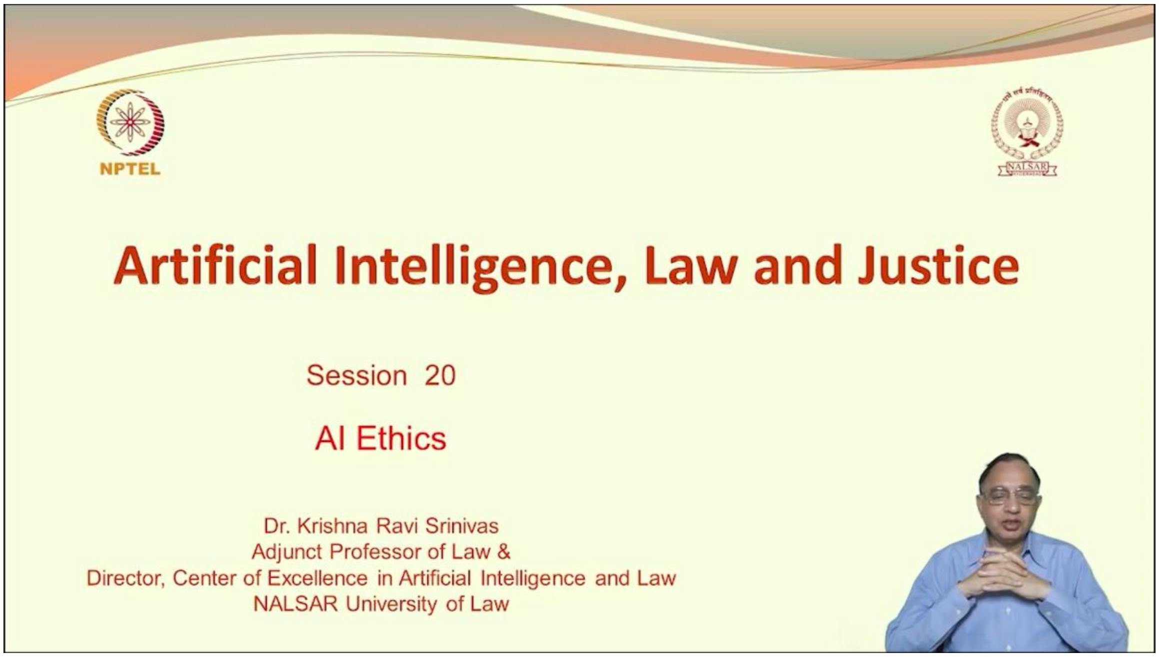
Course Name – Artificial Intelligence, Law and Justice
Professor Name – Dr. Krishna Ravi Srinivas
Department Name – Center of Excellence in Artificial Intelligence and Law
Institute Name – NALSAR University of Law
Week – 04
Lecture – 20



NPTEL

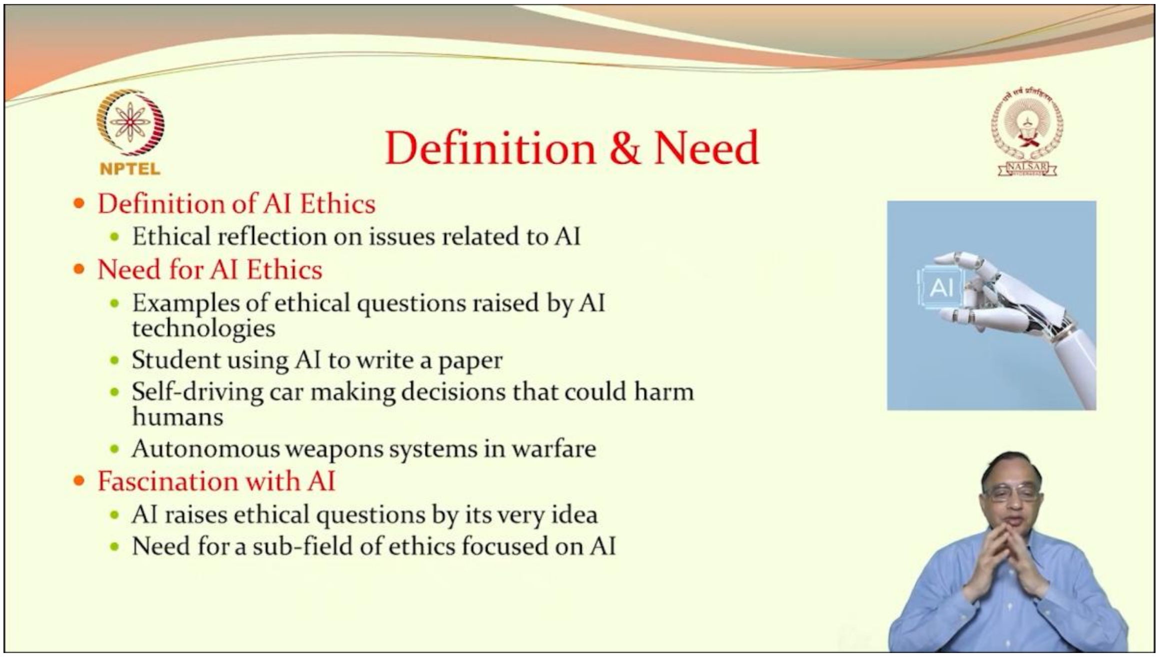
Artificial Intelligence, Law and Justice

Session 20

AI Ethics

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 20, AI ethics.

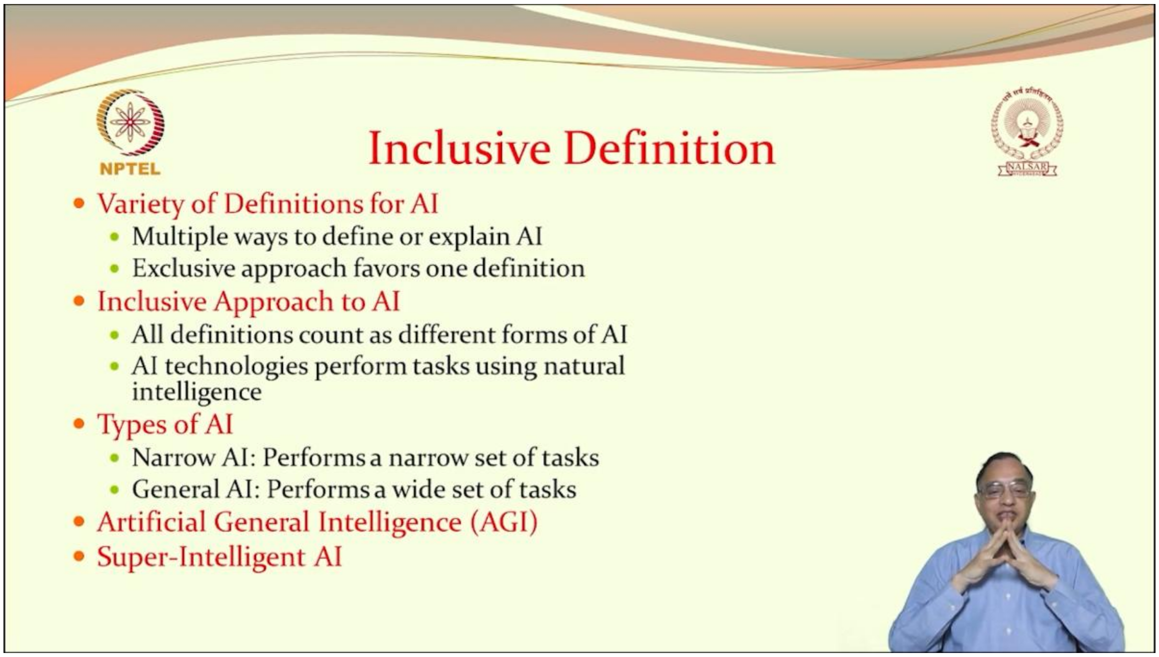


NPTEL

Definition & Need

- Definition of AI Ethics
  - Ethical reflection on issues related to AI
- Need for AI Ethics
  - Examples of ethical questions raised by AI technologies
  - Student using AI to write a paper
  - Self-driving car making decisions that could harm humans
  - Autonomous weapons systems in warfare
- Fascination with AI
  - AI raises ethical questions by its very idea
  - Need for a sub-field of ethics focused on AI

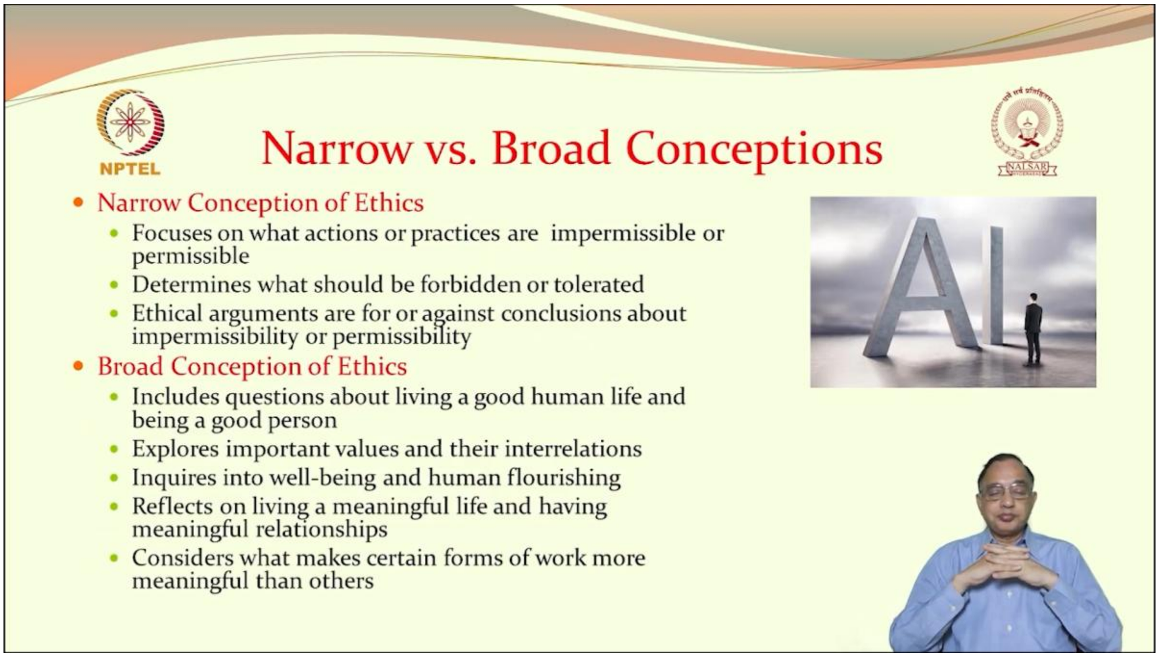
In this session, we will look into AI ethics and then discuss why there is a need for AI ethics. First of all, we need to reflect upon the issues related to AI from the framework of ethics. There are many ethical issues raised by AI technologies; for example, a student using AI to write a paper without acknowledging that he or she has used AI, and then even claiming that the paper has been written entirely by AI, with him or her just tinkering here and there. And also, the question of the self-driving car that makes decisions is that: the self-driving car, knowingly or unknowingly, can harm a person or can violate traffic rules or can result in chaos and accidents. Similarly, in warfare, lots of autonomous systems are being developed, and they have not yet been fully deployed or fully used in war, but a lot of work is going on. They are also called lethal autonomous weapon systems (LAWS). So, when these things come, there is an inevitable ethical question. Is it ethical to use them? Is it ethical to develop them? And if they have to be deployed, what sort of ethical frameworks and guidelines would we need? Again, AI itself raises a whole lot of ethical issues because people are fascinated with artificial intelligence, general artificial intelligence, and the potential of general AI to reach or surpass human intelligence. So, we need a subfield in ethics that focuses on AI. But this subfield on its own cannot find things because it either has to develop from some of the earlier theories or it has to develop new theories.



