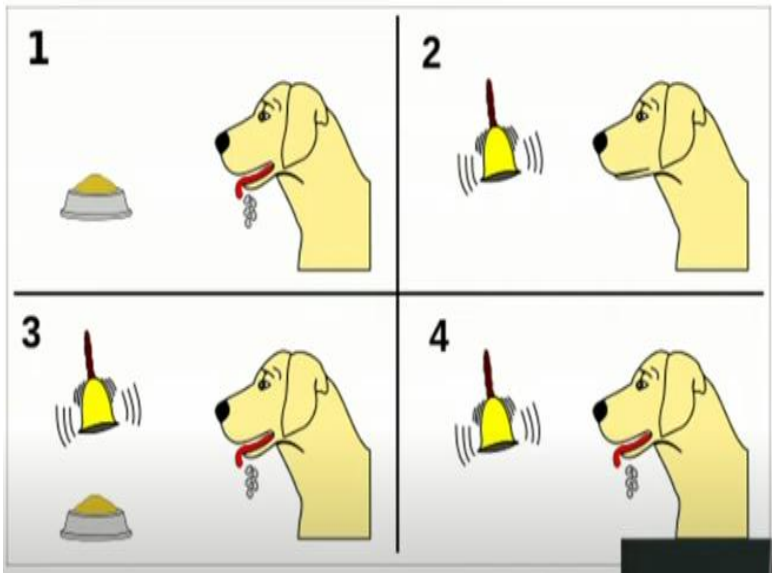
Artificial Intelligence
Prof. Mausam
Department of Computer Science and Engineering
Indian Institute of Technology-Delhi

Lecture - 76
Reinforcement Learning: Background

So today we are going to talk about reinforcement learning. Now this is a direct continuation of what we have been studying in the literature on Markov decision processes so far, right.

(Refer Slide Time: 00:29)



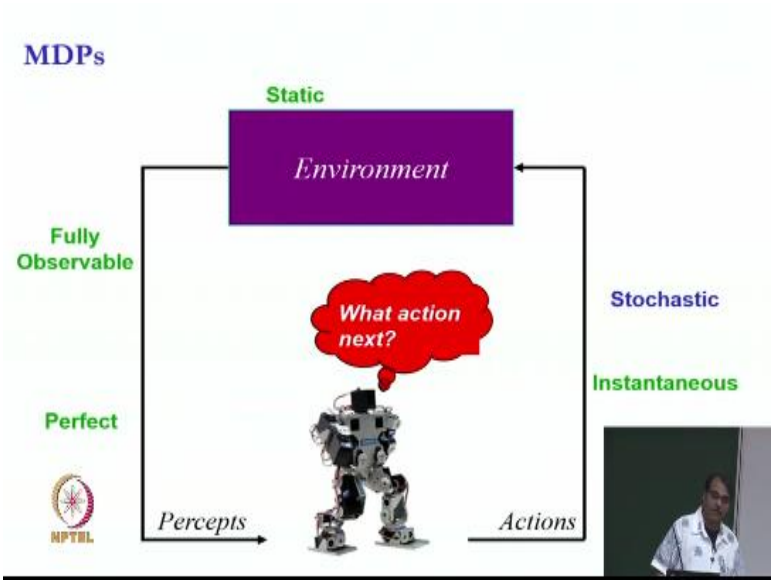
So look at this cartoon, right? So can somebody recognize what kind of what event this cartoon refers to? So it is an experiment. It is a very famous experiment in the field of psychology. It is an experiment by Pavlov and his dog is famously called Pavlov's dog, right. And the experiment was very simple. You know you give food, the dog starts salivating. You give food and ring a bell, the dog starts salivating.

You give food, ring a bell, dog start salivating. And then at some point, you just ring the bell and dog still starts salivating. Because the dog has learned the behavior that by at the time when you ring the bell salivation is the right action to do. And this particular phenomenon is the phenomenon by which we train dogs, or we train elephants, or we train any animals that we are able to train.

And this phenomenon is called reinforcement learning. So the topic of today and the topic in the in this week is going to talk about how humans how animals and our machines learn behaviors, right. And there is a very famous Big Bang Theory episode. You all of you know Big Bang Theory, a very famous TV series. I hope that almost all of you know it. Especially because you are engineers and you will enjoy.

