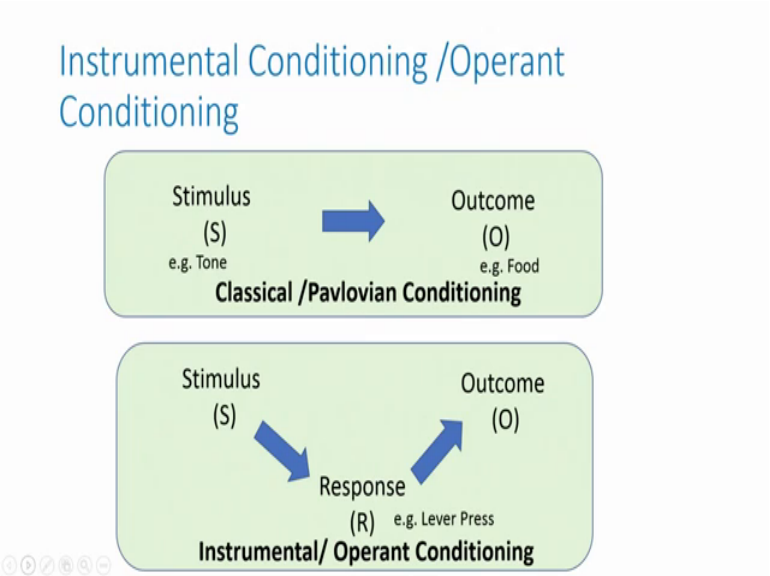
Learning about Learning A Course on Neurobiology of Learning and Memory
Prof. Balaji Jayaprakash
Centre for Neuroscience
Indian Institute of Science, Bangalore

Lecture – 11
Introduction of Reinforcement Learning: Classification Thorndike's, Tolman's views, Skinner box

Hello and welcome to the lecture 11 of the Learning about Learning NPTEL course. And, until now we have been concentrating on understanding, how the associative strength is developed when 2 stimuli. A, stimuli that has a native response and the stimuli does not have a native response in the context we are looking at or co-presented right. And, we developed a model called Rescorla Wagner model and used to explain some of the observations that we have made previously before the model was proposed, and then we also talked about the limitations of this model and where it fails.

Moving on from this, in this set of lectures in the forthcoming set of lectures. We will be looking at a different kind of learning paradigm, if you remember our initial classification of the learning's. We classified that as associative non-associative and within associative we talked about classical conditioning and instrumental learning or operant conditioning ok. The next few lectures is going to be about that and what insights can we get about learning from the experiments done to unravel this set of mechanisms.

(Refer Slide Time: 01:56)



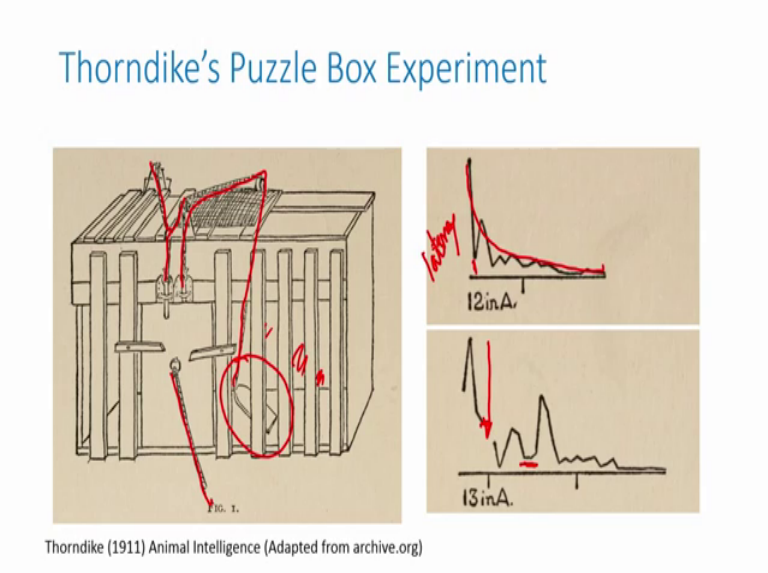
So, just to rehash what we have seen so, far is that when we have a Stimulus for example, a Tone. And, then couple it with an Outcome such as the food we understood we have seen that the stimulus over a period of time develops a response on it is own, that is the tone by itself develops a response on it is own in this particular case it is salivation. And, we call that as classical or pavlovian conditioning, that the difference between such a learning, such versus the instrumental or an operant conditioning is as follows.