Inclusive Definition

- Variety of Definitions for AI
  - Multiple ways to define or explain AI
  - Exclusive approach favors one definition
- Inclusive Approach to AI
  - All definitions count as different forms of AI
  - AI technologies perform tasks using natural intelligence
- Types of AI
  - Narrow AI: Performs a narrow set of tasks
  - General AI: Performs a wide set of tasks
- Artificial General Intelligence (AGI)
- Super-Intelligent AI

So, the first question, or the first problem that needs to be addressed, is how do we define AI? As we saw at the beginning, there are many ways to define or explain AI. So, when we talk about AI ethics, we need to look for an inclusive approach so that all definitions can be taken as different forms of AI, and then they are trying to develop or address them in one way or another. So, all AI technologies, since they perform tasks using and imitating natural intelligence of humans, we need to have an inclusive approach there. Then we know narrow AI is something that does a very narrow set of tasks that are assigned to it or whose capacity is limited. Whereas generative AI or general AI performs a wide variety of tasks, including creative ones. So, artificial general intelligence is something that will go beyond general AI, surpassing the current versions of AI, and it will have enhanced capacity that could be equivalent to, or even surpass, humans in the future. That's what people think. So, when that happens, we will have super intelligent AI. So, the very idea or the very definition of AI needs to be examined when we discuss AI ethics, because unless we define exactly what the technology is that we are talking about, discussing ethics alone won't make sense.

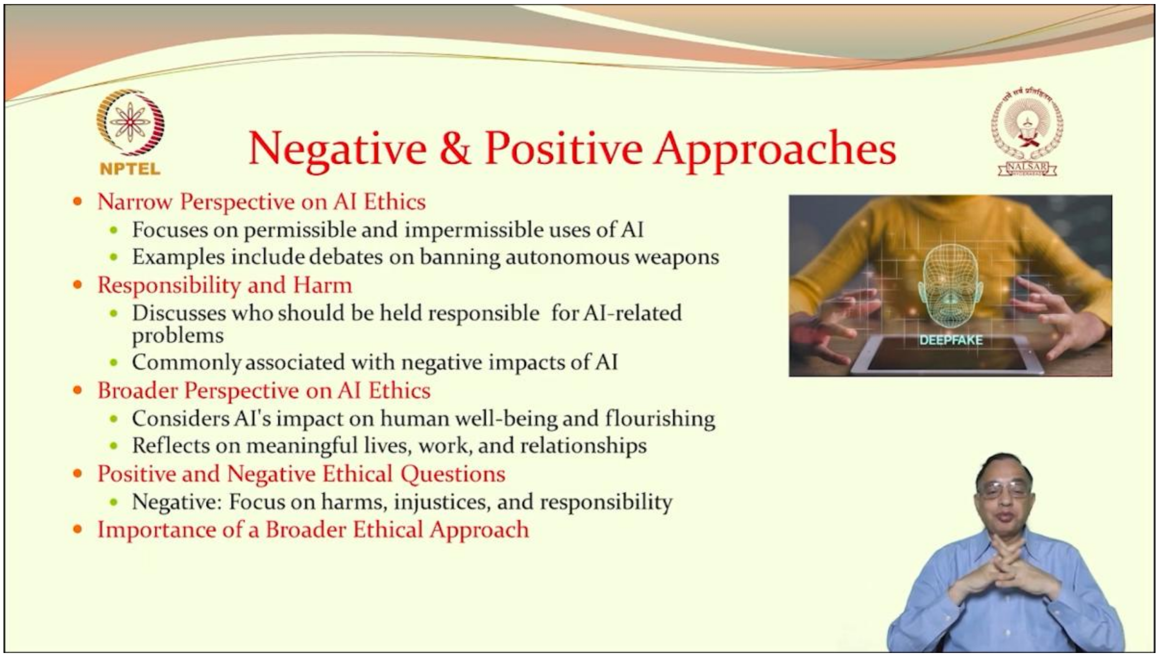


Narrow vs. Broad Conceptions

- Narrow Conception of Ethics
  - Focuses on what actions or practices are impermissible or permissible
  - Determines what should be forbidden or tolerated
  - Ethical arguments are for or against conclusions about impermissibility or permissibility
- Broad Conception of Ethics
  - Includes questions about living a good human life and being a good person
  - Explores important values and their interrelations
  - Inquires into well-being and human flourishing
  - Reflects on living a meaningful life and having meaningful relationships
  - Considers what makes certain forms of work more meaningful than others

Having said that, we need to look at two or three things. To understand ethics, we need to recognize that there are either a very narrow conception of ethics or a broad conception of ethics. A narrow conception of ethics focuses on what actions are permissible or impermissible; it simply tells us whether something is permissible or not permissible, what should be tolerated or not tolerated, or what should be totally prohibited. Then, you build up an argument either for or against one of these decisions: why we should allow it or why we should not allow it, why this should be forbidden or why this should not be forbidden. So, this narrow conception of ethics looks practically at actions, then considers what is forbidden, what is tolerated, or what could be permitted. But we need to go beyond this idea because AI is not a simple technology. AI can contribute to human development and can contribute to society in more than one way. So, AI cannot be reduced to a narrow idea of itself. Then we need to look at a broad conception of ethics. We need to see how living a good human life and being a good person relate to how AI contributes to it or how AI constrains it. So, we need to bring in the important values and norms, and then the relationships between them.

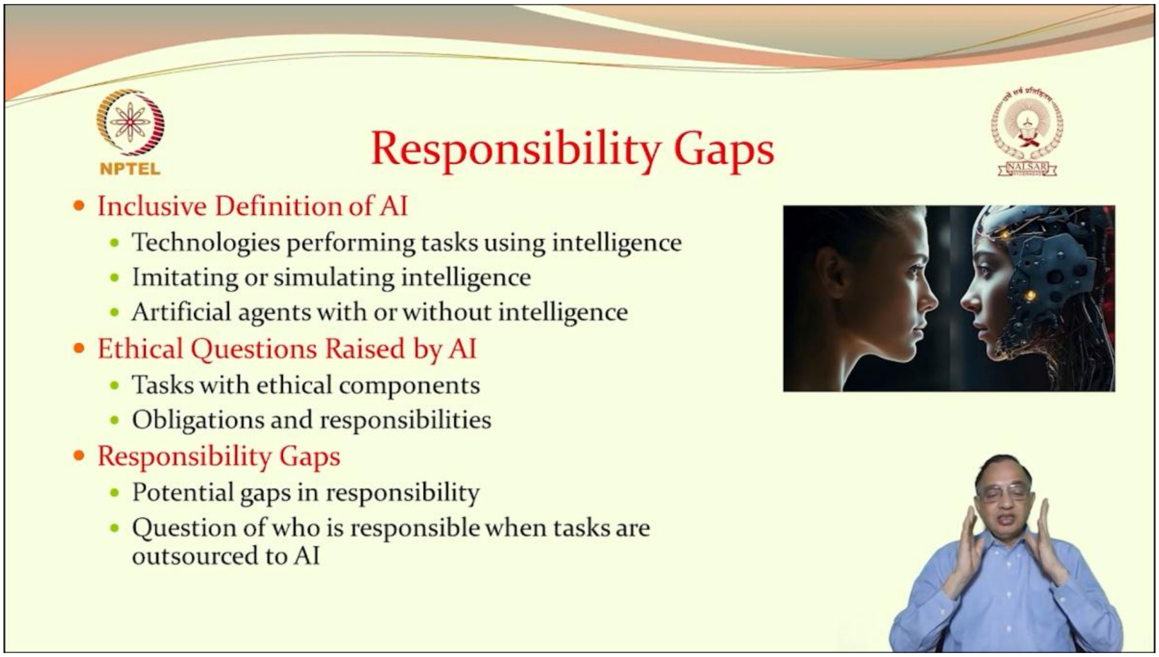
Again, we need to bring in well-being and human flourishing. See, human flourishing here is not something we are talking about in terms of some romantic, poetic, or ideal way. Human flourishing here means humans having the capacity to develop themselves, to contribute good to society, and human flourishing is used here in a very positive way so that society itself allows humans to flourish, recognize their talents, and permits the development of skills and capabilities. It also comes with some responsibilities when society gives them, but then human flourishing can be a key term in it. Whether AI helps us to live a meaningful life and have meaningful relationships is also important. When we talk of a broad conception of ethics, we will bring in living a meaningful life and having meaningful relationships. This meaningful life can be thought of as equivalent to, but not identical to, Article 21's right to life. Then consider what makes certain forms of work more meaningful than others. The reason we need to talk in terms of a broad conception of ethics is that AI, particularly when it comes to work, is going to have a serious impact on work, working relationships, jobs, employment, and the way society is going to reorganize or rethink the various categories of work. So, when we take a broad conception of ethics, we will be able to address the impacts of AI in a much better way and how we can direct AI to enhance human flourishing, contribute to human welfare, and contribute to societal welfare. So, this broad conception of ethics is something we really need to discuss when we talk about AI ethics.



Negative & Positive Approaches

- Narrow Perspective on AI Ethics
  - Focuses on permissible and impermissible uses of AI
  - Examples include debates on banning autonomous weapons
- Responsibility and Harm
  - Discusses who should be held responsible for AI-related problems
  - Commonly associated with negative impacts of AI
- Broader Perspective on AI Ethics
  - Considers AI's impact on human well-being and flourishing
  - Reflects on meaningful lives, work, and relationships
- Positive and Negative Ethical Questions
  - Negative: Focus on harms, injustices, and responsibility
- Importance of a Broader Ethical Approach

