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 – 03
Lecture – 15

Artificial intelligence, law and justice, session 15, Machine learning, Copyright, and Overview.

We will do a recap of the last session, which is the first session on AI and copyright. In the last session, we gave a quick overview of the issues in AI and copyright, then we also looked at some of the concerns of the creators and artists, and we contextualized them in

the broader political economy of copyright, IP, and innovation. In fact, we also made it clear that we should not simply go by the rhetoric or what people claim to be safeguarding the interests of this party or that party. Rather, we need to look at the interests that lie behind this rhetoric and then decode them so that we know what exactly matters in the debate. What matters in the debate is the interest and the economic stakes rather than simply the terms that are being used by the people.

So, what are the copyright implications of data scraping and mining? Data scraping and mining, as you all know, are very fundamental. All of us may not be doing it, but whenever we look at a website or whenever a bot scrapes a lot of information, a lot of data is being picked up. And then a lot of applications, including whenever you compare or contrast, or whenever you run an application that compares prices from different websites or different vendors, do collect data and then put it together; but then data scraping is so fundamental that it happens on a regular basis, done by so many bots that are roaming on the internet. Copyright has actually got a lot of implications for data scraping, mining, and learning because the whole corpus of the material put together for machine learning purposes, developing LLMs for applications, will include a lot of copyrighted works.

And those copyrighted works being included may or may not be acknowledged. So, when an application or LLM is built using digital copies of copyrighted works, it will still impact the economic right of reproduction. So, the copyright holder or the one who holds the license, their economic rights will be infringed because they may not be getting anything from it. And the changes in the material can be considered adaptations. For example, my chapter or my work is being used, or someone else's painting is being used, or someone else's music tunes are being used in the application, and then they are tweaked, or some changes are made.

Whether they slightly changed, fully modified, or to an extent changes happened is a different question. But that could be considered an adaptation because they are being

used as the source of raw material. For example, if I take a tune from a famous film composer and then use it in my application or an LLM, and then I make some changes. When the original tune has guitar, I bring in something else like an accordion or another instrument. If I replace one instrument with that and then tweak it, then that becomes an adaptation to some extent. So, research or text data mining exemptions are available, but they are not universally applicable. And, more importantly, not all activities of researchers and firms are covered by that because there are some serious limitations in the sense that there is no blanket license or blanket exemption for all activities of researchers and firms in the name of data scraping and mining for research purposes. We will see some of these in detail later.
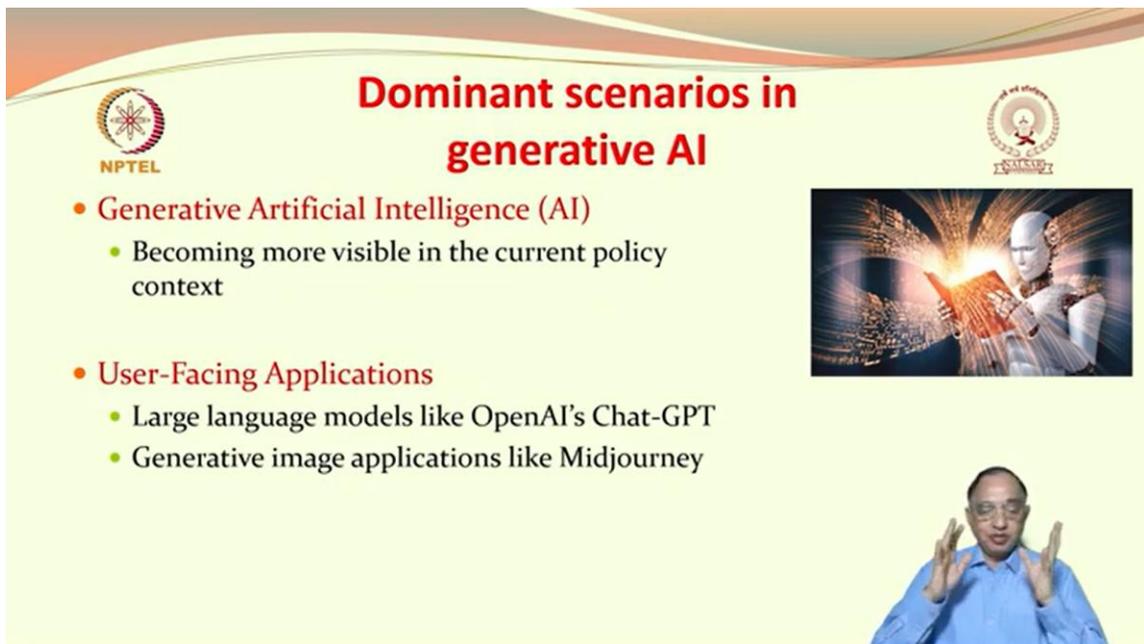

