Learning about Learning a Course on Neurobiology of Learning and Memory
Prof. Balaji Jayaprakash
Centre for Neuroscience
Indian Institute of Science, Bangalore

Lecture – 12
Introduction of Reinforcement Learning - II: Classification, Thorndike's view, Tolman's views, Skinner box (cont)

Hello and this is the lecture 12 of the Learning about Learning lecture series. And so far we have talked about reinforcement learning in the lecture 11. And particularly we were talking about Premack's principle and it is modification by over land and timber lack on response deprivation aspects of that right.

So, we said we took the example of an; different sets of reinforcers ice creams potatoes and spinach. And we said Premack's according to Premack you should be able to modulate the response for the spinach by offering ice cream a higher responsive stimuli. And saying that hey, look if you consume more of spinach I would be able to I would be offering I will be offering you ice cream.

So, that should clearly increase the response of the spinach, but the cono random that we post. And then we try to overcome that using the response deprivation theory is that; if that is true is it possible that in by some means we can make the people eat more ice cream by offering make eat more potatoes by offering ice cream.

See you remember ice cream is the highest in our classification ice cream was the highest rewarding stimuli and then followed by potato and spinach. So, it is natural to assume that you can use the highest to modulate thus lower. But the question that we they were trying to ask is by some means can we actually make them make the lower responsive stimuli. In this case potatoes be rewarding while the natural thing itself is eating the ice cream right.

So, for the in reality it turns out you can actually do that; you can actually offer people certain amount of potatoes and then ask them and tell them; hey, look if you actually eat more I mean if you actually eat more ice cream, I would offer more potatoes people would do that people would increase their intake of ice cream for the offer of potatoes, not just like that.

So, the way you explain that is through response deprivation. Wherein you say that there is a basal level of response there is a basal level of likeliness that each of you each of us have towards these food varieties. And if you actually deprive the person of that likeliness like say in the measure that we have illustrated we said the 63 and 1.

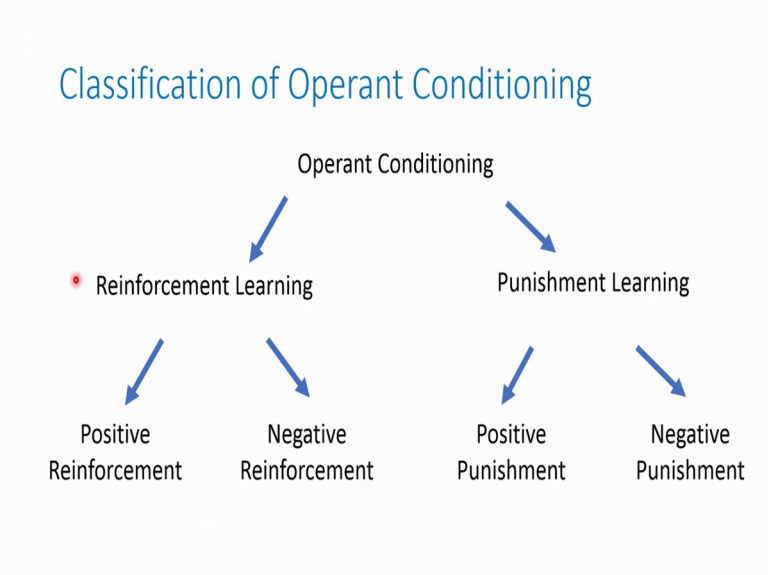
So, if I were to deprive the likeliness of the potato by some means to only 1; I give you only one amount of potato. Then you would have that deprivation in order to overcome that deprivation you would want to eat more potatoes. In that circumstance I can go ahead and say hey; look I would offer you more potatoes only if you eat more ice creams then you will adjust your likeliness of ice cream to or increase your likeliness of ice cream little bit more and then eat more ice creams.

In the expectation of getting potatoes even though potatoes to start with this less preferred there was also an example of how we can use that to in a classroom scenario to modulate the behaviour we did both of that. And then emphasize the fact that in order to understand this we need to take a closer look at the stimulus response outcome behaviour itself that is different from Pavlovian conditioning all right.

So, now in this lecture what I am going to do is that I am going to talk to you in detail about how the reinforcement learning per say is further classified. Just for the clarity sake we will be able to follow it easily and be able to understand the intricacies associated with them.

And then dwel upon the fact that when we have talked about what kind of reinforcements can be used in examples, but in a more organized manner how do you think about this reinforces. And what are all the factors that can actually change this reinforcement all right. So, let us start with the idea of the reinforcement learning and then that categories themselves.