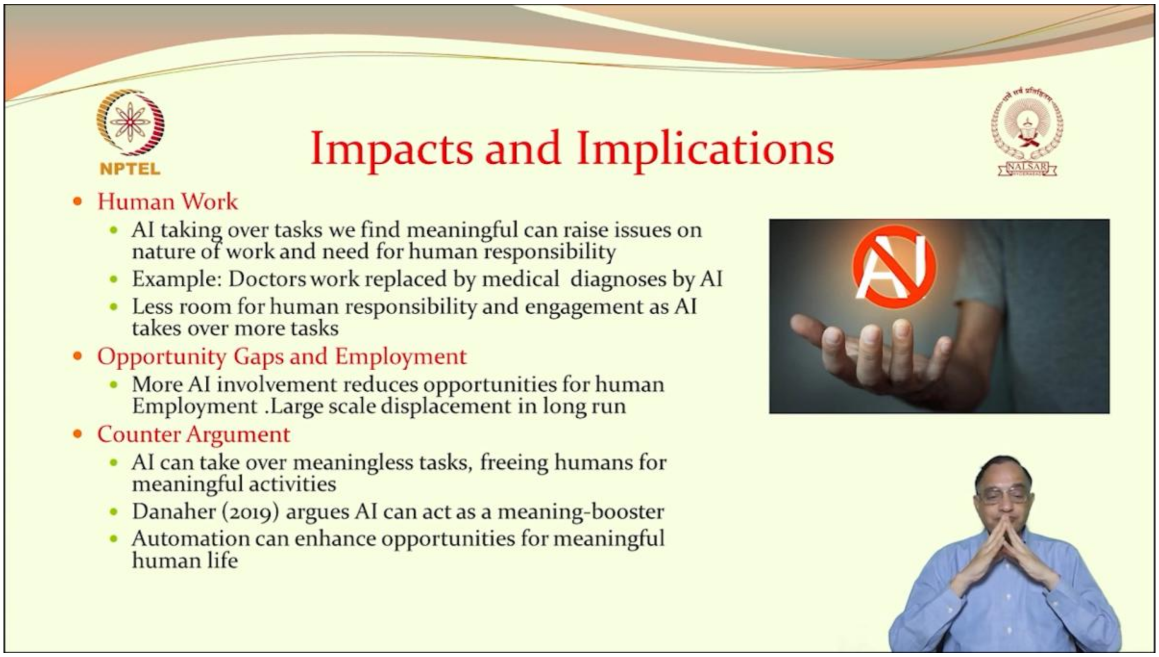
Naturally, from this, the next step would be negative and positive approaches. See the narrow perspective talks of only permissible and impermissible debates on banning autonomous vehicles, and then you go to the next step - the question comes about who should be held liable for the harms done by AI, and then who is really liable. This is again commonly associated with negative impacts. It is also possible that we can address some of these issues from the perspective of tort law, but not all can be addressed satisfactorily by tort law. So, the broad perspective considers AI's impact on the wider human being, flourishing, societal development, and also meaningful life relationships. So, the positive and negative ethical questions arise in different ways. Negative looks more like forms, injustices, and responsibilities, whereas the positive one looks both at the positive contributions that AI could make and then at how AI can also result in constraining or inhibiting some of the potential benefits that could arise from AI on account of the policies we have. So we need to look into a broader ethical approach so that we come out of this dichotomy of positive versus negative, or positive on one side and negative on the other side, and then try to balance, stating this is not good, this is permissible, things like that; whereas the broader ethical approach will look into the larger questions, such as under what circumstances it is ethical to use AI for job displacement or to move people up in the value chain or upgrade themselves in terms of skill. Similarly, a broader ethical approach will also help us understand some of the emerging implications of the convergence of AI with other technologies like biotechnology, synthetic biology, nanotechnology, and others. It is all the more important to have a broader ethical perspective because, as of now, we are not sure what potential ethical questions AI might raise in the future. Right now, we are grappling with, or we are struggling with, the ethical questions posed by AI today. But only a broader ethical approach will help us guide, understand, and then navigate the ethical challenges that could arise in the future on account of the rapid development of AI and its capacities.



Responsibility Gaps

- Inclusive Definition of AI
  - Technologies performing tasks using intelligence
  - Imitating or simulating intelligence
  - Artificial agents with or without intelligence
- Ethical Questions Raised by AI
  - Tasks with ethical components
  - Obligations and responsibilities
- Responsibility Gaps
  - Potential gaps in responsibility
  - Question of who is responsible when tasks are outsourced to AI

Having said that, we should understand two things. There are some serious responsibility gaps that we need to address too. See, technology performing tasks using intelligence is what AI firmly does. So, AI is either simulating or imitating human intelligence. But when it comes to artificial agents, they may not be fully intelligent, or they may not be intelligent, or they may possess intelligence. As we saw in the earlier class, some of the chatbots may not come with any intelligence or any capacity built in for "intelligence." They may be doing some mechanical jobs; they may be replying to some mechanical queries. But it's also possible to develop artificial agents with intelligence, with the capacity to learn, the capacity to respond, and the capacity to identify and then learn from the environment. So, when those things come, the question of responsibility inevitably arises. So ethical questions are tasked with ethical components, obligations, responsibilities, and the potential gaps in responsibility, which also raises the question of who is responsible when tasks are outsourced to AI, especially if society decides that tomorrow AI agents will take care of many things that normal people would do. I outsource the question of booking my ticket to AI, which is something that a normal person does, so that is not something we are talking about. But when the government itself outsources some of its functions, like bail bonds or some functions that people have to face on a day-to-day basis with AI agents, such as customer service or queries, or going to the hospital and then getting registered or obtaining the first level of consultation with an AI agent, a lot of responsibility gaps and ethical issues arise; so these things have to be thought of, contextualized, and understood rather than merely addressed as ethical issues that will always be present because AI is going to raise them.



NPTEL

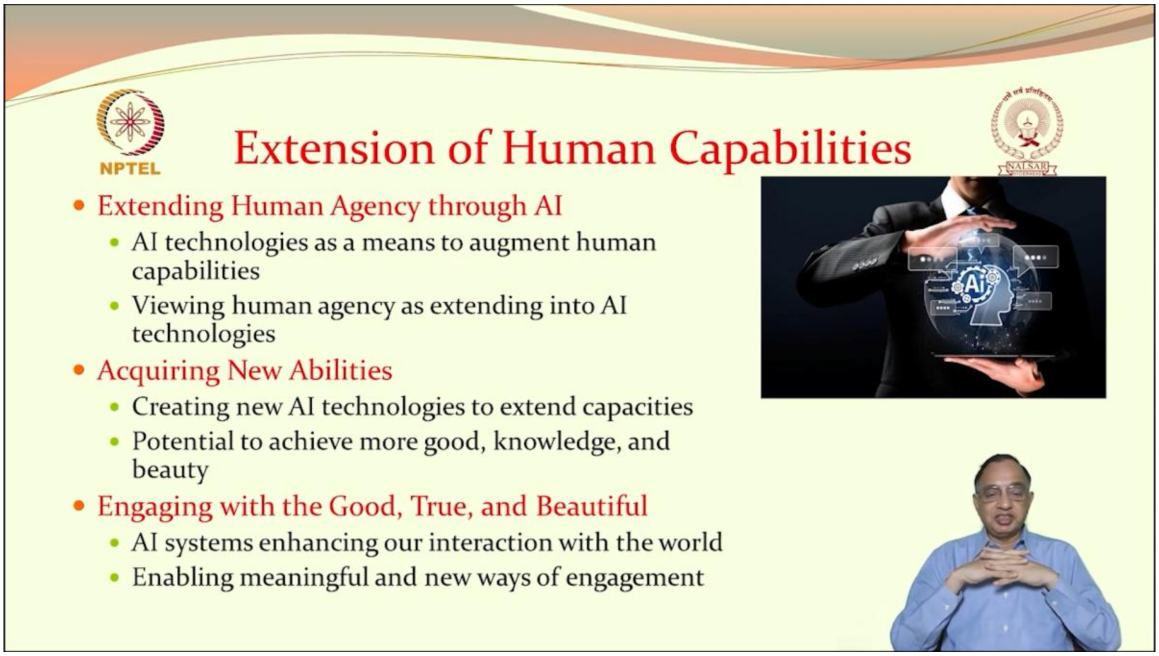
Impacts and Implications

- Human Work
  - AI taking over tasks we find meaningful can raise issues on nature of work and need for human responsibility
  - Example: Doctors work replaced by medical diagnoses by AI
  - Less room for human responsibility and engagement as AI takes over more tasks
- Opportunity Gaps and Employment
  - More AI involvement reduces opportunities for human Employment .Large scale displacement in long run
- Counter Argument
  - AI can take over meaningless tasks, freeing humans for meaningful activities
  - Danaher (2019) argues AI can act as a meaning-booster
  - Automation can enhance opportunities for meaningful human life

So, impact and implications are very important. Let us start with something very fundamental. If AI takes over tasks that we find meaningful and that we do to earn our livelihood and for dignity, it could raise the issue of what exactly AI would impact. AI can take away a lot of jobs that people would need anyway, and at another level, if we are going to replace the doctor's work with medical diagnosis by AI and AI systems, how does it matter, some people may ask. But the problem is, can we replace doctors and their wisdom, experience, and expertise with AI? If so, even if it is desirable, up to what level, and then where do we bring the doctor into the loop? The third question is that if some tasks are to be entirely left to AI and then AI agents or AI systems, where is the room for human responsibility and human engagement? If AI takes over more and more jobs, it could result in a situation where humans are either made less responsible or become just another cog in the wheel or another small component of a larger system. So, the question of whether humans are in the loop, human above the loop, or where exactly humans can come in for decision-making raises a lot of important questions. And then, more AI involvement reduces opportunities for human employment; large-scale displacement in the long run is possible, and this has been a very contentious topic. But we should look at employment not just from a mere economic perspective, because employment also gives people a way to keep themselves occupied, do some meaningful work, contribute to society, and engage with their own skills and expertise. Employment cannot be reduced to the mere economic factor of earning a certain amount of money per month or per year for putting in a certain amount of work in terms of hours or days.

Employment has a moral component and a moral dimension; it has a livelihood aspect, but it goes beyond livelihood because people work not just for money. People work because they find it somewhat satisfying; it enables them to earn money, but it also gives them dignity in society, provides them with responsibility, and, above all, assures them that their self-respect is being recognized and accepted by society. So, beyond employment as a mere economic opportunity, we should look at jobs and employment in a broader way. The counterargument is that AI can really take over meaningless jobs that humans do repeatedly, and then they get really bored or become totally alienated. This will mean that humans can be freed up to do much better, more meaningful jobs.

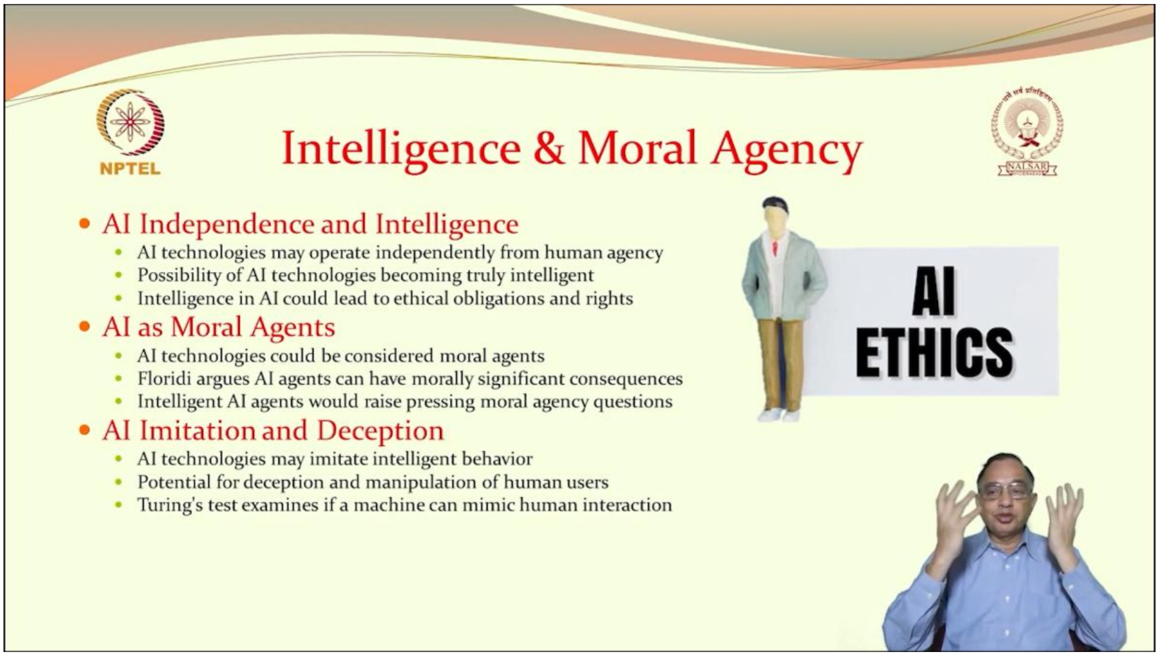
Danaher, in his 2019 book, argues that AI can, in fact, act as a meaning booster, and this is something we cannot really say is totally unacceptable; it is acceptable. But then how do you use AI to boost meaning, and under what circumstances also make a huge difference? Then it's also possible that automation can enhance opportunities for meaningful human life by ensuring that automation keeps a lot of things away from people, in the sense of mundane, boring routine jobs that people do just to earn a livelihood, so that they can be taken away from them, and then they can be given jobs that are much more suitable to their qualifications and expertise. So, even in a matter relating to employment, there are a whole lot of issues that need to be addressed. Of course, unless we address them through the prism of ethics and then responsibility, we may not even know what questions we should ask.



