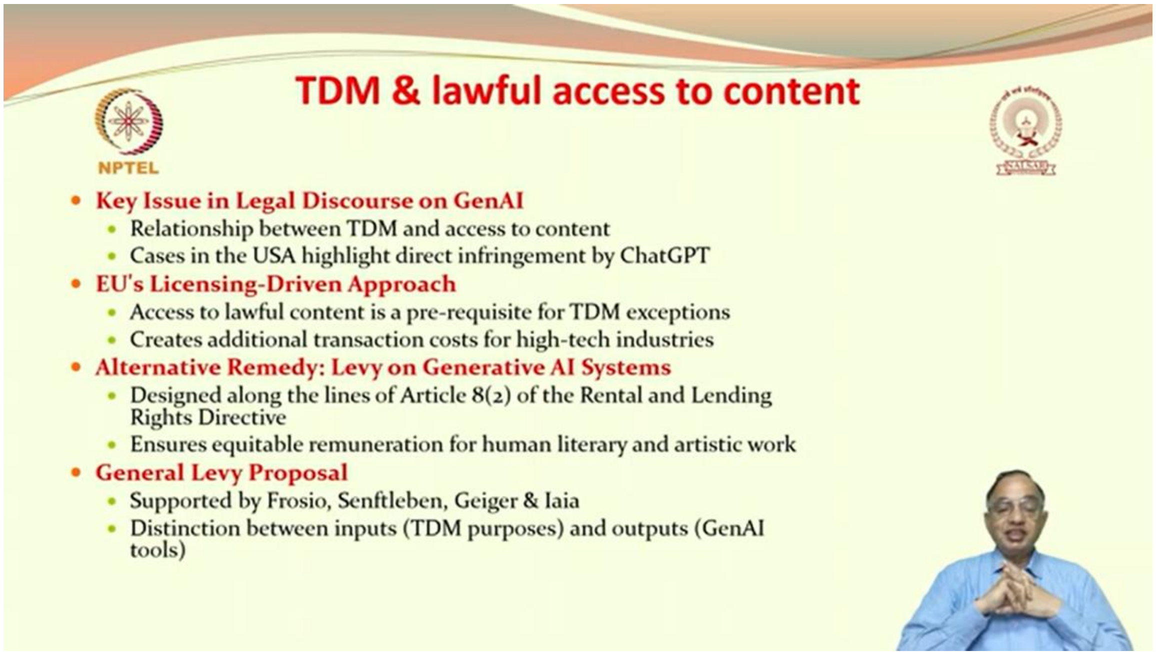
New and Contentious Issues in generative AI and TDM debate?

- **Generative AI's Influence on Creativity**
  - Generative AI systems like GPT-n and BARD generate human-like outputs
  - Quantity is unmatched as no author can be so prolific
  - These systems analyze human creations through TDM
- **Need for Remuneration of Human Authors**
  - Philosophical foundations of copyright support compensating human authors
  - Generative AI's outputs are based on original human works
- **Copyright as a Privilege**
  - Peukert's view: copyright permits third parties to use works
  - This view strengthens authors' bargaining positions
- **Cycles/Loop s of Generative AI**
  - AI systems generate outputs that train further AI systems
- **Impact on Authors' Bargaining Position**

  Not clear

But then, generative AI's influence on creativity is very significant. Particularly when generative AI systems like bots generate human-like outputs, the quantity is unmatched, but no author can be so prolific. And then they analyse human life, creativity, and human creations through text and data mining, so the need to remunerate authors comes from the philosophical foundation of copyright supporting and compensating human authors.

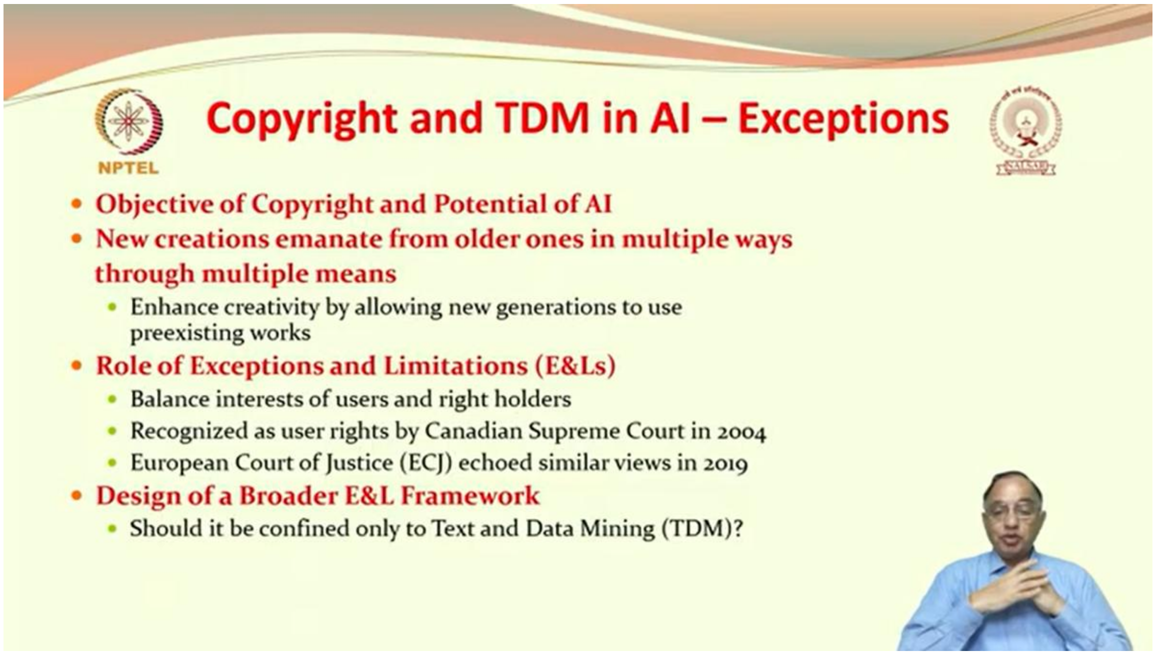
Generative AI outputs are based on original human works; copyright is a privilege, but then copyright permits third parties to use works, but to what extent? This right paradoxically strengthens the bargaining position, but how far authors will be able to bargain is a big question. And then, when the cycles and loops of generative AI happen repeatedly, the impact on authors' bargaining position is not very clear.