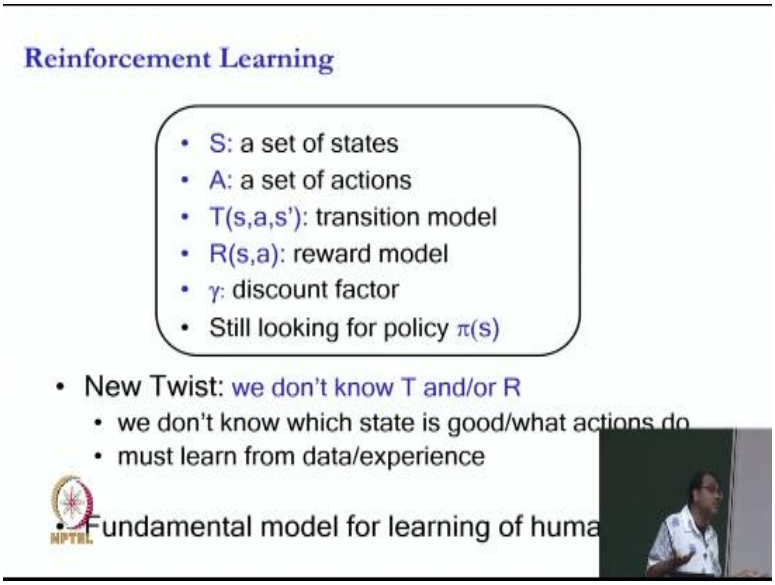
Where, you know Sheldon tries to teach Penny good behavior. For example, not talking loudly on phone when others are present. So what Sheldon does is Sheldon gives Penny small bits of chocolate for every time she does the right thing in the hope that she will learn. Of course, that is a joke. I am not saying that, you know we should teach Penny like that. But the point is made this is how you know lot of dogs and lot of animals get trained.

(Refer Slide Time: 02:44)



From our point of view, the computational point of view we still live in the world of a Markov decision process. So the world is still perfectly observable, fully observable. It does not change by itself and so on so forth, my actions can still be probabilistic.

(Refer Slide Time: 02:59)



Which basically means that I am in this, I am going to be in this standard MDP model where we have states, actions, transitions and rewards and let us say a discount factor. I told you that this is the most common version of MDP in the reinforcement learning literature. And the reason you can think about it is interesting to see whether we are very goal oriented or reward oriented, right? And it is a it is an interesting question.

You should ask yourself this question that are you working towards a specific goal in life? Or many specific goals in life? Or are you interested in just reward? So you eat burgers because you get a lot of reward at that point in time and that is why you know what happens you feel very happy, right? Chocolate, you eat chocolate because it feels good.

You know you feel attracted to a person you feel affectionate towards a person. You know you love your family, because all these things sometimes produce some chemical reactions in your brain, which let you believe that you are feeling good about the particular behavior, right. So it is actually there have been studies which say that people are highly reward motivated.

And in fact the reward has equivalent chemical compounds that gets secreted in the brain. And then there is further analysis of something called intrinsic reward and extrinsic reward. So people, especially when they are children are only interested in the intrinsic reward the rewards that they come with their gene pool, right that that comes with the gene pool. Like I want to eat food when I am hungry.

That is the intrinsic reward. My feeling of hunger gets satisfied. And that is an internal intrinsic reward, right? The fact that I would like to you know sleep or whatever it is, those are intrinsic rewards as we say, food clothing shelter roti, kapda, makan these are sort of intrinsic rewards, right? Maybe not shelter, but at least food. And then there are extrinsic rewards.

The extrinsic rewards are rewards that we over time learn as you know good rewards. We those rewards sort of come from outside to us like giving some money to somebody else if they are in need, like giving some money to beggar, right. Now would you call that good behavior, some of you will, some of you would not, I am not going into the sociology of it that begging is you know organized crime and so on so forth.

But keeping that aside, suppose there were somebody in need and you gave them a few rupees or something, food or something, you will feel good. Now that is not intrinsic reward that is sort of extrinsic reward. That is reward that you have learnt over time that comes from outside where you know you give something you feel satisfied, that satisfaction leads you to some reward.

