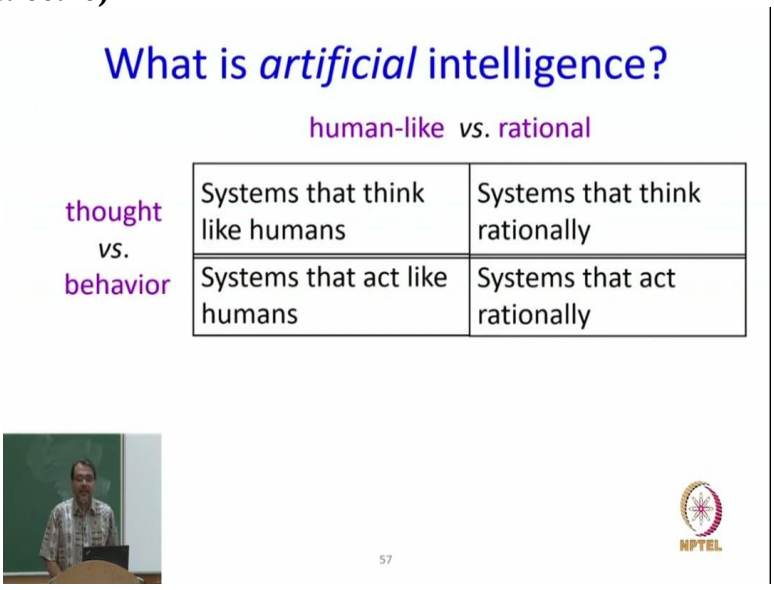
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology Delhi

Lecture-8
Introduction: Definition of AI Rational Agent View of AI, Part-8

So, where are we, we sort of said that we need to define AI and in the process of defining AI. We came up with several candidate definitions like a system that can think like humans is AI system that can act like humans is AI system that can think rationally is AI system that can act rationally say AI.

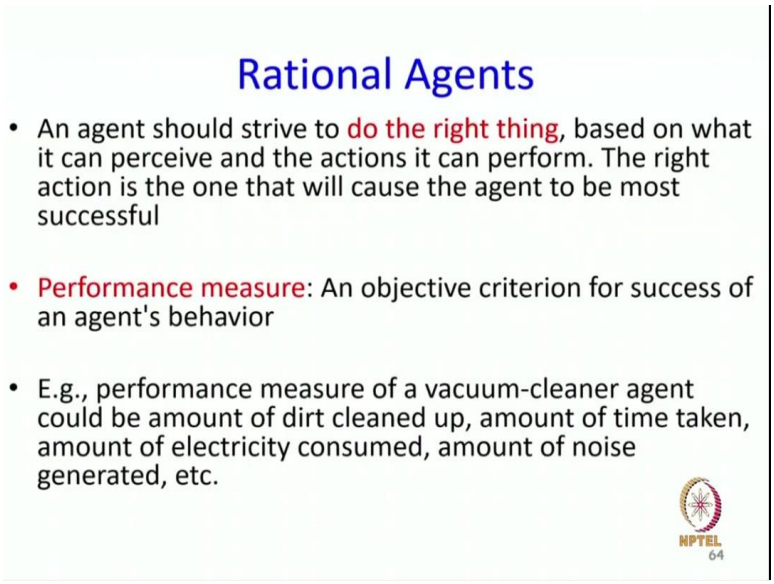
(Refer Slide Time: 00:40)



Now, we could take these definitions and finally, we came to the conclusion or somewhat, that acting rationally is a good idea. We should think of AI as acting rationally because there are less issue with this definition. There are some issues with other definitions. So far, so good. But we still were not able to answer. We have not yet gone to a place where we can answer what is acting rationally in particular, what is rationality?

We said that thinking like humans and so on a non operationalizable definitions, but we need to operationalize rationality at this point and that seems daunting as well.

(Refer Slide Time: 01:23)



So, let us think about the idea of rationality. And I will also say that this is the rational agent view of AI. There are different views of AI. AI is science AI is engineering and so on. But this is the agent view of AI is that there is this organism. Now this organism need not be a living organism. Need not be even a physical organism. It could be just like an internet organism or digital organism for all we know, but think about it from an organism point of view.

