MICHALE FEE: Today we're going to continue talking about the topic of neural-- recurrent neural networks. And last time, we talked about recurrent neural networks that give gain and suppression in different directions of the neural network space. Today we're going to talk about the topic of neural integrators. And neural integrators are currently an important topic in neuroscience because they are basically one of the most important models of short-term memory.

So let me just say a few words about what short-term memory is. So and to do that, I'll just contrast it with long-term memory. So short-term memory is memory that just lasts a short period of time on the order of seconds to maybe a few tens of seconds at most, whereas long-term memories are on the order of hours, or days, or even up to an entire lifetime of the animal.

A short-term memory has a small capacity, so just a few items at a time you can keep in short-term memory. The typical number would be something like seven, the classic number, sort of seven plus or minus two. You might have heard this, so just about the length of a phone number that you can remember between the time you look it up in the-- well, you know, we all have phone numbers on speed dial now, so we don't even remember phone numbers anymore. But in the old days, you would have to look it up in the phone book and remember it long enough to type it in.

OK, whereas long-term memories have very large capacity, basically everything that you remember about all the work in your classes that you remember, of course, for your entire life, not just until the final exam. Short-term memories are thought to have an underlying biophysical mechanism that is the persistent firing of neurons in a particular population of neurons that's responsible for holding that memory, whereas the biophysical mechanism of long-term memories is thought to be physical changes in the neurons and primarily in the synapses that connect neurons in a population.

So let me just show you a typical short-term memory task that's been used to study neural activity in the brain that's involved in short-term memory. So this is a task that has been studied in nonhuman primates. So the monkey sits in a chair, stares at the screen. There is a set of spots on the screen and a fixation point in the

middle, so the monkey stares at the fixation point.

One of those cues turns on, so one of those spots will change color. The monkey has to maintain fixation at that spot. The cue turns off then. So now the animal has to remember which cue was turned on. And then some delayed period later, which can be-- it's typically between three to six or maybe 10 seconds, the animal-- the fixation cue goes away, and that tells the animal that it's time to then look at the cued location.

And so in this interval between the time when the cue turns off and the animal has to look at the location of that cue, the animal has to remember the direction in which that cue was activated, or it has to remember the location of that cue.

Now, if you record from neurons in parts of the prefrontal cortex during this task, what you find is that the neural activity is fairly quiet during the precue and the cue period and then ramps up. The firing rate ramps up very quickly and maintains a persistent activity during this delay period. And then as soon as the animal makes a saccade to the remembered location, then that neural activity goes away because the task is over and the animal doesn't have to remember that location anymore.

So that persistent activity right there is thought to be the neural basis of the maintenance of that short-term memory. And you can see that the activity of this neuron carries information about which of those cues was actually on. So this particular neuron is most active when it was the cue in the upper-left corner of the screen that was active, and that neuron shows no changes in activity when the cued location shows no change in activity during the memory period, during the delay period when the cued location was down and to the right.

So this neuron carries information about which cue is actually being remembered. And of course, there are different neurons in this population of-- in this part of prefrontal cortex. And each one of those neurons will have a different preferred direction. And so by looking at a population of neurons then during the delay period, you could figure out and the monkey's brain can remember which of those cues was illuminated.

OK, so the idea of short-term memory is that you can have a stimulus that is active briefly. And then for some period of time after that stimulus turns on, there is neural

activity that turns on during the presentation of that stimulus and then stays on. It persists for tens of seconds after the stimulus actually turns off. So that's one notion of short-term memory and how neural activity is involved in producing that memory. And the basic idea here is that that stimulus is in some way integrated by the circuit, and that produces a step in the response. And once that stimulus goes away, then that-- the integral of that stimulus persists for a long time.

All right, now, short-term memory and neural integrators are also thought to be involved in a different kind of behavior. And that is the kind of behavior where you actually need to accumulate information over time. OK, so sometimes when you look at a stimulus, the stimulus can be very noisy. And if you just look at it for a very brief period of time, it can be hard to figure out what's going on in that stimulus.