Extension of Human Capabilities

- Extending Human Agency through AI
  - AI technologies as a means to augment human capabilities
  - Viewing human agency as extending into AI technologies
- Acquiring New Abilities
  - Creating new AI technologies to extend capacities
  - Potential to achieve more good, knowledge, and beauty
- Engaging with the Good, True, and Beautiful
  - AI systems enhancing our interaction with the world
  - Enabling meaningful and new ways of engagement

Then comes the tricky, challenging, and interesting question of the extension of human capabilities. So, you know that there is a whole idea called human enhancement, post-humans, humans caring with super capacity, etc. Now extending human agency through AI is also possible in the sense that AI can enhance our cognitive capacity. AI combined with neuro technologies can enhance our cognitive capabilities. So, if AI is going to really extend our capacities and human agency, is it ethical to allow it in all circumstances? what exactly are we talking about here? Is it that AI is extending our human agency, or is AI helping us to extend our agency, and then who is the major player here? Are we depending on technology to extend our agency, or do we choose technology, and then it helps us extend our agency, leaving the option of whether to extend our agency or not to us? Then AI can really help us acquire new capabilities, whether it's learning a language, acquiring a new skill, or anything else that is possible. So, it's possible that in this way, humans can become more productive, more dynamic, more actively engaged, more knowledgeable, and can have a better sense of aesthetics and a better sense of life. So, when AI systems enhance our interaction with the world, they can do a lot of positive things, in the sense that they can make us more artistically inclined, stimulate our artistic thinking, train us to be more artistically sensitive to certain things, and help us appreciate and co-create. So, AI can create meaningful and new ways of engagement not just within us or with our own colleagues, fellows, and families, but also with the external world, including nature. So, in that sense, the extension of human capabilities through AI per se can be seen as a very positive, very welcome development. But of course, there are some ethical issues that cannot be wished away.



Intelligence & Moral Agency

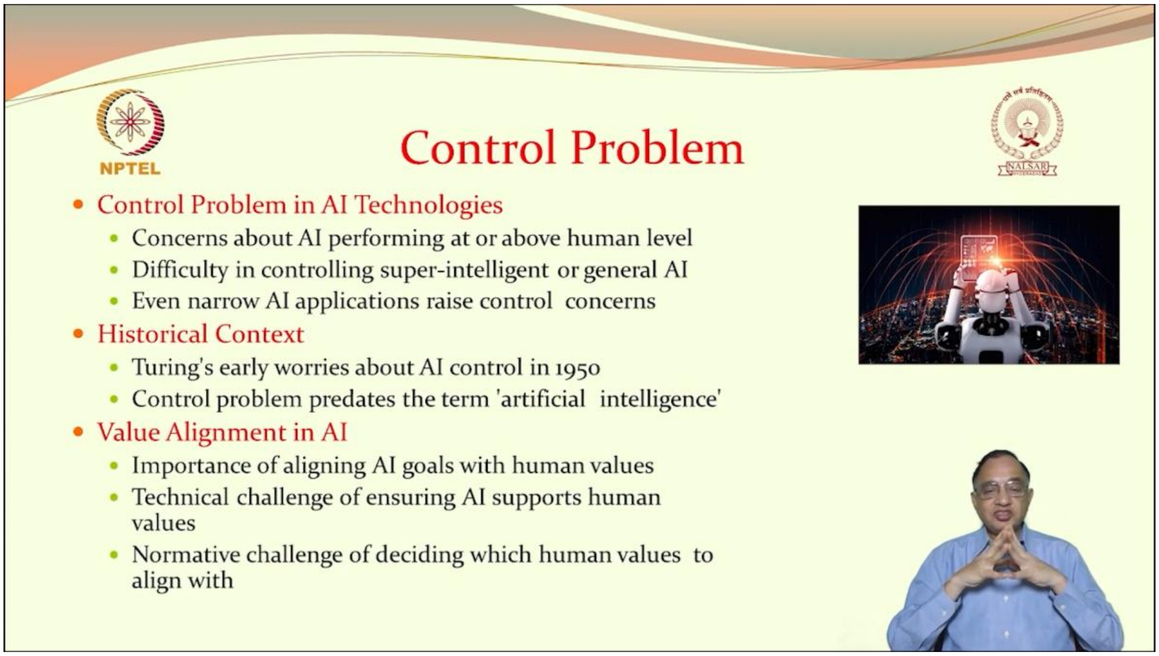
- AI Independence and Intelligence
  - AI technologies may operate independently from human agency
  - Possibility of AI technologies becoming truly intelligent
  - Intelligence in AI could lead to ethical obligations and rights
- AI as Moral Agents
  - AI technologies could be considered moral agents
  - Floridi argues AI agents can have morally significant consequences
  - Intelligent AI agents would raise pressing moral agency questions
- AI Imitation and Deception
  - AI technologies may imitate intelligent behavior
  - Potential for deception and manipulation of human users
  - Turing's test examines if a machine can mimic human interaction

Then what happens to the questions of intelligence and moral agency? If AI on its own operates from human agency, the first question is, who will be responsible? The second question is whether, if they become truly intelligent, it should also be linked to some ethical obligations and rights. In the sense that, can we say that an AI system will have rights? Can we say that robots will have rights? Of course, they will also have some sort of responsibility. They will also have some sort of accountability to us. So, if we are going to concede that AI or artificial intelligence systems are becoming truly intelligent, then they are also ethically obligated. If they are ethically obligated, where do we stop or where do we draw the line between human intelligence and artificial intelligence? Are we going to look as if they will always act as humans do, or are we going to say that we will give them, we will assign them some special rights, but we will not consider them equivalent to human intelligence? In that sense, their ethical concerns and responsibility will be something that is lesser than that of human but more than that of an animal. Or when it comes to some robots or AI systems, should we again differentiate robots technically as something very different and AI systems as something very different? So how do we assess these things? Can we really categorize and label them as robots, or are they better than AI systems but closer to humans? AI systems are better than animals but are moving towards robots.

So, this sort of classification issue will always arise, even if we say that AIs have intelligence, but that itself can be questioned and challenged. Then we also discussed earlier the question of AI acting on its own and being moral agents. AI technology can be considered moral agents. When we say moral agents, we mean that they have a sense of morality, a sense of agency, and the power to act on their own. According to Floridi, he is one of the important philosophers of digital ethics, AI ethics, and is generally considered one of the top-ranking philosophers who have been working consistently on it for more than two decades. So, he says that AI agents can have morally significant consequences because the moment we assume that AI agents have responsibility, and then the moment we say that they are also moral agents in the sense that they have a sense of morality, they are bound by certain norms and values; automatically, their actions can have morally significant consequences. So intelligent AI agents, when deployed in large numbers in systems or as part of a larger ecosystem, will really raise moral agency questions. Now, the moral agency question comes not from an abstract idea but from a practical idea. For example, an intelligent AI agent working in a health system makes a mistake or commits a mistake that results in some harm. So is the AI agent really capable, is the first question to put up.

It is really capable, but then it did something wrong. So, the question of responsibility and accountability is one. But can we really attribute moral agency to that and then prove that it wilfully did it or did it knowingly so that the harm would happen, or are we saying that it was cognizant of what it was doing and that it had the mental idea that it was going to do some harm? So, this, again, is not the question of talking; we are not talking in terms of science fiction, rogue robots or those types of things. We are talking about very real things when it comes to moral AI agents and intelligent AI agents if deployed across different sectors. Then they can also imitate intelligent behaviour, but the imitation may never be the actual behaviour. It could be something that is imitative. So, the potential for deception and manipulation of human resources is present. For example, using an AI agent, I can try to deceive people; I can try to deceive and manipulate humans' emotions, humans' sentiments, and humans' understanding. I can try to make the AI agents act in a very friendly, very casual way and then try to steal some very sensitive personal information or try to do some harm without the other person knowing that he or she is being harmed. It's also possible that AI agents and AI technology per se can be used for many deceptions and manipulations. Of course, deepfakes are only one category of it.

There are many such technical possibilities. So, we need not always think that when we talk about AI imitation, deception, and all that, it is just a question of deepfakes. It is not. There are a whole lot of things; other things can happen. For example, an AI agent may squander itself or can try to mislead a person, say, by stating that it represents some other organizations, whereas in fact it wouldn't be so; the potential is huge. Then, if we go back to the Turing's test of whether machines are intelligent, if a machine can mimic human interaction, we can say that machines have intelligence. But what happens if AI agents are also able to mimic human interaction and then convince a human that the person on the other end is also a human? In the sense that what happens if an AI agent mimics a human and then convinces you that you are talking not to an agent or a system but to a human on the other end, these are the questions that arise when we talk of intelligence and moral agency.



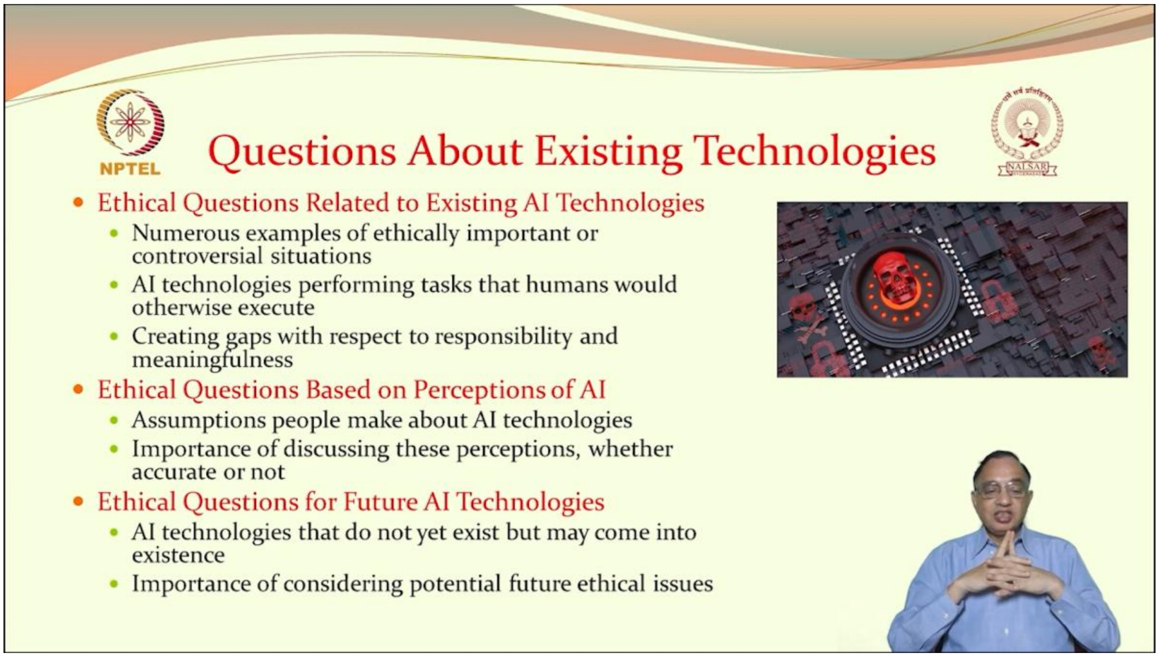
NPTEL

Control Problem

- Control Problem in AI Technologies
  - Concerns about AI performing at or above human level
  - Difficulty in controlling super-intelligent or general AI
  - Even narrow AI applications raise control concerns
- Historical Context
  - Turing's early worries about AI control in 1950
  - Control problem predates the term 'artificial intelligence'
- Value Alignment in AI
  - Importance of aligning AI goals with human values
  - Technical challenge of ensuring AI supports human values
  - Normative challenge of deciding which human values to align with