(Refer Slide Time: 05:33)



So reinforcement learning of the instrumental conditioning or say and we can think of that as consisting of two big classes. One the reinforcement learning and two is the punishment learning all right. So, the operant conditioning the bigger class so reinforcement learning is a part of a bigger class called operant conditioning, which we know is different from the classical conditioning here right. The difference is you would have an outcome only and only if the animal response to a stimulus ok.

If the animal does not respond then there is no possibility of an outcome the fact that there is an a response does not necessarily mean you will have an outcome, but for sure it is true if you do not respond there is no outcome. So, that is the difference that is a key difference and that is that separates the classical conditioning and operant conditioning. Once you make that distinction clear, then we are talking about the subclass of operant conditioning in which you have a reinforcement learning that is one part, and then the punishment learning here all right.

So, that is the part that we talked about reinforcement learning. As the name implies here the outcome is a favourable behaviour for that animal so it is a reinforcing so to say. It is going to encourage the stimulus response associative strengths or the associative associativity. So, that is why it is called as a reinforcement. On the other hand punishment is going to reduce the response of the or the association of the stimulus and the response. I mean I should not say the associative strength here.

But I should say it going to the both of them are going to modulate the response values; one is going to increase the response you call it as reinforcement, other is going to reduce the response you call it as punishment. A very simple example here is a Pavlov's dogs experiment of a bell provided we say that; in response to a bell the dog has to press lever ok, then I am going to give a food.

You know food is going to reinforce this behaviour of pressing the lever in response to the light so that is reinforcement learning. On the other hand I can say I will take a rat or a mice or even a dog. And in response to a stimuli right I am going to present in response to a stimuli dog if we does not do anything I am going to present a little shock food shock right remember our shuttle box training.

So, in there we are going to present a simple food shock here. So, if you do the food shock then the dog has to respond not responding is not an answer there. And that again would come under operant conditioning you are responding to a stimulus, but it would be coming under a class of a reinforcement learning you call it as negative reinforcement meaning.

Let us say if there is a disturbing noise that is continuously present all throughout and that is present all throughout right in respond. In response to a stimulus if you are actually taking away that disturbing noise then that would classify into a negative reinforcement ok. On the other hand punishment here is you have a stimuli the dog responds for that response you are actually providing a shock. So, then you are actually talking about a positive punishment; I mean the punishment the punishment is being presented to it so that it is going to reduce it is response in reduction of the response defines it to be an positive punishment ok.

So, the reinforcement and punishment of positive and negative is defined in terms of how the magnitude of the response changes if the response were to change in to increase the magnitude you call it as reinforcement. If the magnitude goes down you call it as punishment that is true for both positive than negative. Clearly when the dog responds to a stimuli by pressing a lever and you are delivering a shock you are going to reduce it is probability of response later on.

So, it is a positive punishment you are actually punishing the dog for the stimuli. Wherein normally it would have gotten a shock, but because of this response right again the shuttle box training where if it were to jump right in response to a tone or within a certain time period. You take away the shocks for some period of time right. You are actually increasing that you are actually making the dog to perform this more and more so that you can develop this behaviour that is you present a stimuli the dog is going to respond more to avoid a punishment right.

So, the that is a positive change in the magnitude of the response so you call it as negative reinforcement. Again two kinds of learning reinforcement learning and punishment learning, the reinforcement and punishment is totally determined by the nature of the outcome that we are going to present. If it is a favourable one you call it as reinforcement, if it is a unfavourable one you call it as punishment. The idea of me classifying these is to say that when you are talking about reinforcement running or an operant conditioning you can say; hey, what if I were to do this or do that and stuff like that.

But any of that you are talking about any variant that you can think of in terms of the as a relationship between the outcome and the response. The nature of that relevant as it pertains to the animal will fall under one of these four categories. Once you say that this is one of this category rest of it is exactly the same. If take one category and then rest of the analysis that we are going to talk about.