But if you stare at it for a while, you gradually get a better and better sense of what's going on in that stimulus. And so during that period of time when you're looking at the stimulus, you're accumulating information about what's going on in that stimulus. And so there's a whole field of neuroscience that relates to this issue of accumulating evidence during decision-making. OK, so let me show you an example of what that looks like.

So here's a different kind of task. Here's what it looks like for a monkey doing this task. The monkey fixates at a point. Two targets come up on the screen. The monkey at the end of the task will have to saccade to one or the other of those targets depending on a particular stimulus. And a kind of stimulus that's often used in tasks like this is what's called a "random dot motion stimulus."

So you have dots that appear on the screen. Most of them are just moving randomly, but a small number of them move consistently in one direction. So for example, a small number of these dots move coherently to the right. And if the motion stimulus is more to the right, then the monkey has to then-- once that motion stimulus goes away, the monkey has to make a saccade to the right-hand target.

Now, this task can be very difficult if a small fraction of the dots are moving coherently one way or the other. And so what you can see is that the percentage correct is near chance when the motion strength or the percent coherence, the

fraction of the dots that are moving coherently, is very small. There's almost a-- there's a 50% chance of getting the right answer. But as the motion strength increases, you can see that the monkey's performance gets better and better. And not only does the performance get better, but the reaction time actually gets smaller.

So I'll show-- I found a movie of what this looks like. So this is from another lab that set this up in rats. So here's what this looks like. So the rat is poking its nose in a center port. There's the rat. There's a screen. There's a center port right in front of it that the rat pokes its nose in to initiate a trial. And depending on whether the coherent motion is moving to the right or left, the rat has to get food reward from one or the other port to the left or right. So here's what that looks like.

[VIDEO PLAYBACK]

[BEEP]

[CLINK]

[BEEP]

[CLINK]

[BEEP]

[CLINK]

[CLINK]

[BEEP]

[CLINK]

So this is a fairly high-motion coherent stimulus, so it's pretty easy to see. But and you can see the animal is performing nearly perfectly. It's getting the right-- it's making the right choice nearly every time. But for lower-coherence stimuli, it

becomes much harder, and the animal gets a significant fraction of them wrong.

[END PLAYBACK]

OK, all right, I thought that was kind of amusing. Now, if you record in the brain in-- also in parts of frontal cortex, what you find is that there are neurons. And this is data from the monkey again, and this is from Michael Shadlen's lab, who's now at Columbia. And what you find is that during the presentation of the stimulus here, you can see that there are neurons whose activity ramps up over time as the animal is watching the stimulus.

And so what you can see here is that these different traces, so for example, the green trace and the blue trace here, show what the neurons are doing when the stimulus is very weak. And the yellow trace shows what the neurons do when-- or this particular neuron does when the stimulus is very strong. And so there's this notion that these neurons are integrating the evidence about which way this-- these random dots are going until that activity reaches some sort of threshold.

And so this is what those neurons look like when you line their firing rate up to the time of the saccade. And you can see that all of those different trajectories of neural activity ramp up until they reach a threshold at which point the animal makes it's choice about looking left or right. And so the idea is that these neurons are integrating the evidence until they reach a bound, and then the animal makes a decision.

The weaker the evidence is, the weaker that evidence accumulates. The more weaker the coherence, the more slowly the evidence accumulates and the longer it takes for that neural activity to reach the threshold. And so, therefore, the reaction time is longer. So it's a very powerful model of accumulat-- evidence accumulation during a decision-making task.

Here's another interesting behavior that potentially involves neural integration. So this is navigation by path integration in a species of desert ant. So these animals do something really cool. So they leave their nest, and they forage for food. But they're foraging for food. It's very hot. So they run around. They look for food.