Of course, all this comes with the big issue of control as to where the control lies or where the bug stops. Concerns about AI performing at a human level have always been expressed. In fact, whether it is Isaac Asimov's Robo rules or any of the things that are particular to discussions on AI ethics in the context of moral agents or AI ethics doing rogue things, we always question the control problem as to who controls; can the technology or the AI agent come with a control mechanism that can also be inactive or active depending upon the moral choice? Then controlling superintelligent or generative AI is again an issue because we have repeatedly seen the black box problem, but that is only one aspect. The other aspect is the moral responsibility. And then even narrow AI applications raise control concerns as to whether humans are in control, whether humans are part of the loop, or whether humans are somewhere but have no control over it, or whether humans can intervene only in some circumstances where AI has made a decision, but only to mitigate the harm. So historically, Turing rarely worried about AI control; 1950 is there. Then Norbert Wiener wrote extensively about some of these issues in his book. Then came Joseph Weizenbaum, who also published his book Computer Power in the Age of Human Reason. And then he also developed an earlier chatbot called ELISA, which his secretary believed was equivalent to a human being and was misled.

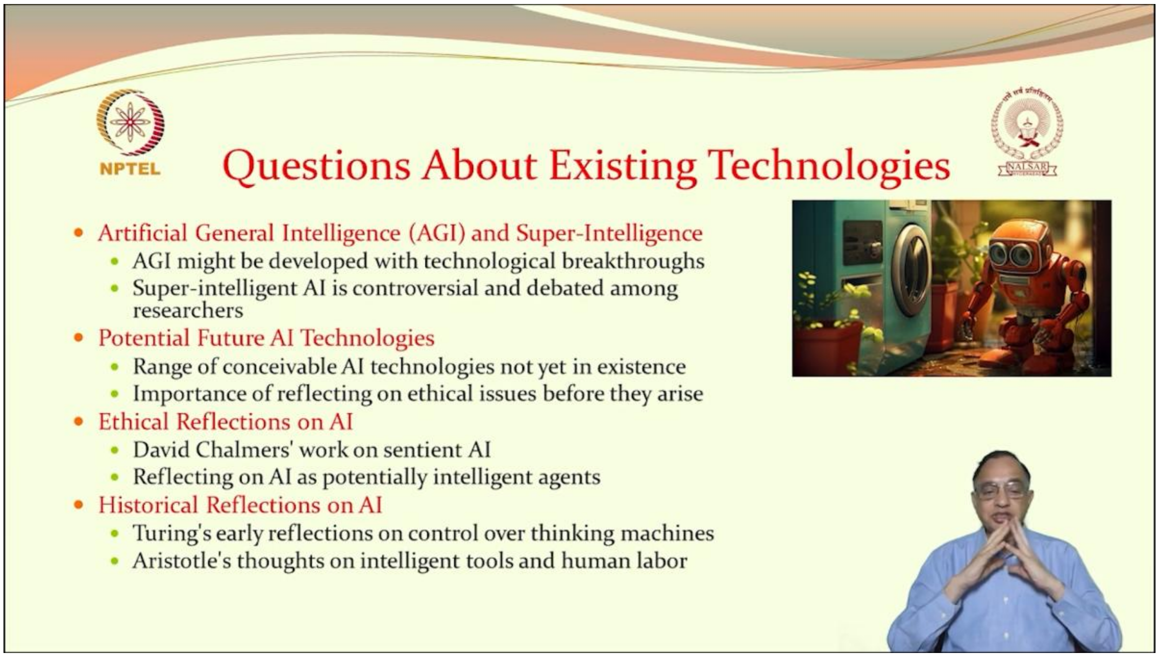
So, if you look at it, right from the 1950s onwards, right from the days of Turing, right from the beginning of discussions on artificial intelligence, some of these ethical questions have been raised again and again, sometimes in a very narrow, specific way, sometimes in a broader way. Questions like those raised by Joseph Weissenbaum and Norbert Peiner brought up much broader questions and issues in the sense that they did not think in terms of mere AI. They thought broadly in terms of how society should interact and how the relationship between society and technology should be guided or governed. So, some of these concerns keep repeatedly appearing not because they were raised earlier but because they are becoming more pertinent now. Then the question of aligning AI with human values has a lot to do with the whole idea of responsible AI as well. So, the technical and ethical challenge here is how we ensure that AI systems respect and support human values. What are human values? It could be transparency; it could be not being a black box; it could be what we see in the case of AI ethics discussions. We will look at some of the things where we will specifically examine some of the specific values as well. So, the normative challenge of deciding which human values to align with is also present. Again, this is a problem because there is no accepted consensus on what exactly human values are and which one among them should be prioritized. And then, depending on the school of philosophy, depending on the school of ethics and other factors, these things vary.



Questions About Existing Technologies

- Ethical Questions Related to Existing AI Technologies
  - Numerous examples of ethically important or controversial situations
  - AI technologies performing tasks that humans would otherwise execute
  - Creating gaps with respect to responsibility and meaningfulness
- Ethical Questions Based on Perceptions of AI
  - Assumptions people make about AI technologies
  - Importance of discussing these perceptions, whether accurate or not
- Ethical Questions for Future AI Technologies
  - AI technologies that do not yet exist but may come into existence
  - Importance of considering potential future ethical issues

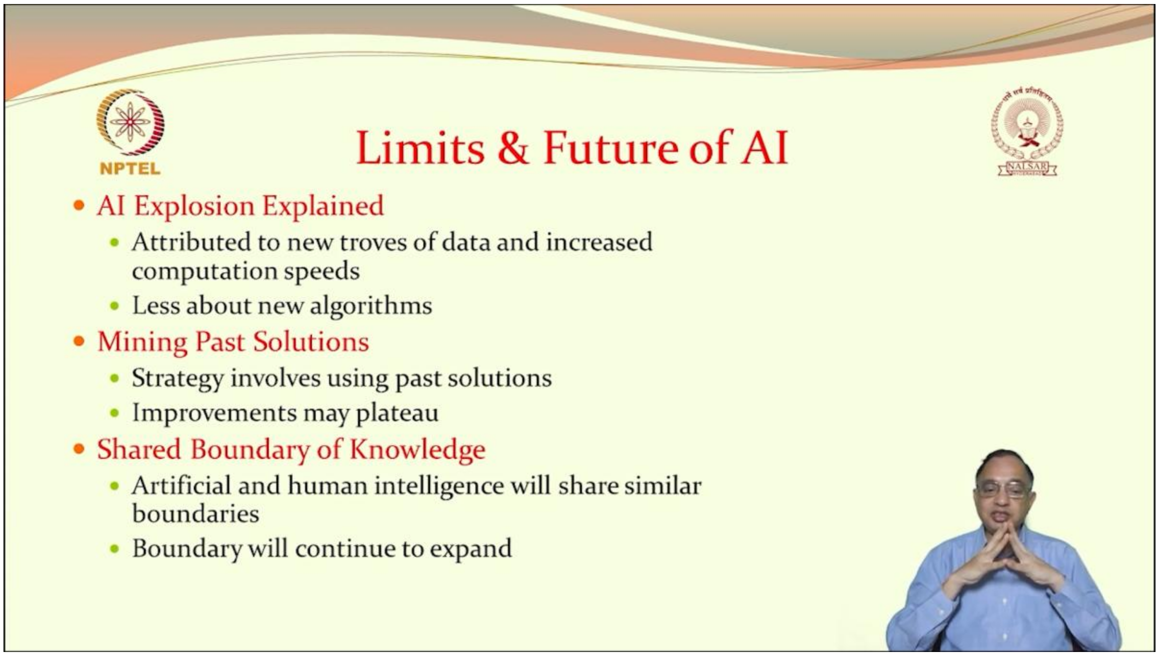
Now, there are a whole lot of existing technologies out there. So, the ethical questions are also raised in them. But in the case of AI, there are a lot of things that are very specific because we have a lot of assumptions, and discussing people's perceptions about AI, AI's capabilities, and AI ethics is also equally important. But we also need to have some foresight and think about what sort of technologies might evolve and what sort of ethical issues may arise in the future. So, it is also better to be prepared or at least cautious about these things so that we know these are some of the potential ethical issues that could arise in the future on account of the development of technology. So, we need to look at things here from two perspectives. One, there are a whole lot of ethical issues with the existing technologies, including AI. So, we need to learn a lot from those debates and then those discourses, say on biotechnology, human genome editing, synthetic biology, nanotechnology, etc. We also need to draw a lot of insights from them, but we should also look into some specific things that are very unique to AI. Because we are not saying that biotechnology on its own will have some of the problems of AI, whereas AI on its own will have the problem of moral agency and the responsibility that arises from moral agency.



Questions About Existing Technologies

- Artificial General Intelligence (AGI) and Super-Intelligence
  - AGI might be developed with technological breakthroughs
  - Super-intelligent AI is controversial and debated among researchers
- Potential Future AI Technologies
  - Range of conceivable AI technologies not yet in existence
  - Importance of reflecting on ethical issues before they arise
- Ethical Reflections on AI
  - David Chalmers' work on sentient AI
  - Reflecting on AI as potentially intelligent agents
- Historical Reflections on AI
  - Turing's early reflections on control over thinking machines
  - Aristotle's thoughts on intelligent tools and human labor