Or somebody else praises you that satisfaction leads to some reward and so on so forth. So that is reward that eventually will lead to intrinsic reward, but at the face of it is sort of extrinsic and there are chains of behavior or chains of reward chains that sort of let you get to the intrinsic reward, right. So people have studied that the whole idea that we are trying to optimize for reward is a good way to think about the world.

Is a good way to think about agents. And so therefore a lot of people in the reinforcement learning, also take that view. They do not take the goal view they take the reward view, right. So we have states, actions, transition model, or reward model and because we will always work in the infinite horizon we have discount factor that is also reasonable.

Because you know burger, eating burgers now is more interesting than you know there are long-term bad consequences that we will have for most people. So you know

having a discount factor is reasonable for people as well. And we are still looking for the policy pi. So we are still looking for the mapping from states to actions such that my expected discount to reward over infinite horizon is maximized.

So we are still in the world of an MDP pretty much as we had been have been defining this last week, except that we are going to add a new twist and the twist is going to be that I may not know the transition function or may not know the reward function or may not know both or aspects of them. Anytime where any of this is not fully specified, we are no longer in inference world. We are in learning world, okay?

And so when that happens, suppose I do not know what behavior makes my mother give me milk. What would I do? You will try many things, right? You will try to cry, maybe I get milk like that I will try to you know move my muscles like this maybe you know that is how we learn how to smile. Smiling is not an intrinsic behavior to us.

We just, you know move our muscles like this and you know everybody starts over everybody around us gets overjoyed. So we feel like maybe that is good behavior, right? If you do like this and they say no, and we have to learn whether no is good or bad, but we slowly learn that no is bad. So we learn. So we all like babies learn using some reinforcement learning.

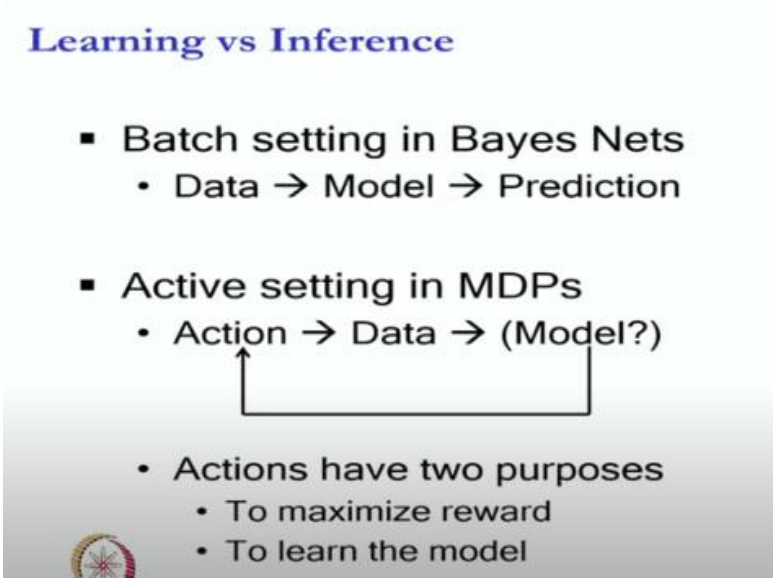
We sort of get reward signals from outside or intrinsically and that allows us to learn behaviors that have made us who we are today. So we will have to try out actions and learn from experience. So we have to figure out what actions give us good reward, what actions do. You know babies like to get everything in the mouth. Babies like to drop everything.

Babies like to, you know I remember distinctly that, you know I gave my friend a very nice toy, where if you, you know throw it on the floor the toy starts you know shaking and doing all kinds of things. The kids loved it. And they tried on all the ganesh morthy's of my friend. They tried to throw it on the floor and see whether it you know shakes. Of course it did not. And so my friend was not very happy.

But that said, you know we do not know the model of the world. We do not know what shakes and does not shake. And so, I have many such stories. If you have either kids or if your friends have kids and you like to observe them, then you will have many stories. By the way, a lot of reinforcement learning, researchers love to have kids. Because they can observe the kids, and they can observe how they learn and then they want to model it computationally, right.

Okay, so and therefore, it is a fundamental model of learning human behavior, right? So we will have to learn some data on experience.