So, what are the real legal grey areas here? Layered protection complexity is present, and that becomes confusing for users and regulators because copyright protection is a layered protection in the sense that it is not something that can be easily understood or put into practice. Why? Because one - the copyrighted materials are available in different formats in a way that could not be made sense of. For example, digital rights are different from the right to reproduce a text. Similarly, the right to reproduce itself is something that varies from context to context, from country to country. The right to reproduce for academic purposes is different from the right to reproduce for commercial purposes. And then the availability of digital copies in multiple forms, without a specific ability to identify the copyright holder or copyright owner, has become one more issue. The second problem is that copyright comes in layered protection in the sense that for what purposes this protection is available to what extent becomes unclear.

There is layered protection for different types of materials such as for different copies, for example, sound recordings, digital copies, and then multiple audiovisuals being made particularly in films, the layered protection becomes very complicated. It will be difficult to identify in the film whether the music is fully protected by the music composer, whether he has licensed it to someone else, or whether the producer is the final owner of

the copyrighted material. And then the layered protection, also depends upon the use, the medium, and the format, making things very difficult. Machine learning, per se, is in a legal grey zone because it learns from established data processing and analysis cycles, and the idea of machine learning itself is not very old. What is making things all the more difficult is that these powerful models, LLMs, are major applications that get integrated into the services of different organizations, and then they proliferate; when they are rapidly put to use, there are a lot of grey areas. In the sense that will anyone know or will any copyright holder know what application has used the materials in what format, for what purposes, and in which quantity, and as generative AI becomes more and more used in consumer-facing applications like Canva or Midjourney or in different sorts of applications in different phases or in combination with them, its visibility increases. But then this has resulted in a lot of lawsuits being filed against OpenAI, and then by publishers against some of the people who have developed and deployed them, and then in some cases by the real copyright holders who themselves are shocked to find that materials have been used in generative applications without their own permission.

So, the dominant scenario here is that generative AI is becoming more and more prevalent in user-based applications, and then generative AI applications like Midjourney can use the material; they can use the paintings, and then the user can also use the materials available to them to come up with something new. So, generative AI, as well as user-facing applications, being the major developments, also results in a lot of grey areas for copyright exemptions and applications.

Continuously Deployed vs. Off-the-Shelf Models

- **Continuously Deployed Models**
  - Rely on current data updates

- **Off-the-Shelf Models**
  - Fine-tuned or aligned for specific purposes
  - Data collection is essentially complete

Then, to simplify matters, we can look at them as continuously deployed and off-the-shelf models. Off-the-shelf models are the models that could be deployed for different applications but for specific purposes. For example, models in insurance, models in healthcare, and models for different educational purposes. Here, they are not like the typical model that needs to be constantly updated in terms of data from other sources. They are not the models that would need constant data from external sources. They would need data, but that could be internal. For example, a health-based AI, once trained, may not need as much data from external sources because the data it needs for further training and analysis would mostly come from within the organization, which could be a hospital or research organization. So, these are off-the-shelf models; they have already been fine-tuned, tested, and then aligned for specific purposes. So, ChatGPT is different from these models, and these models are also proliferating in every direction, left, right, and centre. So, the problem is who will really monitor which materials have been used in training when these things are being continuously deployed.