Here, the stimulus per say does not necessarily mean that there will be an outcome, there is a stimulus we can go back to the same example; example of a tone, but that does not mean just the tone arrival of the tone does not necessarily mean; there will be food the animal need to do something and we call that as a response. In this case it could be a simple lever press; pressing of a button or picking of a key as we have seen in Jenkins experiment by the pigeons, or pulling of a string in the experiments that we are going to see further.

So, the animal needs to make a response after the stimulus has arrived, only and only if the response is made there will be an outcome, which could be the food or opening of a cage or any favourable outcome that the animal is happier to be in. So, now, such kind of stimulus response outcome behaviour when the animal acquires, we call that as the animal has acquired the instrumental or an operant conditioning. This entire process is the process of reinforcement learning, the reinforcement could be of a different kinds we will come to that in a probably in the next lecture.

But, the notion here is that stimulus per say does not predict the outcome, you need to respond and the depending on the response there will be an outcome. Now, why would anybody do such kind of experiments right, in parallel to Pavlov's experiment that was happening in Russia, Edward Thorndike.

(Refer Slide Time: 04:34)



He was interested in understanding, how animal think about things? Especially, he was very much interested in animal intelligence. In fact, he wrote a book called animal intelligence at the end of his study summarizing all this findings. The question he was interested in is, we all know that the animals do develop responses.

If, the animals do develop responses, how do they develop? Do they develop in a cognate do they learn and develop in a cognitive fashion that is do they make an expectation do they understand the process, that ok; look there is a tone I need to be pressing this lever in order to actually get the food. Or, it is just goes by a reflex in a non I mean in a way where the animal does not necessarily learn in a cognitive fashion.

So, you just there is a stimulus and they just plus that lever in a reflexive manner and then you get the food. Idea, here is that if you were actually developing this association by in a reflexive manner, it is much like an a trial and error learning, that is every for every response and every exposure that the animal has experienced a stimular stimulus and response outcome behaviour. Very mean it responses could have happened the animal could have elicited many more responses.

For example, he was using such kind of Puzzle Boxes. The box if you actually pay close attention to you will see that there is a pedal that is present inside this box. That pedal through the string is connected to a door latch alright. And, which of occurs pulls the door open.

The idea here is that he is going to take his cats. He, experimented on his own pets there is a bunch of cats that he used to raise them at home and then he would take them it is about 13 of them. So, he would take them and put them inside this box, they will be food deprived. So, they are there are multiple factors here that are playing in for this animal to actually come out of the caging. Number 1, they do not like to be in a confined space they would like mean when you confine them they would like to come out. And, you they can see and smell the food because there are gaps here, you can actually see there are gaps between these wooden planks. And, through which they can actually see the external world the food is kept somewhere outside.

So, they want to come out and then eat that food because of the food deprivation as we will as for this getting out of this confinement. However, they do not know by themselves how to come out of them. So, they need to figure out they need to understand, that they need to press this pedal in response to that the door is going to open and then you can eat the food.

So, his question was when the animal is actually learning to solve this puzzle is it really learning by making an intellectual association, that is hey look I am inside this box and I need to press this pedal is connected to the door. So, I am going to open or by accident it stumbles upon it and then, it just makes a correlation, but whenever I stumble upon that pedal the door is opening and I am going the food is there. So, that learning is not conscious and in a cognitive fashion here at all.

So, how would you test it? So, his argument to was that look I am going to do this multiple number of times, as animal is not going to figure out in one time. And, when I do it multiple number of times and keep measuring how long does the animal actually takes to come out of this box so, called the latency. Now, if the animal were to really think about it and solve it, you need to solve it only once you do not have to solve it every time in your life.