And as soon as they find food, they head straight home. And if you look at their

trajectory from the time they leave food, they immediately head along a vector that points them straight back to their nest. And so it suggests that these animals are actually integrat-- look. The animal's doing all sorts of loop-dee-doos, and it's going all sorts of different directions. You'd think it would get lost. How does it maintain? How does it represent in its brain the knowledge of which direction is actually back to the nest?

One possibility is that it uses external cues to figure this out, like it looks at the-- it sees little sand dunes on the horizon or something. You can actually rule out that it's using sensor information by after the point where it finds food, you pick it up, and you transport it here to a different spot. And the animal heads off in a direction that's exactly the direction that would have taken it back to the nest had it been in the ori-- in the location before you moved it.

So the idea is that somehow it's integrating its distance, and it's doing vector integration of its distance and direction over time. OK, so lots of interesting bits of evidence that the brain does integration for different kinds of interesting behaviors.

So today I'm going to show you some-- another behavior that is thought to involve integration. And it's a simple sensory motor behavior where it's been possible to study the circuitry in detail that's involved in the neural control of that motor behavior. And the behavior is basically the control of eye position. And this group, this was largely work done that was done in David Tank's lab in collaboration with his theoretical collaborators, Mark Goldman and Sebastian's Seung. OK, so let me just show you this little movie.

[VIDEO PLAYBACK]

OK, so these are goldfish. Goldfish have an ocular motor control system that's very similar to that in mammals and in us. You can see that they move their eyes around. They actually make saccades. And if you zoom in on their eye and watch what their eyes do, you can see that they make discrete jumps in the position of the eye. And between those discrete jumps, the eyes are held in a fixed position.

OK, now if you were to anesthetize the eye muscles, the eye would always just sort of spring back to some neutral location. The eye muscles are sort of like springs.

And in the absence of any motor control of any activation of those muscles, the eyes just relax to a neutral position. So when the eye moves and it's maintained at a particular position, that has-- something has to hold that muscle at a particular tension in order to hold the eye at that position.

[END PLAYBACK]

So there are a set of muscles that control eye position. There's a whole set of neural circuits that control the tension in those muscles. And in these experiments, the researchers just focused on the control system for horizontal eye movements, so motion, movement of the eye from a more lateral position to a more medial position or rotation, OK, so eye posi-- horizontal eye position.

And so if you record the position of the eye, and look at-- this is sort of a cartoon representation of what you would see-- you see that the eye stays stable at a particular angle for a while and then makes a jump, stays stable, makes a jump, and stays stable. These are called "fixations," and these are called "saccades."

And if you record from motor neurons that innervate these muscles, so these are motor neurons in the nucleus abducens, you can see that the neural activity is low, the firing rate is low when the eyes are more medial, when the eyes are more forward. And that firing rate is high when the eye is in a more lateral position because these are motor neurons that activate the muscle that pulls the eye more lateral.

Notice that there is a brief burst of activity here at the time when the eye makes a saccade to the-- into the more lateral direction. And there's a brief suppression of activity here when the eye makes a saccade to a more medial position. Those saccades are driven by a set of neurons, by a brain area called "saccade burst generator neurons." And you can see that those neurons generate a burst of activity prior to each one of these saccades.

There are a set of neurons that activate bursts-- activate saccades in the lateral direction, and there are other neurons that activate saccades in the medial direction. And what you see is if you-- is that these saccade burst neurons are actually-- generate activity that's very highly correlated with eye velocity.

So here you can see recording from one of these burst generator neurons generates a burst of spikes that goes up to about 400 hertz and lasts about 100 milliseconds during the saccade. And if you plot eye velocity along with the firing rate of these burst generator neurons, you can see that those are very similar to each other. So these neurons are generating a burst, drives change in the velocity of the neurons of the eye.