European Legal Analysis of Reproduction Rights

- Right of Reproduction under ISD
  - Contained in Art. 2 of the Information Society Directive (ISD)
  - Temporary exception in Art. 5(1) ISD
  - Enables technological development
- Tension between ISD and CDSM Directives
- Copyright in the Digital Single Market (CDSM)
- Directive, brought into law in 2019
  - Arts. 2/5(1) ISD vs. Arts. 3 and 4 of CDSM
  - Text and data mining exceptions
- Research Use under Art. 3 CDSM
  - Subject to lawful access
  - Contracts involved
- Opt-out under Art. 4 CDSM
  - For non-scientific purposes
  - But there are issues in lack of harmonization between ISD and CDSM

So, if we understand them as reproduction rights, then the European Union's Directives deal with them in different ways; we will look at the right to reproduction of the Information Society Directive (ISD), which is available as a temporary exemption under Article 5(1). This enables technology development because this is a temporary exemption available, but there is also the Digital Single Market (CDSM) of the European Commission. So, there's a tension between them. And then the CDSM directive was brought in 2019, so the question is: Article 2/5(1) of one of the ISD and then Articles 3 and 4 of CDSM—are they really fully compatible, particularly when it comes to text and data mining exceptions? We are just giving an illustration here; we are not going into the fuller details. And then the research use exemption is available under Article 3 of CDSM. But again, this is subject to lawful access, and then contracts are involved here.

Then, Article 4 of the CDSM gives an option to opt out. So, for non-scientific purposes, one can simply opt out of the Research Uses Rights. Then there are issues like the lack of harmonization between ISD and CDSM, and CDSM is again applicable only within the European common market area; it is not globally applicable. So, when Europe tried with the ambitious plan for a single digital market, they came up with a lot of notifications and directives, which they thought would make the European market the single biggest market for digital products, innovation, and services. So, they were trying to pioneer some sort of directive that would balance the interests of the rights holders and those who build innovations on them.

Scientific research uses

- Unclear Copyright in Machine Learning
  - Uncertainty affects research uses, as seen in
  - Lawful access terms control research possibilities and costs
- Licensing Arrangements for Research
  - Providers set terms for valuable data sets
  - Right holders may license material to AI firms
  - Threats to withdraw archives from public interest research
- Live Online Services and Public Interest
  - Unclear line between competitive control and public interest
  - EU's TDM (Text and Data Mining) exceptions and
  - SGDR not successfully drawn
  - What is the Sui Generis Database Right (SGDR) and how does it relate to other rights in Databases?
  - For legal purposes, a 'database' means "a collection of independent works, data or other materials arranged in a systematic or methodical way and individually accessible by electronic or other means
  - https://www.openaire.eu/sui-generis-database-right-sgdr
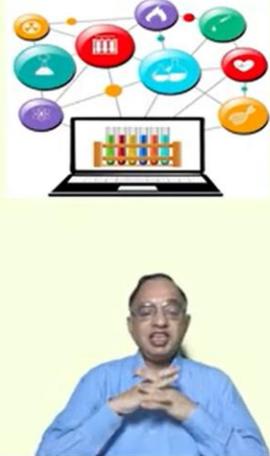
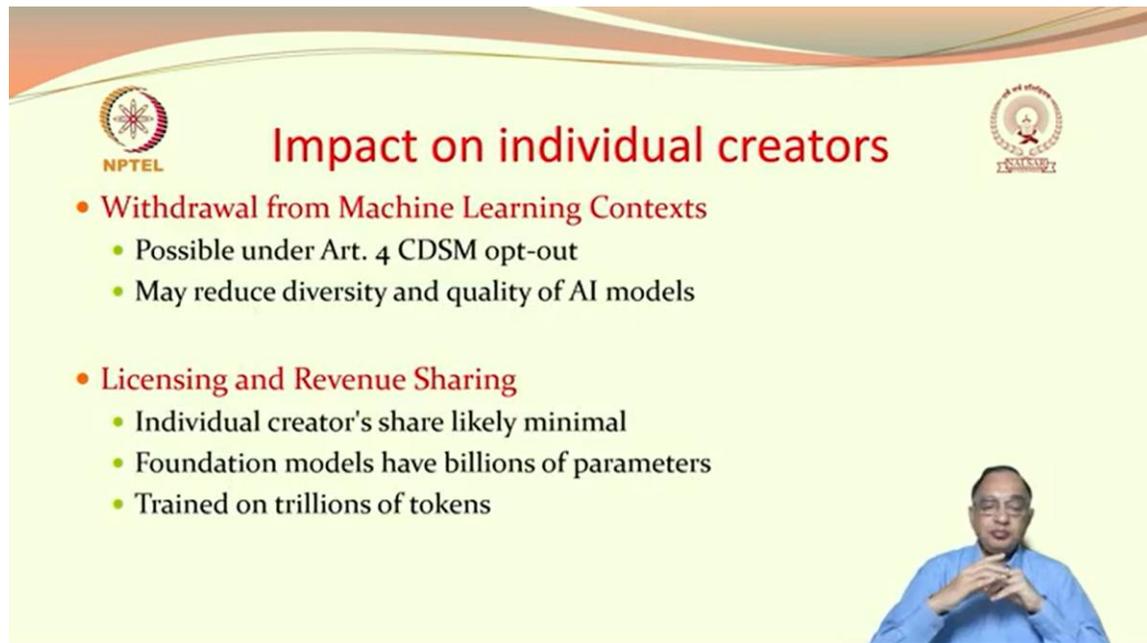But scientific and research uses are also very major applications here. So, when it comes to copyright in machine learning, this is a very grey area because the uncertainty affects the research uses, as seen earlier. Lawful access terms control research possibilities and costs because lawful access comes with costs and restrictions in the sense that the right holder may impose some limitations that this could be used only for the purpose of research and not for further commercial development or commercialization purposes. Then the licensing terms might be available in a very restricted form for the valuable datasets. So, the rights holder has the option to license the materials to AI firms.