So, if you solve it only once and then next time you go into that box you know that this is why this I did that and I got this. So, now, I do it again press the pedal. On the other hand if it is a trial and error learning, you just did not know why I mean that there is no causal relationship, that is obvious and you did not do it with the intention of the causal relationship. Then, so, many different things that the animal could have done so, you the

animal really does not know what is really correlated and it is the association slowly develops. Wherein things that did not matter, but the animal still does.

For example, as soon as the he observes that as soon as the animal is put inside the cat is put inside that box, the cat would be very very anxious and he would run around the cage pull grasping clawing and pushing against every single plank that it can.

In doing so, once in a while it does press on the pedal. So, now, pressing the pedal is the response, but that is that we are interested in, that is associated with the positive contingency of opening the door. So, now, but for the animal very many responses that it did express right like clawing the planks and pushing against the planks all of them and pulling the strings all of them are responses.

So, naturally all of them would associate; however, being the pedal being the contingent one always. So, over a period of time the association develops for just the pedal while the rest of them goes down ok. So, now you would predict 2 different kinds of learning behaviour. In one; where there is a trial and error learning the one that we saw just now the second case, we would see a slow progression.

Once in a while reverting back to the latencies, because accidentally you pressed on that you do not even the animal does not even recollect that it need to press on it. So, it has to do it again to really reinforce the fact that yes it is the case. So, this kind of slow progression with once in a while reversal of not pressing the pedal would be characteristic of an associative learning, which is very reflexive not necessarily a cognitive process.

On the other hand if the animal were to really think about what is happening and then press the pedal you would see that it has solved it once, next time onwards it should do it much more easily. And, it should mean not just much more easily it should never revert back to the place where it is not able to solve that puzzle.

It is a wonderful book animal intelligence it is available in archive I will recommend strongly you guys to go through that. I am illustrate, I am putting out 2 of the trials just illustrate the point that Thorndike wants to get across; which is number one here he is plotting the latency in this axis. So, right and 12 in a; a is one kind of a puzzle box. So, he has a different sets of puzzle boxes.

So, he is actually putting them experimenting them with all the puzzle boxes. So, 12 means the number of the cat the 12th that is an ID ok. So, he is plotting out the latency as you can see latency is high. So, initially the animal does not even know what it is needs to do to get out of it. So, he will leave the animal in the box for about 15 minutes, by 15 minutes if the animal does not come out then he has to go and open and take the animal out.

So, this guy figures it out I mean it comes out it is not right to say it is figure it out, it comes out in the second trial probably, but then that were to be the case of thoughtfully solved puzzle, you would see that the next time around when you are actually putting the animal it should continue further, but it is actually not. So, not doing that it is actually going up, coming back again and doing this oscillatory behaviour slowly approaching towards a point where it can actually get out of the box by itself.

If, you notice this is exactly what you would have predicted. This kind of a curve is exactly what you would have predicted using a Rescorla Wagner model for a classical, for modelling the associative strength in classical conditioning 2. Except here you are looking at the invert I mean, the latency which is the higher the latency is lesser is the performance. So, you are looking at the curve in the reverse direction the down.

This is not just one animal I mean in fact, there have been cases where like for example, the animal 13 were very well into the plateau region right, right about here right where you have the animal has actually reached the plateau, and still it reverts backs to that ok. And, then when whenever you see a gap, this gap is because in this place; the animal even fail to come out of the box by itself he has to come in; so, whenever that happens he introduces that as a gap. I mean in the beginning if it happens it happens just 15 minutes. And, in the in the middle if he had to intervene and then the animal is not able to come out he will put that as a gap ok. At this point the animal even fails to come out of this box.

Based on this he proposed that what it is most likely happening is that the animals are learning through a non-cognitive process in a trial and error through a trial and error mechanism. Of course, with any such hypothesis it drew a considerable flak from the fellow psychologists, one particular thing flak of that is notable flag from (Refer Time: 15:46) in Austria. He was a strong opponent of such an idea of non-cognitive processes.