OK, so if we go from neurons that have activity that's proportional to position, and we have neurons that have activity that's proportional to velocity, how do we get from velocity? So the idea is that you have burst saccade generator neurons that project to these neurons that project to the muscles. You have to have something in between.

If you have neurons that encode velocity and you have neurons that encode position, you need something to connect those to go from velocity to position. How do you get from velocity to position? If I have a trace of velocity, can you calculate the position by doing what?

AUDIENCE: Integrating.

MICHALE FEE: By integrating. So the idea is that you have a set of neurons here. In fact, there's a part of the brain, and in the goldfish it's called "area one," that take that burst saccade generator neuron burst, integrate it to produce a position signal that then controls eye position. All right, so if you record from one of these integrator neurons while you're watching eye position, here's what that looks like.

[VIDEO PLAYBACK]

And so here's the animal's looking more lateral to the nose. The goldfish's mouth is up here. So that's more lateral. That's moving more medial there, more lateral, more--

[END PLAYBACK]

OK, so this neuron that we were just watching was recorded in this area, area one. Those neurons project to the motor neurons that actually innervate the muscles to control eye position. And they receive inputs from these burst generator neuron.

OK, so if you look at the activity of one of these integrator neurons, that's a spike train during a series of saccades, and fixations is a function of time. This trace shows the average firing rate of that neuron.

This is just smoothed over time, so you're just averaging the firing rate in some window. You can see that the firing rate steps up, that the firing rate jumps up during these saccades and then maintains a stable, persistent firing rate. So the way-- think about this is that this persistent firing right here is maintaining a memory, a short-term memory of where the eye is, and that sends an output that puts the eye at that position.

OK, and so just like we described, we can think of these saccade burst generator neurons as sending an input to an integrator that then produces a step in the position, and then the burst generator input is zero during the [INAUDIBLE]. So the integrator doesn't change when the input is zero. And then there's effectively a negative input that produces decrement in the eye position.

OK, we started talking last time about a neural model that could produce this kind of integration. And I'll just walk through the logic of that again. So our basic model of a neuron is a neuron that has a synaptic input. If we put a brief synaptic input, remember we described how our firing rate model of a neuron will take that input, integrate it briefly, and then the activity, the firing rate of that neuron will decay away.

So we can write down an equation for this single neuron, tau dv/dt is equal to minus v. That's due to this intrinsic decay plus an input. And that input is synaptic input. But what we want, a system where when we put in a brief input, we get a persistent activity instead of a decaying activity. And I should just remind you that we think of this intrinsic decay and this intrinsic leak as having a time constant of order 100 milliseconds.

And I should have pointed out actually that in this system here, these neurons have a persistence of order of tens of seconds. So even in the dark, the goldfish is making saccades to different directions. And when it makes a saccade, that eye position stays stable for-- it can stay stable for many seconds. And if you can do this in humans, you can ask a person to saccade in the dark and try to hold their eyes

steady at a given position, and a person will be able to saccade to a position.

Just you can imagine closing your eyes and saccading to a position. Humans can hold that eye position for about 10 or 20 seconds. So that's sort of the time constant of this integrator in the primate, so that's also consistent with nonhuman primate experiments.

OK, so this has a very long time constant. But we want a neural model that can model that very long time constant of this persistent activity that maintains eye position. All right, but the intrinsic time constant of neurons is about 100 milliseconds. So how do we get from a single neuron that has a time constant of 100 milliseconds to a neural integrator that can have a time constant of tens of seconds?

All right, one way to do that is by making a network that has recurrent connections. And you remember that the simplest kind of recurrent network is a neuron that has an autapse. But more generally, we'll have neurons that connect to other neurons. Those other neurons connect to other neurons.

And there are feedback loops. This neuron connects to that neuron. That neuron connects back, and so on. And so the activity of this neuron can go to other neurons, and then come back, and excite that neuron again, and maintain the activity of that neuron. So we developed a method for analyzing that kind of network by [INAUDIBLE] a recurrent weight matrix, recurrent connection matrix that describes the connections to a neuron A in this network from all the other neurons in the network, A prime, input to neuron A from neuron A prime.