(Refer Slide Time: 09:40)



And just a few words on learning versus inference. What is so different? So we have been doing learning throughout this course, right? We have been doing bits of learning, every now and then, like, when we did games, we tried to learn the utility function. When we did Bayesian networks, we tried to learn the Bayesian network parameters and even the structure of the Bayesian network.

So and in all those situations, if you remember data was given, which allowed us to learn the model which allowed us to make the prediction. That is sort of how our model has been. You give me data of the games that allows me to learn the utility function that allows me to make the prediction of what is the next action. Or you give me data, I learn the Bayesian network parameters that allows me to learn the prediction of the next query.

Now what is so different here or is it sort of the same thing? Well, I am going to argue that this is going to be fundamentally different or if not fundamentally strongly different from the way we have been doing learning earlier. And why is it different the way it is different is that we will take an action and we will observe the world. We will observe the reward we will observe the transition.

And based on that we will create data. So we are not in a back setting where somebody gave us a data set. I mean, we can be there are versions of that in reinforcement learning, batch setting RL has been done, discussed and studied. But in the most general version, we will take an action you will get a bit of data. And then we can do whatever with it. We can use that to learn the model.

And then we will use the model to figure out which next action to do, which will generate the next bit of data, okay. So therefore, a lot of people believe that this is as general a learning problem as possible. It is not one step learning or one short learning as people say. It is not that you give me data, I learn the model or whatever I need to learn now my learning is done, I will never learn ever again in my life.

You give me a new query, I will give you the new answer, but I am not learning anymore. I have learned whatever I had to learn. Well this is not how humans and animals behave. And this is not how we want our agents to behave either. So agents want to do what is called lifelong learning. We have an agent which is continually learning. It is never stopping to learn.

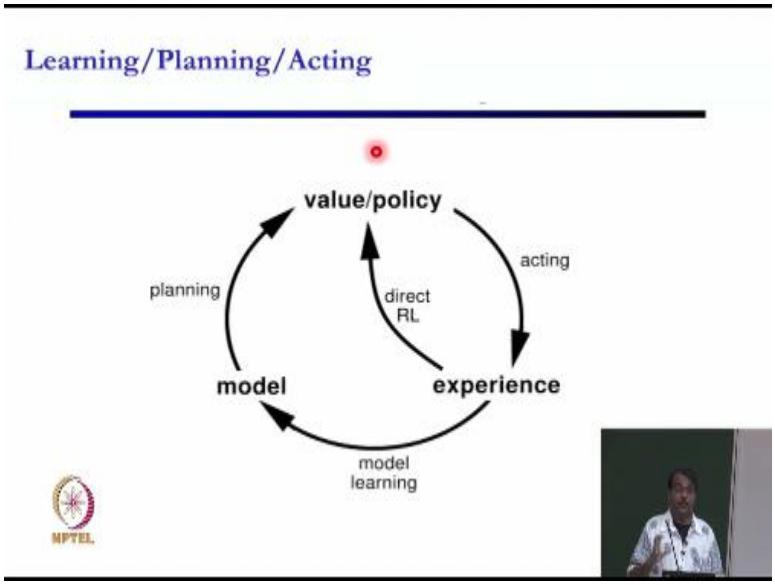
And it is not like that kind of learning where it was learning it was doing something completely terrible. So I generate lots of data and give it again go learn. It is not how we are envisioning this, we are envisioning this every bit of time we take an action there are two reasons why we want to do this action. We want to do this action to maximize reward. But we also want to do this action to learn the model of the world.

Think again of a baby. The baby has figured out that when it when he or she cries in a certain way, they get milk. Now if they feel hungry, they cry in that certain way, get their milk, they have maximized the reward. But now they have got their milk now

they do not have to maximize that part of the world. So they will try to throw things in the ground just to learn the model and see what happens.

So they are not always taking actions to maximize reward. They are also taking actions which they do not know what they do, and using that they figure out you know how the world behaves.

(Refer Slide Time: 12:55)



So this is sort of the setting we are in. There is learning, planning and acting. And what we did last week was we were given the model. We were given the Markov decision process, we were given the transition function, we are given the reward function everything. And then we did planning, also you can think of it as inference to figure out the value function and the policy and that is the edge that we were in last week, this particular edge.