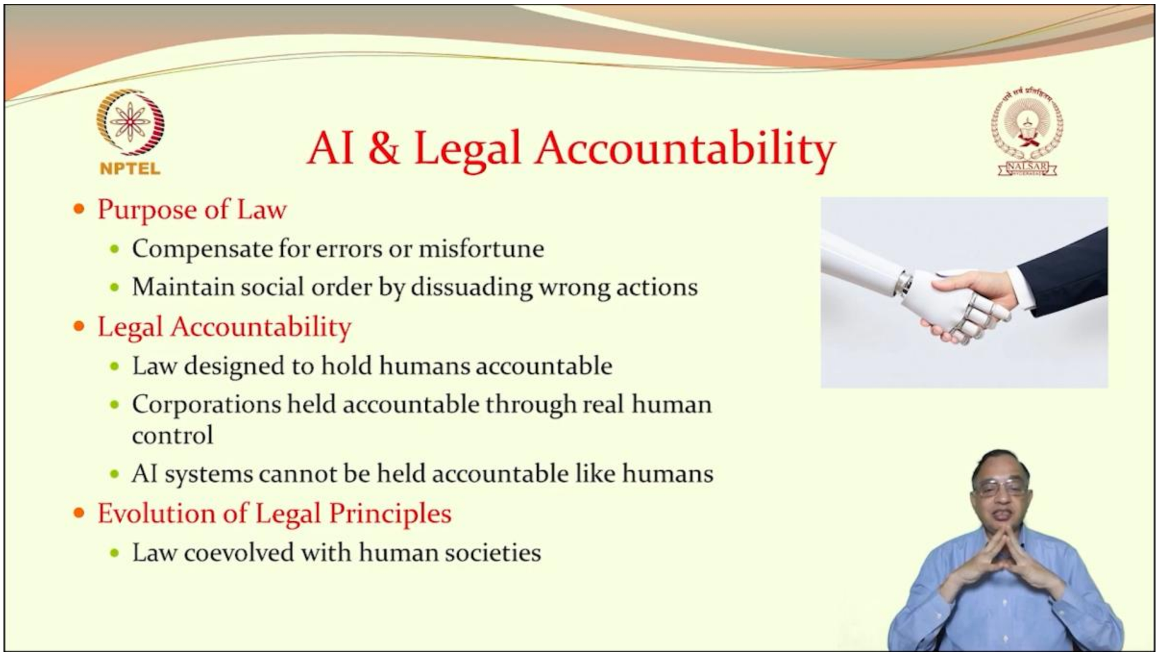
If we talk in terms of AGI, AGI might be developed with super technology. So, super-intelligent AI is a very controversial but somewhat challenging notion. And then we saw future AI. So, there are a lot of people who have worked on it. But one of the things that comes up again and again is what happens if AI becomes sentient beings. Now, if AI results in "sentient beings," should they be treated as beings that are equivalent to things with life, with a sense of morality, with a sense of mind, or should they be treated as yet another machine that is little more than a machine, but not as good as human beings or anything that is alive? So, David Chalmers is one philosopher who has worked extensively on sentient AI. So, it is also possible that AI may evolve into a potentially intelligent agent. As we have seen, Turing's earlier reflection was present, but if you go back, Aristotle's thoughts on intelligent tools and human labour are also important. So, starting from the Greek days of Aristotle, there has also been a lot of discussion on technology and philosophy, technology and ethics, and also on how society should deal with technology when ethical issues arise. So, there are a lot of people who have worked on it; for example, Heidegger has written a lot about it, and then coming to our times, we have at least a dozen or more philosophers who have devoted themselves to addressing some of the issues related to emerging technologies. So, whether it is existing technologies or emerging technologies like AI, there is a whole lot of philosophy of technology, and ethicists, philosophers, and other social scientists are trying to address them from various perspectives. When we talk of AI ethics, we need to draw from the insights of their work.



Limits & Future of AI

- AI Explosion Explained
  - Attributed to new troves of data and increased computation speeds
  - Less about new algorithms
- Mining Past Solutions
  - Strategy involves using past solutions
  - Improvements may plateau
- Shared Boundary of Knowledge
  - Artificial and human intelligence will share similar boundaries
  - Boundary will continue to expand

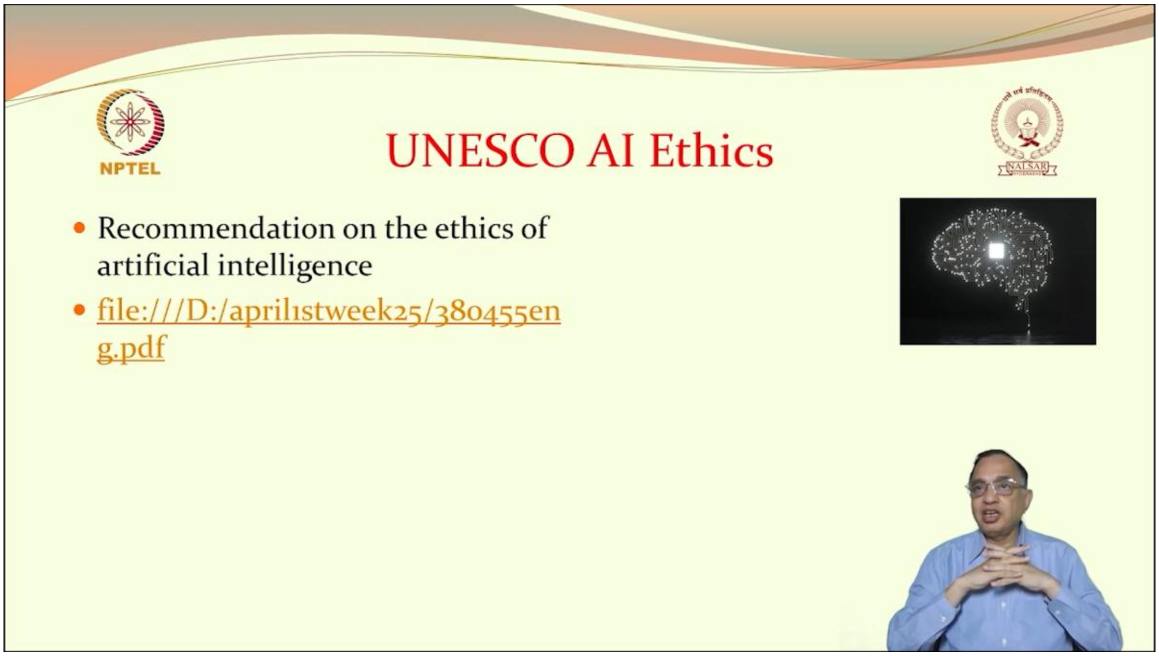
And we also need to take into account the future of AI, so we can do one thing: we can look at past solutions and then identify where some of them have worked and some of them have not when we discuss ethics in technology. But it is also possible that some of the things we talk about today as superintelligent AI or AGI may be high in the sense that there is no guarantee AI will continue to evolve and progress at this level. There could be technical issues, and there could be other issues that at some point AI may plateau or may not go further for one reason or another. We cannot predict that, though. So, the future of AI is something we really need to look into, and then what happens to artificial human intelligence, which shares similar boundaries? So, if you are talking about human intelligence as a paradigm and then compare that with artificial intelligence, that is one way to do it. Otherwise, you look at intelligence and then say, okay, we know we are talking about intelligence. Since we are talking about intelligence, we will first not bring in humans as the paradigm. We will talk in terms of what exactly intelligence is. Then we will independently assess AI as well as humans on those parameters rather than taking humans as the primary one to assess AI's intelligence. So, there are many ways to address these questions. Then the boundary between human intelligence and artificial intelligence, whether they will merge when AI systems merge with some of the controls we have or when AI systems are integrated into our thinking and decision-making in the sense that if we all become cyborgs who also operate with AI systems, will make the question totally different. So, these are some of the things with which AI ethics comes to grapple.



AI & Legal Accountability

- Purpose of Law
  - Compensate for errors or misfortune
  - Maintain social order by dissuading wrong actions
- Legal Accountability
  - Law designed to hold humans accountable
  - Corporations held accountable through real human control
  - AI systems cannot be held accountable like humans
- Evolution of Legal Principles
  - Law coevolved with human societies

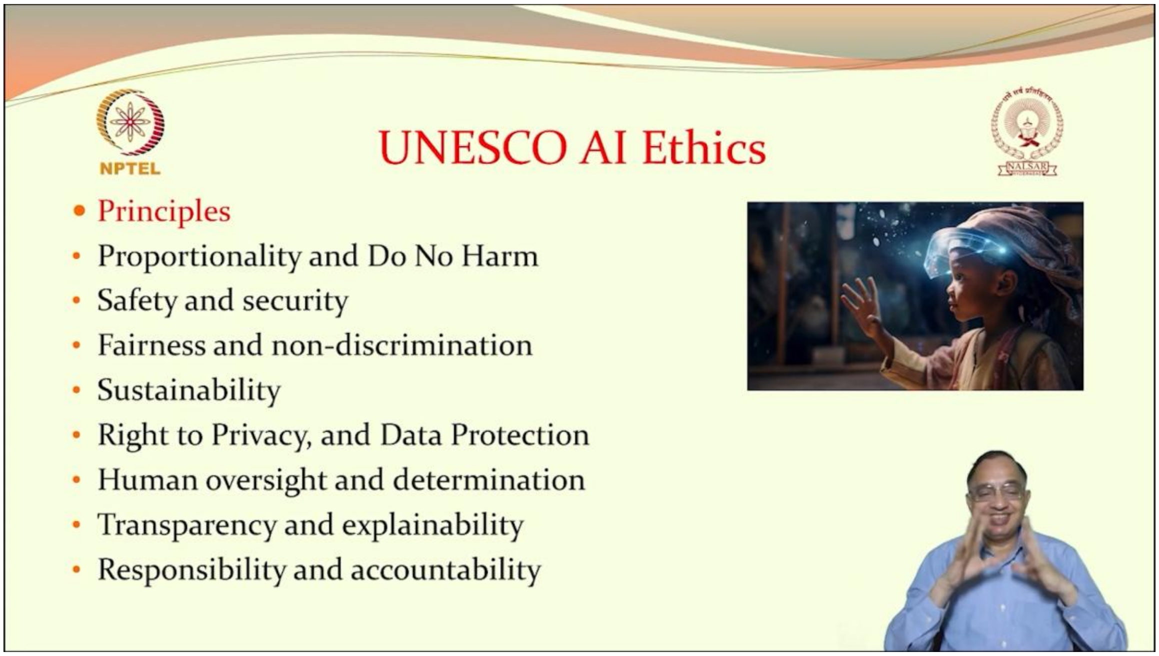
In particular, when it comes to ethics in law, we need to know the purpose of law is to compensate for errors and misfortune and then ensure that wrongful actions are not continuously done or are not disincentivized or decided because of the social order. What would happen if AI or any technology created massive havoc or caused a lot of harm in the sense that, suppose you have autonomous driving vehicles, and then if they caused at least 100 accidents here and there in the city on account of their behaviour, then it could be something of a social order problem that also needs to be addressed. Of course, the law has been designed to hold humans accountable. When it comes to animals, we don't hold them accountable, but the owners. So, the whole lot of tort jurisprudence stems from some of these ideas. Similarly, corporations can be held accountable for human rights violations, although they are recognized as artificial persons or juristic persons. So far, we have not been able to extend that rationale to AI systems and say that they are accountable like humans. But sooner or later, this question has to be grappled with. So, law on one hand evolved with human capabilities, human services, human societies, human aspirations, and human wishes. But when law has to deal with AI, this can be a necessary starting point, taking humans based on our experiences in the past and then assessing how law and AI have to be really looked into. But that alone is not sufficient because we cannot extend by nature or by default whatever humans due to AI and vice versa.