And now we can write down a differential equation for the activity of one of these neurons. dv/dt is minus v that produces this intrinsic decay, plus synaptic input from all the other neurons in the network summed up over all the other neurons plus this external burst input. So how do we make a neural network that looks like an integrator? But how do we do that?

If we want our neuron, the firing rate of our neuron to behave like an integrator of its input, what do we have to do to this equation to make this neuron look like an integrator? So what do we have to do? To make this neuron look like an integrator, it would just be tau dv/dt equals burst input. Right? So in order to make this network

into an integrator, we have to make sure that these two terms sum to zero.

So in other words, the feedback from other neurons in the network back to our neuron has to exactly balance the intrinsic leak of that neuron. Does that makes sense? OK, so let's do that. And when you do that, this is zero.

The sum of those two terms is zero. And now the derivative of the activity of our neuron is just equal to the input. So our neuron now integrates the input. So now the firing rate of our neuron, so there should be a v is equal to 1 over tau, the integral of burst input.

So we talked last time about how you analyze recurrent neural networks. We start with a recurrent weight matrix. So again, these Ms describe the recurrent weights within that network.

We talked about how if M is a symmetric matrix, connection matrix, then we can rewrite the connection matrix as a rotation matrix times a diagonal matrix times a rotation, the inverse rotation matrix, so phi transpose lambda phi where, again, lambda is a diagonal matrix, and phi is a rotation matrix that's [INAUDIBLE] two, in this case, in the case of two-- a two-neuron network, then this rotation matrix has as its columns the two basis vectors that we can now use to rewrite the firing rates of this work in terms of modes of the network.

So we can multiply the firing rate vector of this network times phi transpose to get the firing rates of different modes of that network. And what we're doing is essentially rewriting this recurrent network as set of independent modes, independent neurons, if you will, that described the modes with recurrent connectivity only within that mode. So we're rewriting that network as a set of only autapses. And the diagonal elements of this matrix are just the strength of the recurrent connections within that mode.

All right, so for a network to behave as integrator, most of the eigenvalues should be less than 1, but one eigenvalue should be 1. And in that case, one mode of the network becomes an integrating mode, and all of the other modes of the network have the property that their activity decays away very, very rapidly. So I'm going to go through this in more detail and show you examples.

But for a network to behave as an integrator, you want one integrating mode, one eigenvalue close to 1 and most of the-- all of the other eigenvalues much less than 1. So if you do that, then you have one mode that has the following equation that describes its activity, tau, and let's say that's lambda 1 that has eigenvalue of 1.

So tau dc1/dt, dc/dt equals minus c, that's the intrinsic decay of that mode, plus lambda 1 c1 plus burst input. And if lambda 1 is equal to 1, then those two terms cancel. Then the feedback balances the leak, and that mode becomes an integrating mode.

So when you have a burst input, the activity in that mode increases. It steps up to some new value. And then between the burst inputs, that mode obeys-- the activity of that mode obeys the following differential equation. There's no more burst input between the bursts.

dc/dt is just equal to lambda minus 1 over tau times c1. And if lambda is equal to 1, then this, then dc/dt equals zero, and the activity is constant. Does that makes sense? Any questions about that? Yes, Rebecca.

AUDIENCE: OK, so why does it [INAUDIBLE] need to balance [INAUDIBLE]

MICHALE FEE: Yes, that's exactly right. If this is not true, if, let's say that-- what happens if lambda is less than 1? If lambda is less than 1, then this quantity is negative. So if lambda is 0.5, let's say, then this is 0.5 over tau, minus 0.5 over tau. So dc/dt is some negative constant times c. Which means if c is positive, then dc/dt is negative, and c is decaying.