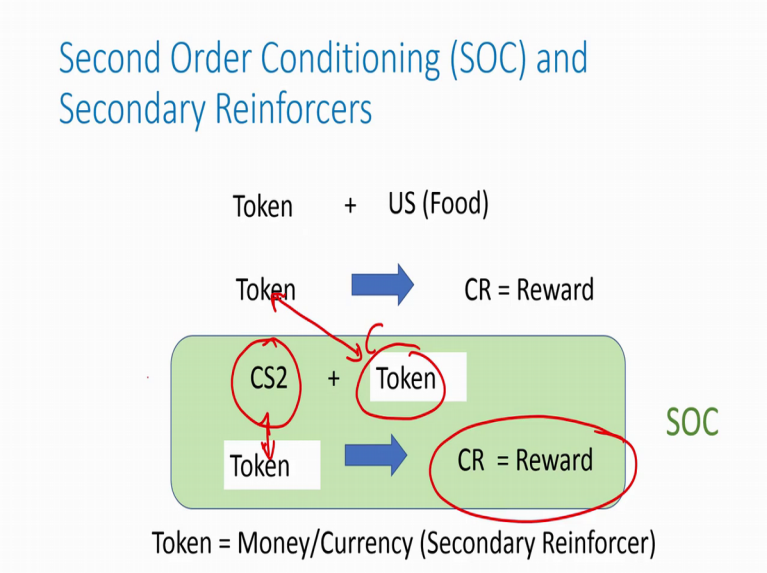
And rest of the analysis we are going to deal with is exactly the same. Except you are going to say the, that sign of the magnitude is going to change. And the nature of the outcome will be different that is all. But the general philosophy will remain exactly the same. So, now what is this general philosophy? The general philosophy comes to the point of knowing what are all there could what are all the stimuli that can act as a reinforces themselves.

Remember the whole idea of this we got into, but to understand; what are all the factors, or what are all the different ways the animal can learn right. One and if there are; different way, multiple ways, of exhibiting learning and then you want to ask what are all the factors; that can bias the learning towards one way or the other.

And in doing so then you are coming boiling it down to understanding the nature of the outcome itself. Nature of the stimuli in the stimulus response outcome curve nature of the outcome itself; how can it bias the learning? Or does it even bias the learning?

So, in order to understand that we first need to understand what are all the stimuli that can be used as a reinforces. Premack clearly said you can use you can look at the behaviour and classify them as more favourable less favourable and so on and so forth. But in a more organized fashion if you are to look at it you would think of the reinforces as primary reinforces and secondary reinforces.

(Refer Slide Time: 13:40)



And what we are going to talk about now is about secondary reinforces. In this context it is important to get the notion of an important way of conditioning an animal. Or a in developing a conditioning to a stimuli is called as a second order conditioning. So, far whatever we have seen has to do with what we call it as a first order conditioning; where you are taking a stimuli let us say CS1 and then you pair it with the US like for example, a food.

And the idea here is you develop an the animal develops an association and thereby develops this response to. Later on if you just present the CS1 you are able to elicit a response it is a CR. So, now, the CR is developed in you can think of that as an expectancy of a reward, or it is a reward response I would like to call it as a reward response.

Not to confuse with the fact that you are providing a reward it is the response exhibited by the animal ok. That is that is that is acquired through reward learning that is what I mean by CR which is equivalent to a reward. This could be just a simple salivation so that is our reward response in this case. Now in this condition it is very simple to say that look I am going to use the food as my outcome right in our stimulus response outcome.

And I use it as a positive reinforcer and increase or develop a new habit; very good. But evolutionarily it turns out having a few of such prime these are called the primary response primary reinforces very analogous to pavolous US right, the native responses. It is good to have them no doubt about it and they are there. But if only they were to be there, then it severely restricts the animals ability to moderate it is behaviour.

So, it turns out the animal can actually form a higher order association; what do I mean by higher order association having learned that the CS1 can to the point where the CS1 can by itself elicit a response you can take this CS1 and then pair it with another CS a new CS here ok. This is the same CS which has developed a response now you pair it with another CS new CS. Such that now the animal can actually elicit another response could maybe same maybe different, but it can elicit another response this again will be a reward response that is the key here.

So, we used the reward response that is acquired after the presentation of the CS and US; to form another response for us different stimuli altogether a different stimuli again a reward response. Now, what use is this? I mean it is good to talk in terms of CS1 CS 2 and all that? But what practical value does it have which I said it is evolutionally important and so on and so forth.

Now what practical value it has even evolutionarily right let us say I switch the name of CS1 to a token all right. So, let us say if you where to give me a token I am going to give you a food. Now what is going to happen is that the token itself later on is going to develop a notion of a reward feeling of a reward right. If you get a token it is equivalent to getting a food right you will develop that reward. And in fact, that happens because of this association all right.

So, you can take something else some other stimuli and again that could be another token. Thereby making I mean this is a reward this itself is eliciting a reward right. So, you could this token is actually you can think of this as a response elicited by the token.

So, now, this CS 2 acquires a new meaning in terms of the reward itself. In fact, the token here I am talking about this could be just money or currency all right.

So, at some point in time the money here was needed for you to buy the food. But later on you develop an association you can develop an association wherein I do not have to present the food at all anymore to you. I can actually give you taken completely new stimulus and then pair it with a token right the reward eliciting token right it is a money here right so pair it with that token. Now the CS 2 itself would develop a response which is a reward response.