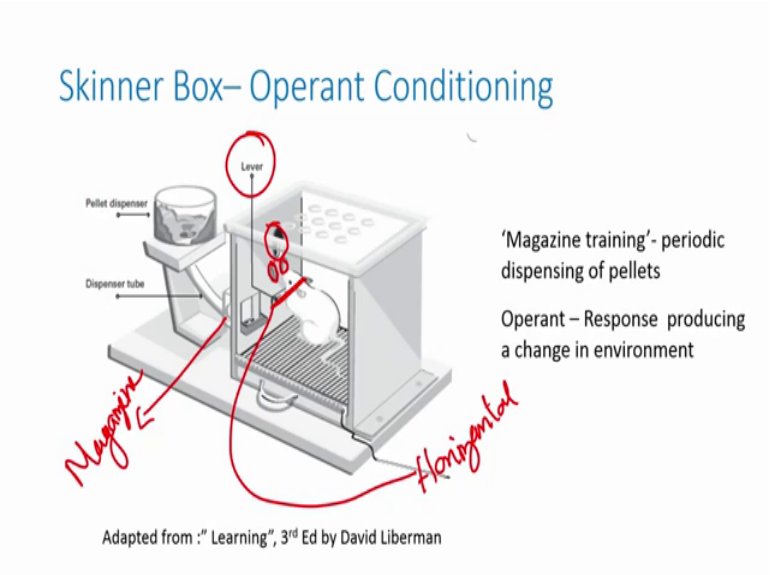
So, he is a base for making that claim was he worked with chimpanzees to show that really chimpanzees can actually cognitively think and go about solving these problems. But, the main drawback of Thorndike's experiment he would argue is that, the causal relationship is not very apparent. That is the cat never sees that makes the connection that pressing the pedal somehow releases the door.

So, it is because of that the animal never learns. In as I was showing in any such claims there is claims and counter claims, but the point being whether the animal is how and if the animal is learning cognitively, how is it learning cognitively or if it is not learning cognitively, why is not learning cognitively is of considerable debate. And, there are occasions where you would say that it is of one kind versus there are occasions where you think that it is not of the other kind. Towards the end of these lectures, we will come and revisit this very idea and then see probably what is the current view or what is our view in this regard.

But, needless to say you are able to ask these questions, ask such questions and probe these phenomena, because of this development of the stimulus response outcome behaviour.

Of course, when you do that behaviour by itself brings in it is own set of phenomena and parameters that one need to characterize and understand. And, for doing that such kind of puzzle box, while it is a very very important first step is not by any means the final step.

(Refer Slide Time: 17:55)



So, Skinner developed much more user I mean much more laboratory use or friendly apparatus, this known as a Skinner box for operant conditioning; wherein the animal is in here it is a rat. It is exposed to a set of cues here; one is the contextual cues where the animal is being present with the grades and the box and so on and so forth. And, the animal is having the opportunity to press a horizontal bar here right, this horizontal bar besides that animal is called the lever right. So, that is marked here, there is a horizontal bar that the animal can actually press with it is paw.

And, there are set of cues, visual cues. So, these are cue lights or a ton,e a speaker that can be used to play a sound. More importantly whenever the animal is engaged in any of these responses, the outcome for that would be a food pellet dispensed through this sort of an arrangement called us magazine. So, whenever the animal is actually pressing the lever after a few minutes the food pellet rolls, and then makes a characteristic sound of it coming down in into and the holding into this box.

Now, that sound together with the fact that it has to retrieve that food in a certain period of time, if food is not going to stay that for a forever. These 2 facts together draws the animal to actually develop the behaviour of pressing the lever and then associating with the fact that it is going to get the food.