But now we will allow ourselves to act in the world. Acting will generate experience or data and from there, we can improve the model and then do planning again or and this is very interesting. We can choose not to learn the model and directly figure out which policy to take. Think about life before Newton. We know that if we leave something, it falls on the ground.

We know that if we leave something our thing may break. It is not a good thing to leave something in the air and throw it out. But we do not know why that happens. We do not know the reason. Newton comes up with a model of gravitational force to

explain why something has happened. Okay, now it is a one-step detached model. We are talking about something more shallow, but you understand the importance.

If based on observation, I simply learn an action without recognizing why we are doing it or what is the reason, what is the transition function, what is the reward function, how this action becomes the optimal action to optimize my policy. That would be called direct reinforcement learning, or model free reinforcement learning, I will not learn the model of the world.

You know my mom says a lot that a lot of these rituals, you know we follow a lot of rituals in the Hindu culture. A lot of these rituals had a reason once upon a time for why they were defined. Whoever came up with a ritual came up with it with a certain, you know with a logical justification that okay this is happening because of this that is happening. So let me counter it by doing this particular action.

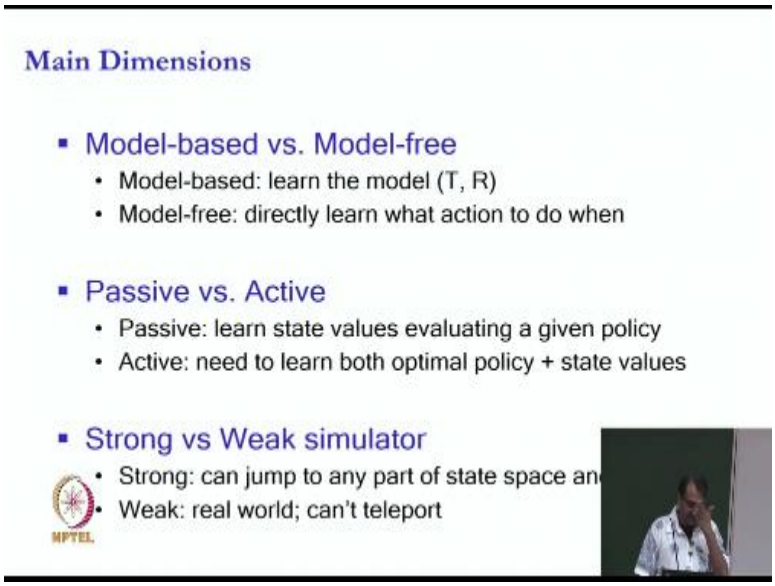
And now five people started doing it hundred people started doing it, it became a ritual, and then three generations down, nobody knows why we are doing it. And we are simply following that ritual, that we hear the ghanti and we start salivating. We have forgotten the reason why we were doing it. And now we are asked to follow rituals. And people say why should I follow this ritual or people follow the ritual, right.

Either which way they are not trying to get to the reason why that was defined in the first place, ask the question whether it is still relevant, and then make a decision on whether it should be followed today or not. We are just following it because it is religion, or culture, or whatever it is, right. So one is model free learning that we were we looked at the observation.

We decided this is bad action this is good action, we just start doing that. Now that the model changes tomorrow we will not know how to behave. If the transition function changes tomorrow, reward function changes tomorrow we will not know how to behave in that world, because we have not learned this action and we are just following that action sort of lightly.

Whereas, if we had a model and then our model changed, then we would have an easier way adapting to it right. So that is called model based learning.

(Refer Slide Time: 16:34)



So we will talk about model based versus model free in this discussion. Model based is when we are learning the model. The model means transition function and reward. Model free would be we directly learn which action to do without the intermediate step of learning the transition function and the reward function. There are two other dimensions that we will talk about one is passive versus active.

So passive is learn the state values evaluating a given policy. So this is like policy evaluation. Active is when we have to learn the optimal policy as well as the state values. You have to learn the optimal policies as well as the model or whatever it is right. So that will be called active setting. And then a very important thing strong versus weak simulated. So we are working in the world. So let us say we have the simulator of the world.