An organism as like we have hands and mouth and eyes and ears, eyes and ears of our sensors hands and mouth will be our effectors will be the way we can change the world. And in the same way an AI agent or an AI based agent will be situated in some environment and it will have some sensors and those sensors will tell the agent the world, that it can observe aspects of the world that it can observe, and it will have effectors and these effectors will be able to change the world.

So, now, we need to create some computational part in this agent, which is going to decide what should this agent do? What should this organism do, the part that would decide with that would understand the world that would decide the actions and so on, so forth. We can informally call it the brain, but that will be the AI part of the organism. So in reality, when people start building AI systems, they eventually they realize that AI is a small part of the big picture. Like if you think about us as people, how big is the brain, it is not tiny.

It is not like a small, insignificant part of the body, but it is not the full body. There is lot of organs in the body, they all have their needs, and you know if anyone organ goes bad, the whole organism stops to function. So everything is important, our heart is important, it gives the energy, our legs are important, these are our effectors and so on, so forth, but in the brain is also extremely important.

And some people will argue that the brain is slightly more important than other organs because your brain goes bad, the whole organ goes and you can, you know, we can debate that. But in the very same way, if you think about digital organism, let us say which lives on the internet, then it must you know, connect with the internet using networking. And that would be part of the organism; it may have some sensors it may be able to read it may be able to look at the images, etcetera.

All those would be part of this organism, it may have various layers of data management layer, the data basis layer where you can query it, they may be plugins with how to self distribute a particular data across many, many servers, etcetera. And all these various aspects that you study in computer science are going to be important when you create this real organism. But how do you control this organism? When do you figure out what to store? When do you do computation? How do you choose what you know answers to give to the quality.

That has come out; all of those will be considered or maybe considered the brain or the AI part of the Agent. So again, the point that I was trying to make is that if you really think about creating an AI startup or maybe AI system someday you will first have to create all these various aspects. Before you can start thinking about creating a high quality brain, high quality brain is extremely important.

But at the same time it is important that the other parts of the organism are developed so that the brain can actually control something and sense from something. So with that in mind, let us think about the rational agent now. So we have now got an agent, we now want to create its AI brain. And the question we have to ask is what actions should it do? And what how and we want to

define rationality in this process. So, the idea of rationality is more or less should do the right thing. But of course, that is not it is very vague.

That what is right, what is right for me can be wrong for you, We have so many multitudes of people in this country and in the world and you know what is right for one community is usually not right for the other community, we all know this. So how do we formally computationally define the idea of right And the way we are going to define it is by defining a performance measure. The performance measure in this particular case would be an objective criterion for success of an organism. This objective criterion would be given to this system for now.

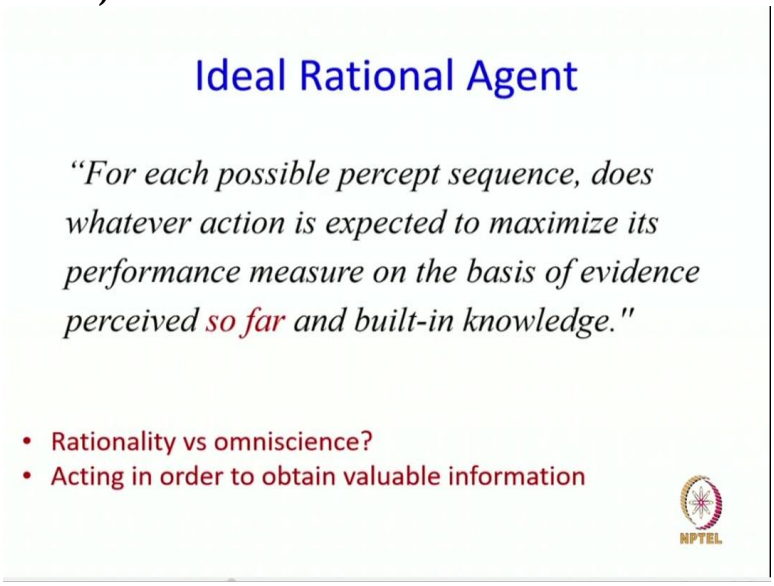
So we are not going to think about what is the right objective function because that leads us into territory which is extremely complicated. We will not we will for now consider the AI system has been given an objective function. It is like you have been told by your parents. This is how you have to live your life. This is right this is wrong. This is what you need to optimise and you are just living it. Now of course that does not make for a really intelligent person. But for now we will say that we have quote unquote a master a human.