But the problem here is that when they want to give them a license, the problem arises when the rights holders find that the materials which are already in archives will be available but not for public interest purposes, and then they may try to club different types of users; if you are going to use an earlier archive, which I allowed earlier as an exempted one for public interest research, it will no longer be available to you. So, they could withdraw it, they would put it into a different category where you really need to go and ask for some license to use them. And then there are a lot of live online services available, which could be news broadcasts, which could be educational broadcasts and telecasts, which could be simple podcasts, which also use audiovisual material. The public interest component would be there, but there is a blurred boundary between them in the sense of what constitutes competitive control, commercial interest, and public interest. And then, to what extent the TDM thing will be applicable here is not fully known.

Unfortunately, Europeans are getting struck because of this because they pioneered what they call Sui Generis Database Rights, and then the Sui Generis Database Rights are something unique in Europe. No other country or continent in the world recognizes something called Sui Generis Database Rights, and that is a really complicated matter. Because, for legal purposes, a database is a collection of independent works, data, or other materials arranged in a systematic or methodical way and individually accessible by

electronic and other means. So, once the database is created, that automatically grants the database creator or the database license holder a lot of rights. So, Europeans wanted to try something called sui generis database rights, but in the context of electronic digital products and other things that has reached a sort of confusing reality for them. So, this again gets reflected in the AI development and the training purposes, what will be the impact of these SGDR is not clear.

So, one option would be to withdraw the machine learning context under Article 4 of CDSM. One could opt out. But this opt-out, if done on a large scale by different authors for different entities holding copyright, is problematic. This will reduce the diversity and quality of AI models because ultimately AI models depend upon the availability of large and well-developed data and other materials for training and other purposes. So, when so many of them opt out, stating that my material is not available for the AI development models, AI Models may be poorer in terms of content, or they may not be very rich in terms of diversity of the content. Okay, even if the creator or the one who holds a license is not willing to opt out and still wants to stay stating that mine is available, the individual creator's share is likely to be very minimal because the foundation models have billions of parameters and are trained on trillions of tokens. So, when so much of them are there, the value or the share of the individual creator's copyrighted material being used will be very minimal. It could not even be something like 0.00002% of that, so when the most minimal is being used, the creator cannot expect too much from them when the share is so minimal.

Policy Choices for AI and Copyright

- Obligations to disclose training data
  - Importance of transparency in development and use
  - Legal and ethical considerations
  - Moral rights of authors and creators
- Collective licenses for machine learning
  - Shared access to datasets
  - Collaboration among organizations
- Collective licensing is a simple way to manage the reuse of small extracts of published copyright material. It provides an efficient, cost-effective means of satisfying the two sides of the equation: those who wish to use extracts of content, and those who hold the rights to it.