UNESCO AI Ethics

- Recommendation on the ethics of artificial intelligence
- file:///D:/april1stweek25/380455eng.pdf

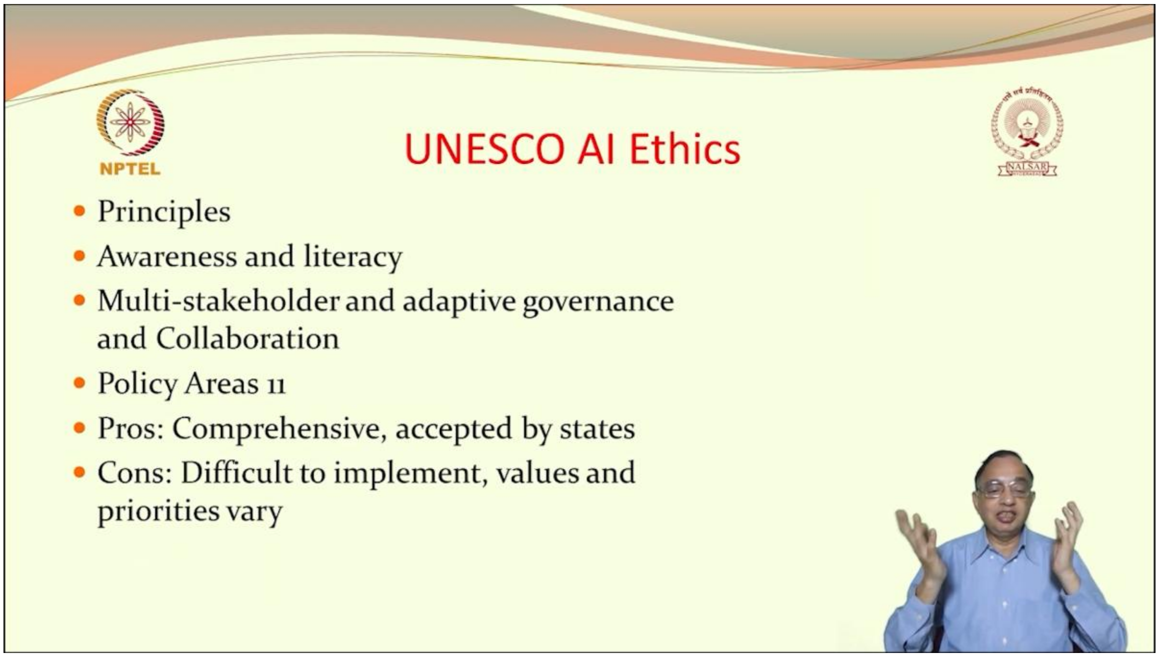
So, having discussed all these, we should now get into something that makes a difference. UNESCO's AI ethics, or UNESCO's Recommendation on the Ethics of Artificial Intelligence, is the first, or perhaps what I would call the most important, ethical principle described and accepted by all the countries in the world. UNESCO started this work much earlier for the simple reason that it has been working on ethical issues relating to technologies and science for many years. So, in 2019, they came up with these UNESCO AI ethics principles, and then they further developed them with guidelines. They also did a lot of work on connecting them to the concept of legal law principles and then the rule of law. So, we need to look at AI ethics because this is something of an issue that everyone has accepted, but as we see, it is not something that everyone has accepted. So, the problems have not been solved.



UNESCO AI Ethics

- Principles
  - Proportionality and Do No Harm
  - Safety and security
  - Fairness and non-discrimination
  - Sustainability
  - Right to Privacy, and Data Protection
  - Human oversight and determination
  - Transparency and explainability
  - Responsibility and accountability

Broadly, UNESCO AI ethics has these principles: proportionality and do no harm, safety and security, fairness and non-discrimination, sustainability, the right to privacy and data protection, human oversight and determination, transparency and explainability, responsibility and accountability.

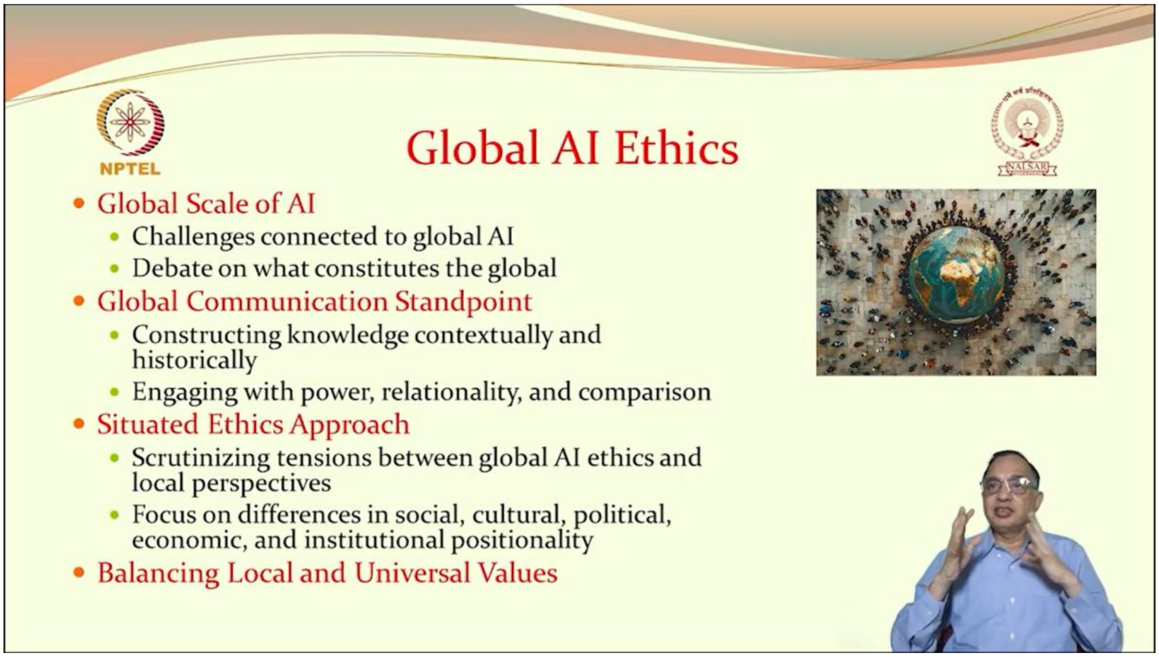


UNESCO AI Ethics

- Principles
- Awareness and literacy
- Multi-stakeholder and adaptive governance and Collaboration
- Policy Areas 11
- Pros: Comprehensive, accepted by states
- Cons: Difficult to implement, values and priorities vary

Awareness and literacy, and multi-stakeholder and adaptive governance and collaboration. So based on this, UNESCO's AI ethics has been identified in 11 policy areas where UNESCO has come up with ideas and suggestions so that the AI ethics can be translated into actionable points and an actionable agenda, or can be integrated, or can be made part of the national ethics frameworks as well as the governance frameworks for

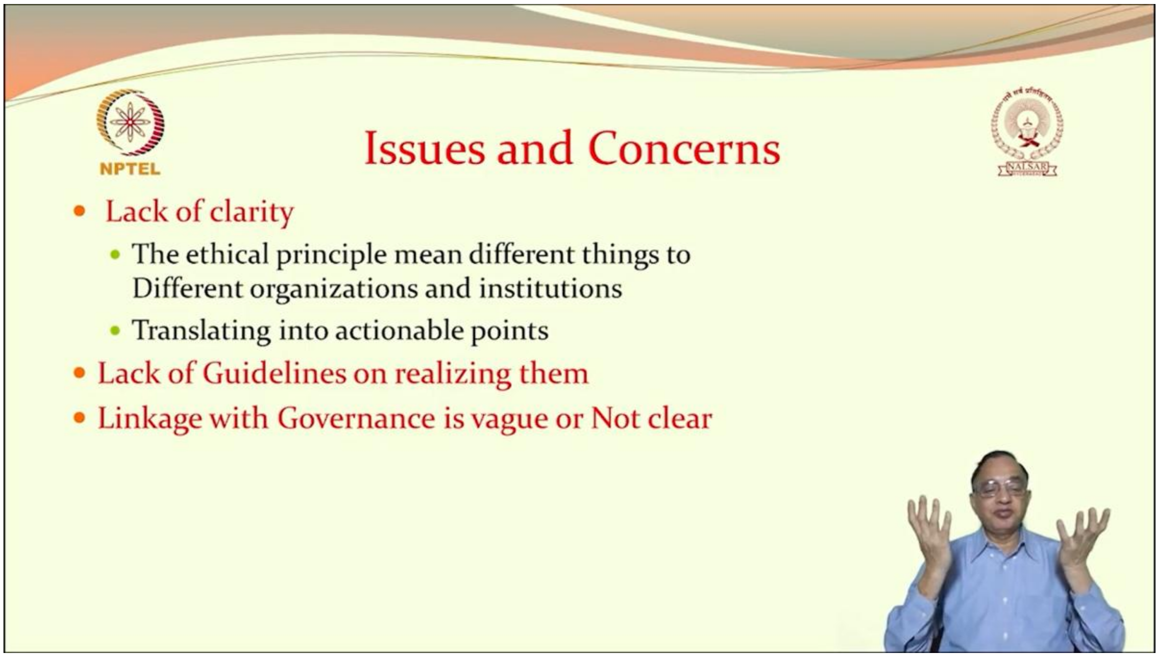
AI. But there are things that we need to understand about UNESCO's AI ethics. First of all, it is a soft-law instrument. It does not have the backing of a hardcore AI treaty, AI convention, or AI agreement. As a soft law instrument, it is not binding on the countries to abide by it or to implement it as prescribed by UNESCO at all times. And being a soft law instrument, one cannot go to court or sue UNESCO if someone has violated this, nor can one go to court and say my country has ratified or accepted AI ethics but has not implemented it. And more importantly, it is very comprehensive; it's accepted by states, and then UNESCO is also working with all the nations to bring them into the policy arena to include them as part of the AI governance, particularly in law and justice. But the other side is that it is very difficult to implement, particularly when values and priorities vary. For example, when we say awareness and literacy, multi-stakeholder and adaptive governance, and collaboration, these may look very fine. But certain things, like transparency and privacy, are contested values. And then not all countries have very similar value systems for privacy, transparency, and accountability. Why? Because of the first reason that there are cultural differences. Second Reason, there are a lot of differences in the way the country's laws and policies, and then the overall governance frameworks, look at some of these things. In some countries, privacy may be a cherished value. But then that again will not help the person when the state itself violates a person's privacy. So how to translate privacy from an absolute value into something workable is a big question.



Global AI Ethics

- Global Scale of AI
  - Challenges connected to global AI
  - Debate on what constitutes the global
- Global Communication Standpoint
  - Constructing knowledge contextually and historically
  - Engaging with power, relationality, and comparison
- Situated Ethics Approach
  - Scrutinizing tensions between global AI ethics and local perspectives
  - Focus on differences in social, cultural, political, economic, and institutional positionality
- Balancing Local and Universal Values

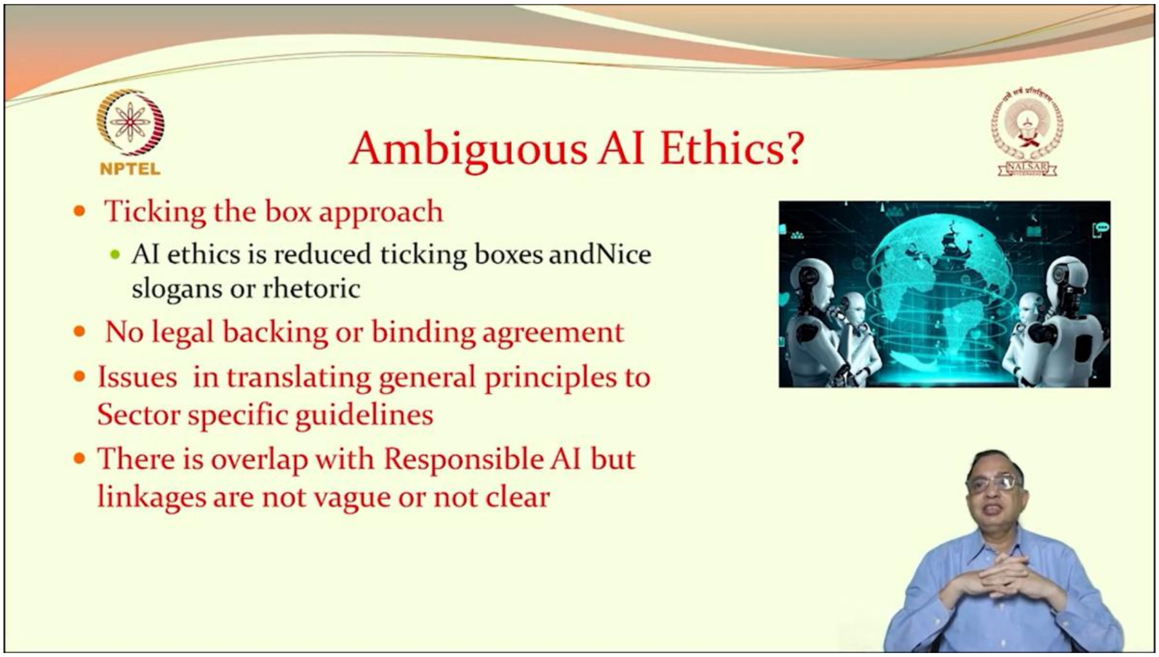
But if we talk about global AI ethics, the challenges are how to connect global AI. Debate on what constitutes global is also important because it is very easy to say global AI and global AI ethics. But what is global? Particularly when AI policy systems are at the national level. And then you need to always look at the contextuality and history rather than merely talking in abstract terms. More importantly, we need to examine the power relationships and rationality between the actors, as well as the compare. Furthermore, AI ethics can be a broadly global issue, but what exactly are the local perspectives? Do the locals in different countries and stakeholders entirely agree with UNESCO's AI ethics principles? Even if they agree, are their understandings the same and identical? That's the question. As there are huge differences among countries in social, cultural, political, institutional positionality, and values, talking about abstract AI ethics as enunciated by UNESCO or any agency and then claiming that this will always be applicable in all contexts will be a meaningless exercise because that will never happen; instead, countries will try to pick and choose where they can accept it in full and where they cannot accept it in full. They will try to adopt something but then negate it or mitigate it in one way, or try to ensure that what is stated there is taken into account but not implemented fully. So, the universal values, which could again be questioned, need to be contextualized and balanced; how to balance this is again a big question because who will decide what should be balanced and which value should be prioritized.