So now what is the nature of the simulator of the world. A strong simulator says you can start from any state and I will start simulating. So you just tell me which state you start from and what action you do, I will simulate and give you the next state and reward that would be a strong simulator, I can basically teleport to any state and start executing. Whereas a weak simulator is like the real world.

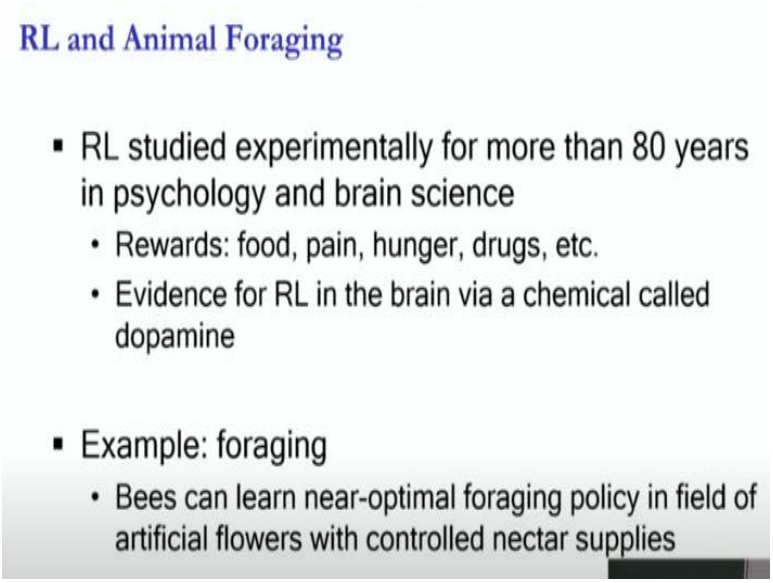
You are in it today. Right now you are in the state you cannot go anywhere else other than by taking an action and then you can only go to the state that you result in. After taking that action you cannot come back to the state you cannot teleport to another state. This is the closest that you can have to the real world. So the reinforcement learning setting where you are doing the active version, and you have a weak simulator is closest to how you and I learn and live our life.

At least this is what RL researchers believe. And because it is so general problem right, inference is part of it, planning is part of it. Acting and data generation is part of it. Learning from data is part of it. Trying to figure out which action to do such that I learn the best data is also part of it. Trying to figure out which action to do so that I maximize my reward is also part of it.

Trying to figure out how to trade these off is also a part of it. In the process, I may have to do constraint satisfaction to figure out you know whatever. So basically, the point is that people who study RL feel that this is one of the most general problems in the field of AI, at least in the field of single agent AI, because in multi agent there is you know minimax and defeating the other person kind of a thing defeating the other agent.

Here it is a single agent version in the most general form. And therefore, they love to work in RL because they feel like they are working in AI. So a lot of people believe that AI is RL right? Because this is as general a problem that you would want to define to. And if you can solve this then you have solved field of AI because you have solved at least the human learning version.

(Refer Slide Time: 19:22)



And not surprising, RL has had a lot of connection with psychology and brain science, right? So RL has been studied experimentally, for more than 80 years in these fields, behavioral psychology, brain science, how animals take an action how humans take an action. They have defined rewards like food, pain, hunger, drugs, right. So food would be positive reward hunger would be negative reward, pain is negative reward.

Some kind of drugs which give a certain kind of chemical response to your brain right will also be reward. And in fact, people believe that all our brains have a chemical called dopamine, which acts like a reward signal. So you can in fact, just check the dopamine concentration or activity in your brain and that tells you whether you are feeling high whether you are feeling that you are getting high reward or low reward or what.

So therefore, there is even a stronger connection between the RL formulation and people. And again, many, many examples that people have seen in animals behaving in crazily optimal ways like it said that bees can learn near optimal foraging policy where to which flower to go then which flower to go and so on. In a field of artificial flowers with controlled nectar supplies.

So bees sort of figure out the optimal path and how do they figure out. You know what processes are going in their brain? Are they doing model free learning or model based learning? Are they doing the shortest path computation or not? We do not

know. And therefore, it becomes intriguing because if bees can do it, the machine should be able to do it very fast, right. Okay, so with this background, let us get started on our first problem.