Again this itself can again act as another proxy for the token. So, there are several proxies for the token here CS1 and CS 2. Now, but the bottom line is the token can act as a proxy for the food. Now this conditioning is called as a second order conditioning. And the reinforcers such as this token right the CS1 token are these reinforcers are called the secondary reinforcers. And they are very very important as you can see money is important and that is how you develop value to that money or a currency.

Um as a child you do not have much value to that right. Over a period of time you develop that association only through this kind of second order conditioning. So, now you can clearly see this kind of a situation allows in an a natural while to money is a civilized phenomena; phenomena that happened because of civilization.

But you can think in terms in the natural while it can actually provide a very valuable way of expanding it is repeater I mean for changing it is modifying it is behaviour. And in fact, people have tested this in one of the famous experiments by Wolf and co-workers. What they have done is that; they have trained the chimpanzees to elicit on a vending machine, where; you actually the chimpanzee drops in a token it gets a grapes.

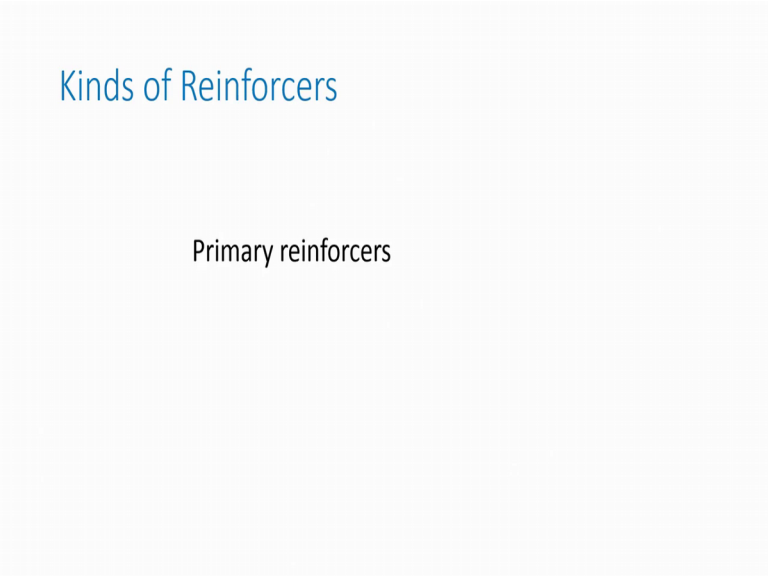
And it is food deprived so it likes to get the grapes and eat the grapes too. So, in such a case the token itself is being associated with the food. Then what they do is that they present lever which is hard to operate. And when you and then when the animal operates a lever you get a token. What they show is that the chimpanzees would work almost the same or in fact, even more on the lever when you are presenting the grape or these tokens.

The moment you learnt the chimp you make the chimpanzee learn that tokens can give you grapes. And what more interesting even more interesting is that; the dominant male would like to snatch the coins from the other chimpanzees when they are actually working together; this is not thought it is actually in there and it is coming out. So, such is the importance of the secondary second order conditioning and the development of secondary reinforcers.

And as an experimental what it allows us to do is that it gives us immense flexibility in terms of; how we use these reinforcements? And what we use as the reinforcers? Apart from this while I mean skinner was a skinner is the kind of the torch bearer for doing the studies that I am going to talk about later and also much of the studies. So, while he was a graduate student and he is actually doing this experiment; it turns out that like any graduate student he works really really hard he works the weekday's weekends and every hour every minute.

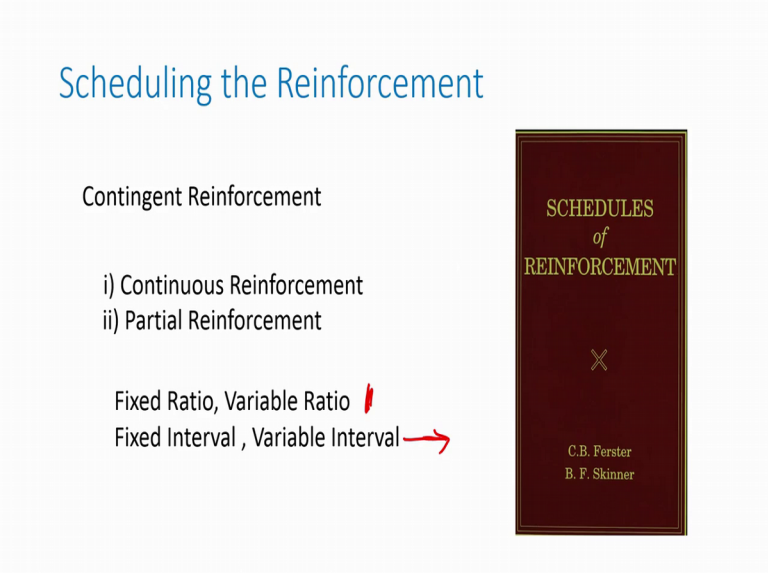
So, he did not want to waste any time so in one such instance he is working on a weekend and then suddenly realized that he is running out of pellets foot pellets to be used for the training. There he decides to offer the animals this reward not for every response of the click, but let us say every ten response. Thereby he discovers a plateau of findings that were entered actually a separate book. What he discovered was the effect of scheduling the reinforcement.