For example, which will tell who will tell the AI system Look, this is what your goal is to optimise. This is what you need to optimize and that is your goal. For example, if it is a vacuum cleaner robot, it could be the amount of dirt it has cleaned up, although there are challenges and we can talk about that later, when you get into human AI interaction. But there could be other parts of the objective function like the amount of time it takes the amount of energy electricity it consumes, amount of noise it generates.

For example, we can hear some noise generated by the top floor because of some construction and that is disturbing. So that is not optimizing our objective function today. So we will have create a value system we will have the AI system. We will create an objective function for the AI system. Which will have these various components and then we will somehow combine them into a single objective function.

Now, given this objective function, what should AI system do? It should take an action such that this objective function is optimized. If it is a minimizing function, it gets minimized if it is a maximizing function it gets maximized.

(Refer Slide Time: 08:31)



So that therefore becomes the definition of AI or definition of an ideal rational agent. So again, so it is in a world it is an environment. So it is observing so it is the definition says that for each possible percept sequence, what is a percept? It is what you observe from the world in our ways, in a world, my IC, what my ESC and the sequence of observations like that is the input to me. What is the input what I know what I can see in the world what I can sense in the world that is my input.

Now, for each possible percept sequence, the agent has the ability to take an action and it should take the action that maximizes is optimizes its performance measure. So, the agent is given some optimization function the agent sees and observes the world and on the basis of that the agent decides this is the action I should take. But there are 2 more components in this particular definition. The first component is on the basis of the evidence perceived so far this is extremely important.

See how many times we have been in a situation where we did something and after we did this something wrong happened or something bad happened, something that we did not expect happened. And then they felt like man, I did not make the right decision. Have you been in such

a situation? Do you feel bad that we did not make the right decision? How many of us feel bad? Sometimes, there have been instances you are all wrong.

Why are you wrong? It is very important to realize that this whole idea of feeling bad about the past decisions needs to be thought again. You must not regret regretting a past decision is not a very intelligent and wise choice and I will explain why. See when you took the decision; you did not know how the cards are going to come out in the future. Fair enough. You only had some expectation about how the world will look like once you take this decision.

You cannot predict the future. You are not astrology. And by the way, even astrologers cannot predict the future. They say they can. At least that is what I will. So you think that the future is going to be like this, because you do not know the future, you only have some model about how the future looks like. And you took the decision based on whatever you knew at that point in time. And now, after you made that choice, you know, various things happened in the world and your action ended up being quote unquote, what you think is not the right action anymore.

But that does not invalidate the action that you took to once upon a time based on not knowing what is going to happen in the future. The best you can do is to use whatever your prior knowledge is about how the world looks like, of how the world will behave when you take this action. Moreover, at least from a psychological point of view, you are comparing a life that you lived. And you know each little detail about that particular life once is a life that you believe you would have lived.

If you had taken a different decision, then this is not even a fair comparison. Because you do not know what little things would have happened if you had taken a different decision in a long back? Does that make sense? Is it too, so the point is, and I learned this firsthand when I was a young kid, and I had an uncle, who came back to India, you know, 20 25 30 years ago, from the US. This is 3 my days. This is I was a young, young child. And so once I asked him that do you think you made the right decision?

And he gave me a very philosophical answer. He said of course, had to decision then I did not know exactly what is what is the life I would be leading when I come back to India and join a particular job at a particular place. We did not know that. And today, if you asked me what decisions I would take, I would take a different decision, he said. But that does not invalidate my past decision. Moreover, this comparison is not fair.