So, response, outcome, that association we are actually forming. And, such kind of so, training is very very important when you are doing an operant conditioning thing and this is called as a magazine training basically. So, what you can actually do is that when the animal is many many times it is auto shaping they will shape by themselves. But, when they you have to forcefully safe shape them, what you do is you periodically dispense the pellet us, just making sure that there is that noise, and then withdraw the pellet us in a short period of time. So, that they understand that the food is being delivered and slowly couple it with the lever process.

So, such kind of pretraining or shaping one can do and it allows us to study now a stimulus. And, then when you present a stimulus animal leans to learn in response to that stimulus and that stimulus alone you need to press the lever and not to something else. So, such kind of training can be nicely done here and then now you can see how many times the animal is responding to the stimulus and how many times, it fails to respond to

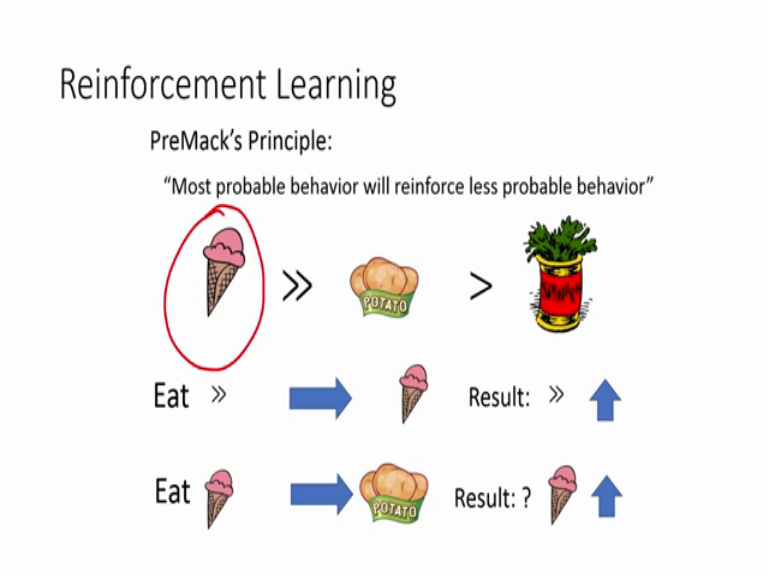
the stimulus and that is our readout. Of how we will the animal has established that association or how we will the learning has occurred.

Alright so, since the animal is learning to operate a lever operate a stimuli that is present in an environment to modulate it is environment, it is also called as an operant the response producing I mean this is response by the animal, that is producing a change in the environment. So, that is why, it is called as an operant conditioning to. So, this Skinner box really helped to make the transition from the very customized puzzle box arrangement conduct made to a laboratory sitting where you can actually go ahead and measure, these and then study these behaviours. We well now that brings to an important point, now what kind of outcomes would really reinforce a behaviour or not reinforce a behaviour right.

So, it is not necessary that you have to give food these are called the primary reinforces, because they very much like our Pavlous US right. They by themselves have a native ability to send in a reward stimuli right, send in act as a reward or a punishment or whichever we are presenting, or designing the experiment.

So, they by themselves can act, apart from this one of the phenomenal discoveries that came out from here is that, there are secondary reinforcers stimuli to start with do not have reinforcing behaviour associated with them. However, can acquire these reinforcement behaviours over a period of time, when they are coupled with primary reinforcers in some or certain ways. So, coming to the bigger question of what is actual I mean, what can actually act as a reinforcement right.

(Refer Slide Time: 23:53)



So, to this PreMack he put forward very very vital and a basic principle, which states that you just look at the behaviour. You do not have to look for anything else; you just look at the response of the animal itself. That is good enough to actually tell you which you can use it as a reinforce, and which you cannot use it as a reinforce. In one line he summarized as most probable behaviour we will in reinforce less probable behaviour, at the start it might look like very circular and confusing.

What do you mean by a most probable behaviour and less probable behaviour? Now, let us take an example here. Let us say we are offered with a 3 different choices of food. An ice cream, a potato, boiled potato, and a spinach. Now, the responses that we would have for these 3 different stimuli would be very different right, most of us would like an ice cream. So, you would tend to prefer the ice cream more than a potato, more than a spinach.