The value for licence holders
For organisations that wish to reuse copyright content without breaking the law, a collective licence—sometimes referred to as a blanket licence—provides access to a rich variety of material in return for an appropriate fee. Schools, universities, central and local government and businesses all use blanket licences to secure permission to use extracts of content by copying, scanning and other means
https://www.pls.org.uk/collective-licensing/what-is-collective-licensing/
Promote Open source
Benefits of open source software
Community contributions and improvements

And then this results in the question of policy choices for AI and copyright, particularly in this context of training data. The obligation to disclose training data is in the public interest. But should that be made compulsory by the policyholders? Because if that is done, there is transparency in development and use, as everyone will know what the sources used for training are, from where they have been taken, and what sort of materials went into building the models. But then there are legal and ethical considerations in the sense that if it is known that much of the material has been taken without any legal licenses and in an unethical way, in the sense that the user or the copyright holder has not authorized it, it has been taken in an infringing way or it has been taken from depositories that are, in fact, violations of copyright; it is totally illegal and then unethical. The other unethical issue arises when the material is taken in full form for training purposes and then, in developing outputs and for other purposes, that material is mutilated or damaged, or its artistic integrity is denied, or it is totally damaged in the sense that the artistic integrity of a painting, sculpture, or work of art is denigrated or mutilated or damaged in one way or another; it affects the moral right of the author. The moral right of the author will be available even if the whole copyright is assigned, or copyright ownership is not a precondition for asserting the moral rights of the author.

But for training purposes, when models are being used in AI or being developed, it is a possibility that moral rights could be violated or ignored in the process. So how can the authors assert their moral rights, particularly when they are not even aware that their material is being used and is being violated? One solution that has been given is collective licensing. Now collective licensing is a valuable option in some circumstances when I need material, but only small extracts from different sources. For example, I want to publish a book on cubism. And then I want to illustrate only one or two paintings of a painter; I don't need access to the hundreds of paintings, nor do I need access to many of the works.

I need only one or two to highlight certain things. Then I can negotiate because this is a small extract, but if I also want to reuse it. So, when I want to reuse small extracts from different sources or copyrighted materials, I can go for collective licensing. And this is an efficient collective way because it fulfils the need of the person who wants to use them for a legal purpose, only the extracts, not the full one, and then who holds the right to it. For organizations that want to reuse copyrighted materials without breaking the law, collective licensing helps in the sense that I can still compile, I can still publish, I can create something new by assembling the works of others in a different way, in a small way, and then I will still be able to do those things.

And this gives access to a rich variety of materials in the sense that I want to pick some samples of music tunes from somewhere else, I want to pick some paintings from somewhere else, I want to pick some copyrighted text from somewhere else, I want to pick some audiovisual material but not in full form. So, I want to access only three minutes of a popular film song, just the incoming interlude of that, or I want to access only the first humming voice of a popular film or music, or I want to access only the specific notation in a concert or in a song, only to that extent. Then I can go for a collective licensing mechanism. But the problem here is that a collective licensing mechanism would again need negotiations with so many rights holders, and it is not something that should always be attempted when you are not sure of the real owner. In the sense that if you don't pursue collective licensing and that option becomes difficult, it is better not to venture into using too many materials from different sources, assuming that since I am only using very little material from here and there, I am not violating copyright; that will be lethal. In the sense that it could be too damaging because too many people can sue me for that, and so in the context of AI, this becomes all the more problematic, although there are a lot of legal grey areas there. One option would be that policy choices for AI and copyright are to promote open source, open-source software, and open-source materials; further innovation is possible, and there are community contribution improvements that could be recognized and rewarded.