Because who knows if I lived in the US what problems I would have encountered there, and how my life would have looked like today. Maybe at that point, I would have felt that, I should have moved back to India along but you never know these things. And inside in our in our world where we live only one life. And we cannot backtrack our life to a previous point in time, time travel has still not been operationalized in practice, maybe someday.

It is impossible for us to really do this comparison. And ever whenever you get into the situation, where you feel like you took a bad decision long time back based on which you know, you suffering today, think again most likely you did the best decision you could have taken then now, not always. There are people who just you know who are not able to think clearly or who just go into a very emotional state of mind and make choices without really thinking about what it is that they know even at that point in time.

And then yes, you can feel that you can say that I was a fool, wrong them back. That is, so what is very important at 2 points here, what is the basis of making my decision and the definition of idle rational agent says, on the basis of evidence perceived so far, so if something else happens in the future, that and I could not predict this, at least in a high probability way, in the now that does not invalidate my decision, and also on the basis of my built in knowledge.

So there are some things I know today, they may be more things I will get to know in the future, but I will only assess my decision based on whatever it is that I know today. So therefore, it is important to realize, by the let is talk about built in knowledge, by the way, but it is actually an important point. So think about a 4 year old child. Let us say you have just started talking to them about, you know, a one digit addition, you have just said, 2+ 3, and the child says 5, and so on, so forth.

And just for fun, you ask the child, what is 8 + 7? And the child says 15, you are like, wow. And then you say, what is 15 + 12. And the child says, 27, and you are completely in all these things do not happen very easily. But if they happen, what would you call this little child? You will call this little child very intelligent and of course. The word intelligence has to come in every soft. We will call such a young child intelligent.

But suppose I asked this gentleman here and say, what is 15 + 12? And he says, it is 27 would we call him intelligent, he may be we are not questioning his abilities, but this particular demonstration would not lead us to making the conclusion that he is intelligent. Why is this difference in our assessment of intelligence with a young child versus a grown up man and the difference is very obvious, the difference is that we have all learned to digitization, maybe in class 1 or class 2 at some point.

Now that we have learned the 2 digit addition, the ability to execute it and give me an answer does not qualify as an intelligent behavior. It may be a fine behavior for that particular question, but that does not constitute intelligence. On the other hand, if there is a concept I do not know. And then I figured out how to do this that may be considered intelligent. You may find some friends or family members.

Who do really well in known environments but who get extremely uncomfortable in a new environment? Where, let us say they have to start playing a new game and they cannot figure out how the game works, or they have to start talking in a new setting and they take a long time to learn. I would say that that is an intelligence question, how quickly can you adapt? How quickly can you infer similarly, for theorems, if I gave you the proof, you memorized it and put it down in the exam that does not qualify intelligent behavior.

On the other hand, if I gave you a new conjecture, which you had never seen before, and then you are able to prove it that will be called intelligent behavior. So in the very same way, what is intelligence, therefore, is heavily dependent on what it is that you know ahead of time, if you already know the answers to the questions that are going to come your way and you just really those answers and output them then that would not be considered intelligent.

So knowledge and how much knowledge Have I given to the system is absolutely important in figuring out what is intelligence and what is not intelligence. So, so just to summarize this picture summarize the definition of rationality. The goal for an ideal rational agent is to maximize or optimise its objective function. It would optimise it objective function by taking actions so if it does not have any power to take actions, you know, fall from the point of view of rational agent as ceases to exist. Because of a definition of a was acting aesthetic.

On what basis would the agent choose its actions on? the basis of what it knows already about the world built in knowledge, how to do to digital edition or not on the basis of evidence perceive so far as possible perceptive sequence. And therefore an ideal rational agent need not be only saying it need not know everything about the future it need not be an astrologer, rationality has nothing to do with what is going to happen in the future or what it is that I do not know.

It is only about I will do the right thing based on what I know. And moreover, this definition also enables the possibility of you acting in order to obtain information because there is something you know, and there is something you do not know. But then there is something you know that you do not know. Like, for example, you do not know which class is going on in the neighboring room. And for if for any reason, there was a reason to figure that out, you can go inside that room and figure out what class is going on and that would be considered intelligent behavior.