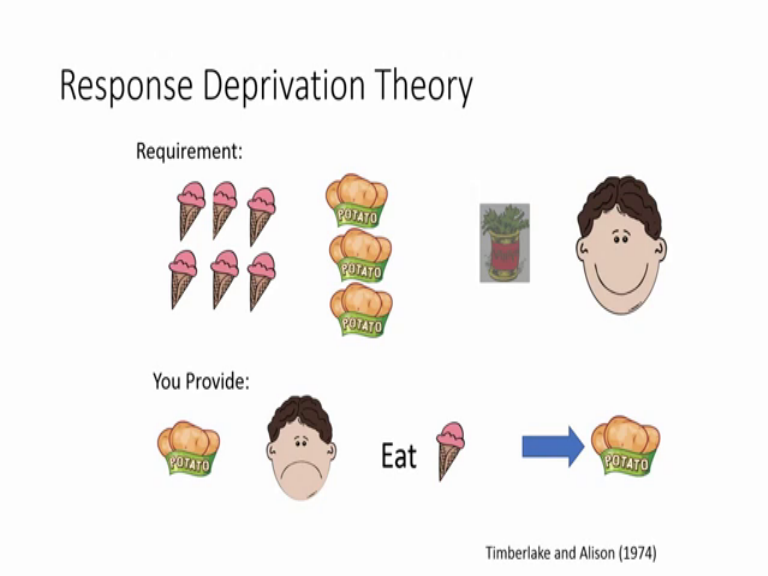
So, what PreMack said is that I can actually make you or me eat spinach by saying hey, if you eat a can of spinach I would present you a ice cream. So, that is what he means by saying most probable behaviour, that is a behaviour that were a response that will elicit a most probable behaviour. That is a stimulus, that will elicit a most probable response, that is in here a stimulus of an ice cream, that is actually eliciting the behaviour or response of consuming it.

You can use that as a reward. For eliciting a response for a stimuli that normally would have been not that favoured right eating a spinach. So, that is pretty simple right. So, what you are saying is that hey look if you tend to eat more of this ice cream. So, you definitely like that ice cream.

However, what I am going to do is that, if you want to eat more of the ice cream then I actually want to you have to eat the spinach, that is how I actually make sure that you eat the spinach more that is very simple no brainer at all, but there is a problem with it. The problem is what happens, if it is kind of intuitive also to say that hey look for you to eat the spinach I am actually offering you more ice cream. So, it is perfectly fine, no problem at all that is understandable. Can I under some circumstance offer or make you eat more of a ice cream offering potato. So, already you are responding high to an ice cream. So, now, that is not intuitive at all.

So, if why would you eat potato or more spinach to get I mean to get more ice cream, you will not do that right. So, that is the idea here right. So, I you have an ice cream which is a highly favourable stimuli and can I make that highly favourable stimuli go more by offering a less favourable stimuli. If I do that then that is kind of saying no to PreMack's principle it turns out you can actually do that.

(Refer Slide Time: 27:32)



The idea here is that we each one of us have a basal line of preference. Let us say that is our example Joe and he likes ice cream by some measure 6 quantities, potatoes by 3

quantities spinach let us not talk about it. So, now, what Timberlake and Alison said is that it is not just the how much you are favouring one food over the other, but it has to be in relation to the basal requirements. Right now, you are saying I would be satisfied with 6 ice creams, while 3 potatoes would be fine for me and 0 spinach it is good.

So, what they would do is that they would deprive you of potatoes; they will give him only 1 potato. So, clearly he is expecting 3 and he is given only 1 potato. In such a case if you say, if you eat more ice cream I will give you more potato, than mister Joe actually starts to eat more ice creams more than 6, that normally he would have consumed to get that potato that he is wanting to get and he is not able to get.

So, that notion of you can actually play around with these stimuli right. These are just stimuli that are rewarding at some different levels. And, you can actually present them in this stimulus response outcome manner by ingenious manner; to alter slightly less flavoured stimulus to get more. It is such kind of phenomena and behaviours, you would know only if you were to able to study the stimulus response outcome behaviour.