Issues and Concerns

- Lack of clarity
  - The ethical principle mean different things to Different organizations and institutions
  - Translating into actionable points
- Lack of Guidelines on realizing them
- Linkage with Governance is vague or Not clear

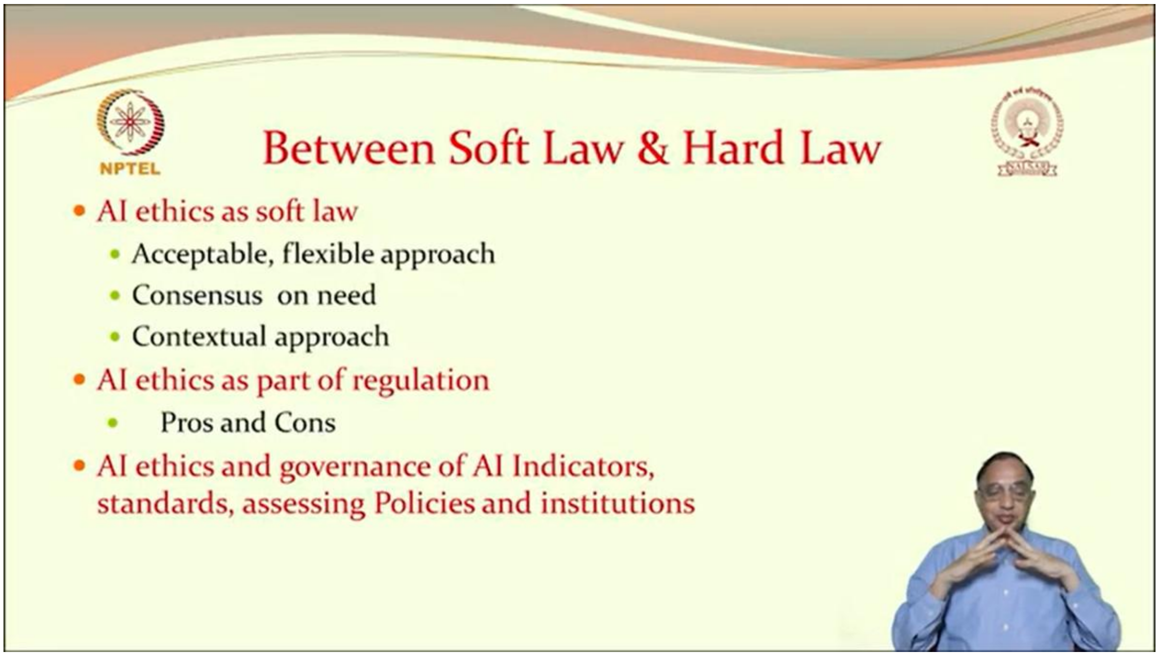
The questions here are a lack of clarity: ethical principles mean different things to different organizations, and then translating this into actionable points—for example, how will you translate this into ideas like responsible AI, explainable AI, or into drafting laws or governance principles and guidelines—there are not many guidelines in realising them. And then, although UNESCO's AI ethics principles look very nice on paper, how we link them with AI governance is not very clear. It is very vague. UNESCO, as an organization, is doing a lot of things related to AI ethics, AI governance, and AI and human rights. But many of the things that should happen at the national level are left to the national governments, where the understanding of that and the practice of that may be totally different from what UNESCO would have envisaged. So, the linkages with governance are vague or not clear. This does not mean that UNESCO's AI ethics are useless or that they are too vague or too difficult to implement; rather, the problem is translating them into actionable points and then putting them into specific laws and rules.



Ambiguous AI Ethics?

- Ticking the box approach
  - AI ethics is reduced ticking boxes andNice slogans or rhetoric
- No legal backing or binding agreement
- Issues in translating general principles to Sector specific guidelines
- There is overlap with Responsible AI but linkages are not vague or not clear

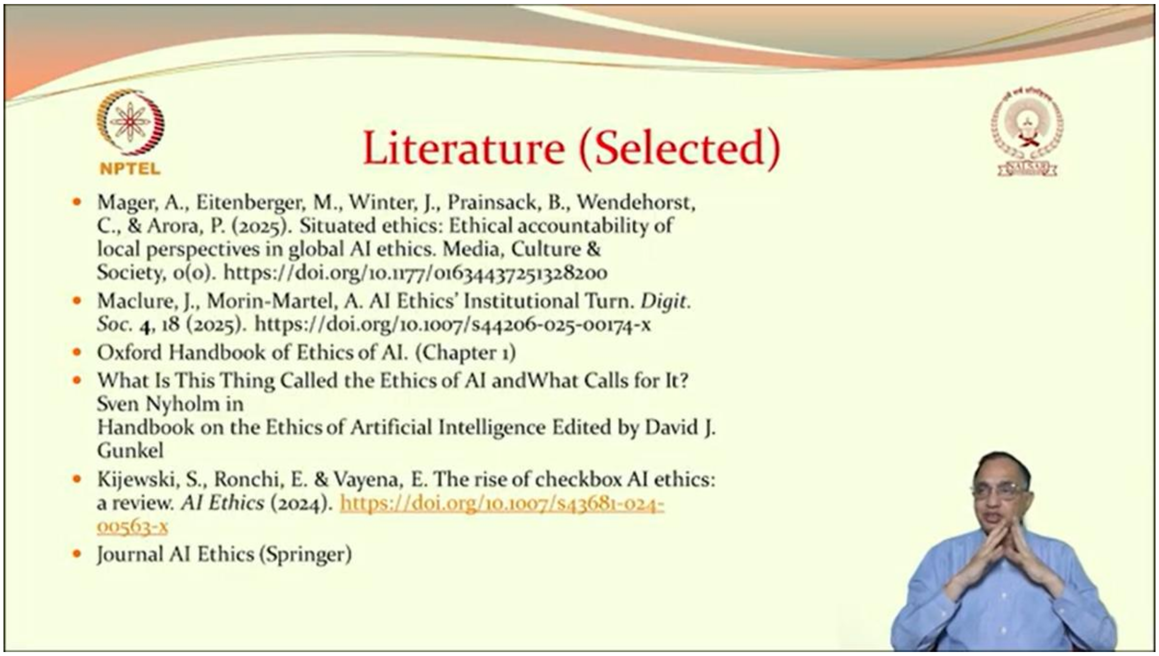
So, another point that has been discussed in the literature is that I have given the relevant literature at the end. Today, since there is a whole lot of discussion about AI ethics, there is a view that AI ethics has been reduced to just ticking the boxes when you ask questions and then say nice things or raise some nice slogans and rhetoric. Because there is no binding legal obligation for it, the question of translating that does not arise in the sense that, you know, you can simply tick boxes; you can simply say that we have been aware of the AI ethics principles we are trying to incorporate, but never work without doing anything. The overlap with responsible AI is there, but the linkage between them is not very clear or is quite vague. So, the ambiguity in AI ethics does not mean that AI ethics are not relevant. The ambiguity in AI ethics stems from the fact that AI ethics is either being reduced to ticking the boxes or not taken seriously enough to translate into actionable points and then get incorporated into various policies and frameworks.



Between Soft Law & Hard Law

- AI ethics as soft law
  - Acceptable, flexible approach
  - Consensus on need
  - Contextual approach
- AI ethics as part of regulation
  - Pros and Cons
- AI ethics and governance of AI Indicators, standards, assessing Policies and institutions

So, if it is part of soft law, it is acceptable. But if it is part of the regulation, then it becomes all the more difficult because of which one to choose and how to incorporate it, as I said, is a difficulty. For example, if it is a question of transparency in AI ethics, how do we translate that? Should it be part of the data protection regime alone, or should it be part of the broad AI governance framework? The EU AI Act is the closest one that addresses some of these things in a comprehensive way. But then, not all countries have comprehensive regulations, a comprehensive act, or a system to govern AI. So, the problem is that AI ethics, even if it is recognized, may not be fully included in the pool of regulatory frameworks or governance. It could be more like a guideline and may start and stop with it. Again, there is a serious lack of indicators, standards, and assessment policies and institutions. In the sense that there are no metrics to assess whether AI ethics have been translated by country or how we assess the various things that are claimed in the name of AI ethics.



Literature (Selected)

- Mager, A., Eitenberger, M., Winter, J., Prainsack, B., Wendehorst, C., & Arora, P. (2025). Situated ethics: Ethical accountability of local perspectives in global AI ethics. Media, Culture & Society, 0(0). https://doi.org/10.1177/01634437251328200
- Maclure, J., Morin-Martel, A. AI Ethics' Institutional Turn. *Digit. Soc.* **4**, 18 (2025). https://doi.org/10.1007/s44206-025-00174-x
- Oxford Handbook of Ethics of AI. (Chapter 1)
- What Is This Thing Called the Ethics of AI andWhat Calls for It?
Sven Nyholm in
Handbook on the Ethics of Artificial Intelligence Edited by David J. Gunkel
- Kijewski, S., Ronchi, E. & Vayena, E. The rise of checkbox AI ethics: a review. *AI Ethics* (2024). https://doi.org/10.1007/s43681-024-00563-x
- Journal AI Ethics (Springer)

So, this is the literature that I have given. There is a very interesting journal called AI Ethics from Springer. It has a lot of very useful information. I have cited only very limited literature here because the literature on this is exploding, and I said we need to start from the days of Aristotle, also looking at the philosophy of technology and then a whole lot of other literature on it. So, we will stop with this on AI ethics.

Next

- AI Ethics in Law and Justice

In the next class, we will extend this discussion into AI ethics in law and justice. Right now, what we saw are the broad common principles of what AI ethics is, why there is a need for AI ethics, and how UNESCO's AI ethics addresses that. But then what are the problems in making AI ethics workable in different contexts? We will take this forward by looking at AI ethics and law and justice in the sense of how AI ethics can be translated into law and ethics, or whether it has really been translated into law and justice; we need to examine this closely.