Collective Licenses for Machine Learning

- **Potential Benefits**
  - Prevents innovation hold-ups

- **Feasibility Issues**
  - Difficulty in assembling sufficient rights
  - Becomes difficult when it involves different types of contents such as text, sound, audio-visual, data from multiple sources and using small portions of them

So collective learning for machine learning is a good option because it doesn't cause too much bother. It promotes innovation and prevents innovation holdups due to the lack of available licensing. But the difficulty comes: I need to assemble sufficient rights from different stakeholders and different copyright owners. In the sense that it will often be very difficult to really identify the owner. For example, in the case of some works like films, it's a legal grey area in some contexts as to who exactly holds the rights. And then the right holder may not be available, or the right holder might have expired, or the right itself might have been transformed, or it might have gone from one hand to another, or it might be available with someone who asserts his or her claim, but then I am not sure whether that person is the ultimate right holder or not. And then a lot of orphan works become a major problem because nobody holds the rights to orphan works, and even identifying the rights holder becomes too problematic. In some specific instances, particularly for films and then for theatrical works like drama, who exactly holds the right will be difficult to decipher and negotiate.

And then multiple parties may stake their claims. And then how much of the sufficient right I need to really look into becomes a problem. In the sense that is it enough if I assemble the right only for this purpose, or should I assemble the right for reuse or multiple purposes, this is also becoming an issue. So, when I use or when I need access to different types of materials, like text, audio, visual data, and multimedia resources, it becomes very difficult. The feasibility issue is something where I need small portions of lots of multimedia works, different works from different sources and different collections. Collective licensing as a mechanism is good. Although there are a lot of feasibility issues involved.

Levy approach for equitable remuneration

- Levy Approach Proposal
  - Suggested by Prof. Martin Senftleben
  - Focus on equitable remuneration to authors
- "Generative AI and Author Remuneration"
- *International Review of Intellectual Property and Competition Law* 54 (2023), pp. 1535-1560"
- https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4478370
- EU's Rental Directive as a Model
  - Levy paid by providers of generative AI systems
  - Funds directed to social and cultural funds of collective management organizations
  - Purpose: Fostering and supporting human literary and artistic work
- Challenges of Levy Approach
  - Bureaucratically challenging
  - Issues around levy efficiency remain unresolved
  - Litigation on who pays and who receives

Coming to the next point: how do we solve this problem? Don't go for collective licensing; instead, levy in the sense that I levy 2% or 3% of the value, and then levies, as we all know, are very small percentages, but when done in sufficient quantities, they become sizable amounts. So, Professor Martin Senftleben said, we can go for a levy approach because that will give equitable remuneration to authors. So, he came up with this idea in 2023, which has been further developed and written by him in "generative AI and author remuneration". And then the available sample model here is the EU rental directive as a model. Here, the levy is provided by the providers of the generative AI systems.

And then the funds that are directed to the social and cultural benefits of the collective management organizations. The idea here is to foster and support human literary and artistic work. But the real challenge is that organizing this is very difficult in the sense that someone really has to identify who will be the ones who will pay the levy, then who will collect it, who will distribute it, and even to identify cases where the levy is really collectible or applicable is a major problem for the simple reason that too many things are happening, too many models are being developed, and a lot of data is being acquired for or being put to use for training purposes in a thousand possible ways. So, issues around levy efficiency remain unresolved in the sense that is a levy a really efficient way because there is a whole lot of cost, including developing and institutional mechanisms to assess, collect the levy, and then to distribute it; that is the first point. Then, is it really going to be an efficient mechanism if, ultimately, at the end of the day, the amount that is collected as a levy is not that big or is not substantial compared to the people to whom it should be distributed and compared to the cost of collecting the levy and identifying the people who will pay the levy? Then, if the authors or creators are not going to benefit substantially from it in the sense that what they will ultimately get from the amounts collected as a levy is peanuts, or say $10 or $20 an author will get from an LLM or from a model that has been developed using his or her material, that doesn't make much difference. So, litigation is also there regarding who will pay, who will receive, and who

will distribute it. So, the levy approach is good in theory, but putting that into practice has many implications.