And finally, to illustrate the usefulness of that, I will bring in one more example; a very very practical example done in a real world and it is been reported it is a one page paper, I strongly recommend that you read that too.

(Refer Slide Time: 29:38)

This is to do with a Homme Homme et al in 1963. This is to do with how you can actually change the environment or a behaviour in a crèche or a childcare setting. We know as in kid's toddlers they love to run around and not sit in one place. We also know that they like to scribble, play with colours and paint and draw we engage in other activities.

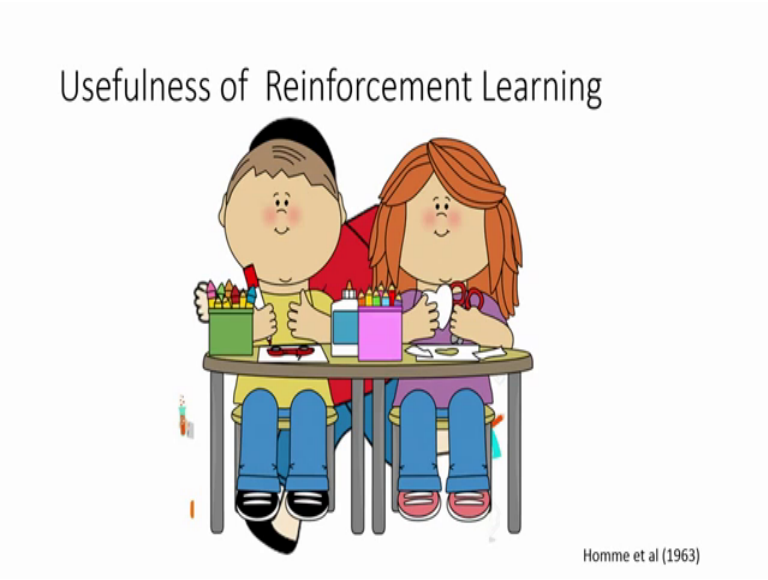
Now, these activities require them to be on a table. If, you are a child caregiver then the idea here is that you somehow want them to be not running around in a complete random chaotic manner, but to bring order into the system make them sit and then do their work on the table. There are 2 ways to go about doing this, one is through fear saying that do. If you do not do it there is a consequence, there will be a punishment and that is what is going to happen.

And, there are undesirable effects to that nobody wants to do that, but we still engage in them. The other way that these people found is that they realized the favourable behaviour for the kids is to run around really. So, they decided to make use of that and, then said hey kids for every for every kid or to every kid, who is not engaging in this activity, but rather sitting in a table alright you do not just chaotically do that, but if you sit in a table and do some activity. For every 5 minutes or some x number of minutes, you engage in that activity you will be given a free run for about 2 minutes or some y minutes.

(Refer Slide Time: 31:42)

So, the idea here is this is the most favourable behaviour right, this is your ice cream here. So, you are actually giving that as a reward and then asking them to go into a behaviour where they can sit in the table and engage, in an activity that is the behaviour that you would like to reinforce.

(Refer Slide Time: 32:26)



So, they are reinforcing with this running and then no wonder in about 3 days' time, the report that all the kids are sitting there completely engaging in the activity. And, then probably that comes a notion of a dedicated period of physical education were you are given or where the kids are given opportunity to run around.

So, I hope with this set of examples; I illustrated the importance of the stimulus response outcome and then some intricacies associated with them, we like to end this lecture with the idea that stimulus response outcome behaviour or a phenomena is very very useful. And, you can \use it to study various aspects of learning. One is that what are all the stimuli that you can use it for reinforcement; how exactly does the learning happen? It does it happen through cognitive or through a reflexive manner.

In the next lecture we will probe little further into it, we will start with the classification of the reinforcement learning itself, and probe little further and then proceed forward and see what theories do we have to explain stimulus response outcome learning.

Thank you.