(Refer Slide Time: 23:39)



So, so far what we have said is that whenever we are we can use different kind of reinforcers. And when we use the primary reinforcers we have always been using it in a contingent manner all right.

(Refer Slide Time: 23:48)



That is whenever there is a response you are presenting it no doubt about it. That by itself gives us some contingency, but you can also do this contingency remember go back to our discussions about Rescorla Wagner. You can also have this contingency without having a continuity right. You can have breaks in the continuity break you do not. And still be contingent contingency is about predictability right whenever you are having the response the likelihood that you are going to have the food.

You can have the likelihood being high and consistent without having to give the food at every response. And that is exactly what skinner did right. So, when he did that he realized depending on how you vary this contiguity; we call it as partial reinforcement and as against the continuous reinforcement. You can have a different kind of a paradigms; four of them I am naming.

One is which of the dominant ones we will be discussing about the. So, I am naming them here list them here it is a fixed ratio presentation, variable ratio presentation, fixed interval presentation, or variable interval presentation. It is also called fixed ratio scheduling, variable ratio scheduling, fixed interval scheduling, or variable interval scheduling.

Now what are all these things? Fixed ratio as the name would imply he would skinner would reward the animal for after fixed amount of responses have elapsed. So, if the animal have responded some x number of times since the last reinforcement he would set that ratio for the x amount of time no reward I am going to I mean the he would set that fraction x by total number of stimuli total number of responses he is expecting.

So, as long as that x number has happened he is going to give the reward, until the x number happens no reward so that is called as a fixed ratio. So, you realize that here time does not have much sense here because the animal could take it eternity to respond to that number. Until it responds to that there is no presentation of the food. The variable ratio is a little bit tricky where he would say it is not a fixed number every time. I am going to change I am going to vary it around a mean.

So, he is going to he will determine that x number right once every 10 times or once every 100 times he will determine that number. About that mean he will generate the random numbers and then he will pick this random numbers. And then he will he will try to match for every. So, now, I have presented this reinforcement next my random number is this he would wait till that point then present that reinforcement. When you do it that way you call it as variable ratio presentation or variable ratio scheduling.

Here again time per say does not matter right, it you have to wait till that particular presentation that particular expression of the response number has happened. So, as against these where you are actually counting the response in some sense these are about time elapsed time; time that has elapsed since my since my last reception of the food or the reinforcement. If I were to present at a fixed interval what, what would what we would do is that you would fix a time t. And after this time t you are going to provide this reinforcement.

However, you have to understand in order to get that reinforcement at that the instance and later on the animal has to respond that it let me come let me repeat it again. So, the experiment starts here all right and we are following in time ok. What this fixed interval means is that you are determined. So, this is the instance there is a first instance let say. At first instance where we have reinforced there is food or from this time until sometime tau that is our fixed interval right.

There will no matter how many times the animal responds there is no food there is no reward. You would the reinforcement will become ready in other words at time t equal to tau and later on for all time t greater than tau the reinforcement will be ready. For the animal to perform the response and get that until then even if it responds no food after that it has to respond to get the food. So, now, that is a fixed interval presentation.

On the other hand in a similar manner as that of the variable ratio we can also do something called a variable interval. In this case we are actually varying the interval about a mean right so you are taking a distribution of random numbers about this tau. And in this first instance you are having the food ready or the reinforcement ready in an interval that is t plus some delta t or plus or minus delta t and so on and so forth. Now, that is a variable interval presentation. Now do the apart from giving names do these different scheduling paradigms matter. Clearly if they were not to have any effect we would not be discussing it. And in fact, it had.

So, much profound effect that I as I told you skinner had to come up with this book of schedules of reinforcements. It is an entire book very well written just describing about the schedules of reinforcement between ferster and skinner. With this we will stop here in this lecture. Thank you I will continue in the next lecture on the effects of this different scheduling. And then venture into asking the question that we started on. Do they learn in a cognitive manner or in a reflexive manner?