TDM & lawful access to content

- **Key Issue in Legal Discourse on GenAI**
  - Relationship between TDM and access to content
  - Cases in the USA highlight direct infringement by ChatGPT
- **EU's Licensing-Driven Approach**
  - Access to lawful content is a pre-requisite for TDM exceptions
  - Creates additional transaction costs for high-tech industries
- **Alternative Remedy: Levy on Generative AI Systems**
  - Designed along the lines of Article 8(2) of the Rental and Lending Rights Directive
  - Ensures equitable remuneration for human literary and artistic work
- **General Levy Proposal**
  - Supported by Frosio, Senftleben, Geiger & Iaia
  - Distinction between inputs (TDM purposes) and outputs (GenAI tools)

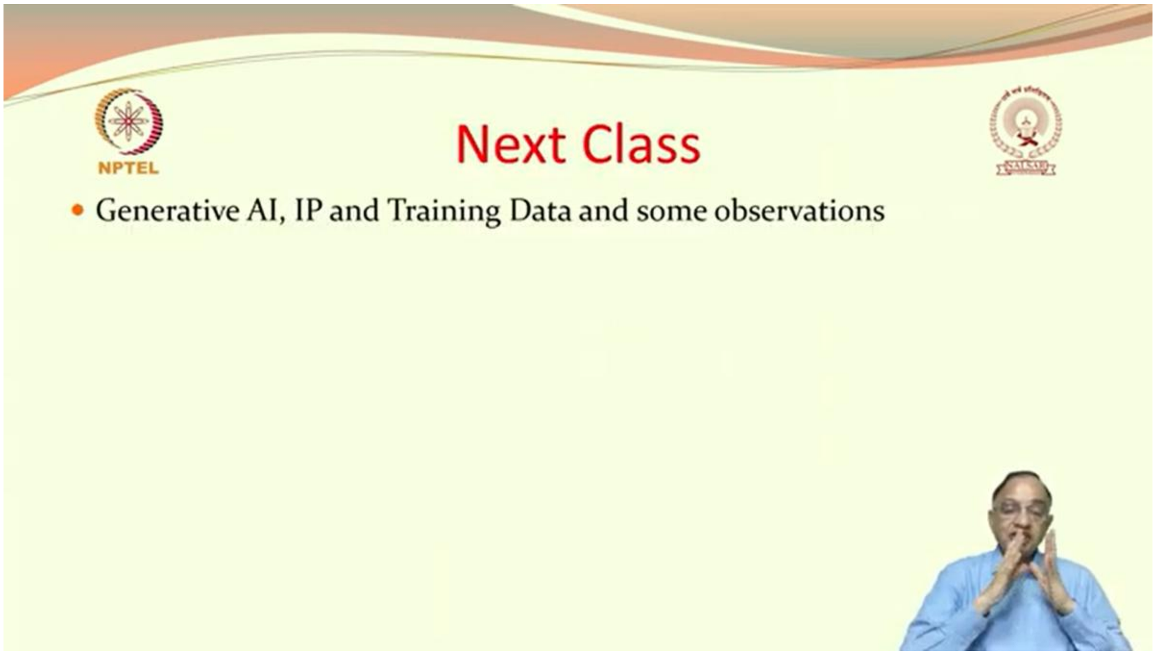
A key issue in legal discourse is that the relationship between what comes under TDM and what access to the content is. Cases in the USA highlight the infringement by ChatGPT. Again, the EU's license-driven approach gives lawful content access as a prerequisite for the TDM exception. But then, there are additional transaction costs for the high-tech industry when it wants to engage in text and data mining. The alternative remedy we saw is a levy. Again, it's problematic. How to ensure equitable remuneration is again a major issue. Again, we said, "The support base, the levy, general levy is supported by some authors, but the distinction between inputs and outputs is not very clear in this."



Copyright and TDM in AI – Exceptions

- Objective of Copyright and Potential of AI
- New creations emanate from older ones in multiple ways through multiple means
  - Enhance creativity by allowing new generations to use preexisting works
- Role of Exceptions and Limitations (E&Ls)
  - Balance interests of users and right holders
  - Recognized as user rights by Canadian Supreme Court in 2004
  - European Court of Justice (ECJ) echoed similar views in 2019
- Design of a Broader E&L Framework
  - Should it be confined only to Text and Data Mining (TDM)?

So, the objective of copyright is the potential for AI to create new works that emanate from older ones in multiple ways through multiple means; enhancing creativity is feasible by creating new generations to use pre-existing works, and then the role of exemptions and limitations is to balance this, but how to do that? If the Canadian Supreme Court in 2004 recognized this as a user right, the European Court of Justice echoed a similar view in 2019. But to design a broader exception and licensing work, should that be confined only to text and data mining, or should that be extended to another use, particularly in the context of AI? It's a big question.



Next Class

- Generative AI, IP and Training Data and some observations

So, in the next class, what we will do is look at generative AI, IP, and training data, and then we will make some final observations in the context of copyright and AI.