One way is to promote open-source models because the advantage of open source is in the public interest. And then there is transparency because open-source models can outperform closed AI systems. And then open-source applications have been used in the past in protocols and operating systems. And more importantly, the licenses were well-defined and also well-recognized in terms of commercial applications. So, the common training sources can easily be built upon. Depositories can be governed by open-source licensing; for example, Wikipedia and GitHub. Different types of licenses are available, such as those provided by Creative Commons, which offers various licenses for different purposes. So, it is easy to identify the purpose, tailor the license, choose it, and make it applicable. In the sense that the open-source licensing mechanism is also easy to negotiate with, we know about Creative Commons licensing, and there are many licensing forms available. Since there are huge licensing models available and there has been a lot of experience over the last two and a half decades in using open-source model licensing for different purposes, this is a viable option where the rigidities of copyright can be removed and a lot of flexibility is also being made available.

Lifecycle of Machine Learning and Legal Implications

- **Legal, Technological, and Contractual Opacity**
  - Leads to undesirable allocation of licenses and obligations
- **Risks of Training and Deploying Unlicensed Models**
  - Currently risky in the EU
  - Will remain risky for the foreseeable future
- **Movement Towards Fully Licensed AI Copyright Environment**
  - Regardless of available exceptions
- **Key Question: Obtaining Suitable Licenses**
  - Where to obtain a license
  - Under what conditions
  - Global Deployment of AI/ML models vs. National Variations
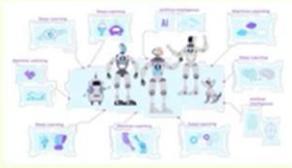
So, the legal technical opacity in machine learning is also present, and unlicensed models always come with a lot of risk. It is better to move towards fully licensed AI copyright, where you know that everything is clear in terms of legal licensing, legal rights, and responsibilities. So, in respect of the available exceptions, it is better to move towards a fully licensed copyright environment so that you know both the persons who run them, who bought them, as well as who developed them know that their license fulfils the copyright terms and that they are crystal clear as to what they are dealing with. But then the question will be where to obtain the right license, identify the right license, determine under what conditions, considering the global deployment of AI and ML models versus the national variations in terms of copyright exemptions and licenses, and recognize that the organizations that deal with them make it very difficult to navigate or make things quite complex.

Public Benefit and Copyright Works in AI Training

- **Public Benefit of Using Copyright Works**
  - Uncertainty about the societal impact
  - Potential for a fully licensed AI environment
- **Current Copyright Solutions**
  - Controlled by major right holders and large AI firms
- **Alternative Approach**
  - Machine learning as a general purpose technology
  - "Knowing if an emerging technology is general purpose is of significant strategic importance for managers and policymakers. Such general purpose technologies (GPTs) are rare and hold potential for large scale economic impact because they push the production possibility frontier out several times (Bresnahan and Trajtenberg 1995). Examples of GPTs include the steam engine, electricity, computers, and the internet (Lipsey, Carlaw, and Bekar 2005)." Avi Goldfar, Bledi Taska, Florenta Teodoridis
    https://www.nber.org/system/files/working_papers/w29767/w29767.pdf
- **Challenges for Copyright Law**
- **Novelty of technology and its ramifications – its more than copying or reproduction**
  - Balancing market entry, open source innovation, and creator remuneration
  - Necessity to address these tensions

One way to look at this issue is that there are a lot of public benefits in using copyrighted works. But we really do not know what exactly the public benefits are when they are used in AI training because AI training itself is not even two and a half or three decades old by now. The current copyright solutions are once again controlled by major holders and large firms, which we saw in the previous session. An alternative purpose or solution that has been proposed is to treat them with machine learning itself as a general-purpose technology. General purpose technology is a technology that could further promote innovation and can go much beyond that, like the steam engine, electricity, and the internet. These are good examples of general-purpose technology where they not only add value, but also cause societal revolutions. They bring in economic transformation and societal transformation as well. So how can copyright law deal with a general-purpose technology if we develop machine learning or identify it as a general-purpose technology? So, the novelty of technology's ramifications is more than copying and reproduction. But we need to balance market entry costs, market entry benefits, open-source innovation, and then the creative remuneration, creator remuneration. And then we also need to really identify these tensions when we talk about the public benefit that would derive from copyrighted works and then when we apply open-source licensing for in-copyright AI training.

Next Session

- Generative AI and Copyright

So, with all these things coming together, we will go to the idea of generative AI and copyright in the next session, where we will look at the specific issues that need to be addressed in the interface between generative AI and copyright.