If in the long run you are able to optimise your objective function. So in other words, sometimes I may have to act in order to learn about the future or learn about the present. Things that I know that I do not know. And there, it is very valuable to take those actions because in the long run, and I am going to optimize the objective function and therefore in an idle rational agent view, we can we often talk about long term intelligence, long term optimization as opposed to very immediate and short term optimization.

And we will talk more about this particular facet of AI when we get to Markov decision processes. So everybody with the definition of rational agents? Yes, question. What if we do not know a single valued performance measure Watson in Michigan? So that is a very good question. What if we cannot figure out a single value performance measure? In fact, one can

argue that we as people have multi objective criteria that we are optimizing we all want to be rich. Is there anybody who does not want to be rich?

We all want to be famous. Is there anybody who does want to be famous? We all want to be healthy. Is there anybody who does not want to be healthy, we all want to have a spouse somebody we really like or love or at least a good partner we can spend our life with we all and many of these things are not achievable. So, so, you will start finding that there are many, many things that you optimizing.

But we will ascribe with different importance to different of these sub objective functions, like some people will say, I am willing to let go of a lot of money, if I can get fame and they become professors. And somebody may say, I need a famous word. But I did rather you know, make a lot of money for my family, and they become engineers at Microsoft. And then there are some people who say, either becomes rich and famous, who care for anything less than that? They start a startup and very few of them succeed, but that.

They have higher risk taking ability. Some people say I can take a lot of risk, they start startup, some people say I do not take a lot of risk, and they only get jobs in large companies. Some people say that, you know, at healthy is important, but if I have to choose, I did rather die at 40 and be famous and, but instead of, you know, living until 65, and not doing anything worthwhile in life, they make decisions when they become hurting, or whatever it is. So the point I am making is we all have slightly different ways in which we make those decisions.

And if I had to ask you, you know, how much more newspaper articles are worth 1,000 dollar of a month, you will not be able to answer this question, what the heck does that mean? How do I compare fame in some criteria versus you know, money in some criteria that just does not make any sense whatsoever? So, that is the philosophy. So coming back to the question that we should do that. What can you do if they are really multiple object is for an AI system like time and noise and amount of dirt and electricity and so on so forth.

There are many, many models that study multi objective, one model says, create a linear approximation linear interpolation or some kind of combinations to create a single value function, that is a combination of all objective functions, like point 5 into fame plus point 3 into money. And so, then we will say that that is very hard. So, I will create a constraint optimization problem. So I will say, I want to optimise my money, subject to spending no more than 1 year of time.

Whatever I want to optimise one value function subject to my battery being less than my maximum battery for example, that that would be a typical way of doing this. These are called constraint optimization problems and these are very hard to solve the other objective function way is to say, I am not going to ascribe any value to how much one is important and how much the other is important, I will just optimise for all the criteria in a way that I output a Pareto optimal set of solutions.

Now, this is a, this goes into advanced AI, which we will talk more about, but the point is that they define non dominating sets of solutions. Were in a solution where I make, you know, a million dollars and my healthiness is point 5, whereas in another one I make, you know, 100,000 dollar, but my holiness is pointed. You say that, look; I cannot compare these 2 solutions. Because in one case money is higher, health is low in one case health is higher money is low, there is no way to do this comparison.

So I will output both of them. And I will let a downstream master figured out which solution is appropriate for them. All these non dominated set of solutions are called Pareto optimal set of solutions. Like for example, if I have one solution million dollars point 5 health and other solution, half a million dollars and a point 4 health then they will say. The second solution is obviously terrible, because it has even worse number for health and worse number for money.

So, I will not output it, but anything which I cannot compare directly I will output it and that is called Pareto optimality. So, there is a branch of AI. Which studies multi objective criteria, anything that we discuss, you can create a multi objective version of it, but we will not talk about

it in this course good question. Any other questions? So, this is a good point to stop and we will meet tomorrow thanks