Does that makes sense? If lambda is bigger than 1, then this constant is positive. So if c is positive, then dc/dt is positive, and c continues to grow. So it's only when lambda equals 1 that dc/dt is zero between the burst inputs.

OK, so let's look at a really simple model where we have two neurons. There's autapse recurrence here, but it's easy to add that. And let's say that the weights between these two neurons are 1. So we can write down the weight matrix. It's just 0, 1; 1, 0 because the diagonals, the diagonals are 0, OK, 0, 1; 1, 0.

The eigenvalue equation looks like this. You know that because the diagonal elements are equal to each other and the off-diagonal elements are equal to each

other because it's a symmetric matrix, then the eigenvalue, the eigenvectors are always what?

AUDIENCE: [INAUDIBLE]

MICHALE FEE: 45 degrees, OK, so 1, 1 and minus 1, 1. So our modes of the network, if we look in this state space of v1 versus v2, the two modes of the network are in the 1, 1 direction and the 1, minus 1 direction. What are the eigenvalues of this network?

OK, so for a matrix like this with equal diagonals and equal off-diagonals, the eigenvalues are just the diagonal elements plus or minus the off-diagonal element. I'll just give you a hint. This is going to be very similar to a problem that you'll have on the final. So if you have any questions, feel free to ask me. OK?

OK, so the eigenvalues are plus or minus 1. They're 1 and minus 1. And it turns out for this case, it's easy to show that the value for this mode is 1, and the eigenvalue for this mode is minus 1. And you can see it. It's pretty intuitive.

This network likes to be active such that both of these neurons are both on. When that neuron's on, it activates that neuron. When that neuron's on, it activates that neuron. And so this network really likes it when both of those neurons are active. And that's the amplifying direction of this network.

And the eigenvalue value is such that the amplification in that direction is large enough that it turns that into an integrating mode. All right, so I'll show you what that looks like. So the eigenvalues again are 1 and minus 1. If you just do that matrix multiplication, you'll see that that's true. lambda is 1, and lambda is minus 1.

You can just read this off. This first eigenvalue here is the eigenvector for the first mode. This eigenvalue is the eigenvalue for that vector for that mode. So here's what this looks like. So this mode is the integrating mode. This mode is a decaying mode because the eigenvalue is much less than 1. And what that means is that no matter where we start on this network, the activity will decay rapidly toward this line. Does that makes sense?

No matter where you start the network, activity in this direction will decay. Any state of this network that's away from this line corresponds to activity of this mode, and

activity of that mode decays away very rapidly. So no matter where you start, the activity will decay to this diagonal line.

So let me just ask one more question. So if we put an input in this direction, what will the network do? So let's turn on an input in this direction and leave it on. What does the network do? Rebecca?

AUDIENCE: [INAUDIBLE]

MICHALE FEE: Good. So we're going to turn it on and leave it on first. The answer you gave is the answer to my next question. The answer is when you put that input on and you turn it off, then the activity goes back to zero. That's exactly right.

But when you put the input-- when you turn the input in this direction on, the network will-- the state will move in this direction and reach a steady state. When you turn the input off, it will decay away back to zero. If we put an input in this direction, what happens?

AUDIENCE: It just keeps going on.

MICHALE FEE: It just keeps integrating. And then we turn the input off. What happens?

AUDIENCE: It [INAUDIBLE]

MICHALE FEE: It stops, and it stays. Because the network activity in this direction is integrating any input that has a projection in this direction. Yes.

AUDIENCE: So [INAUDIBLE] steady state [INAUDIBLE] to F1, so if anything that has any component in the F1 direction will either grow or [INAUDIBLE] over 90 degrees [INAUDIBLE] F1?

MICHALE FEE: Yep.

AUDIENCE: Would it [INAUDIBLE]

MICHALE FEE: Like here? So if you put an input in this direction, what is the component of that input in the integrating direction? If we put an input like this, what-- it has zero component in the integrating direction, and so nothing gets integrated. So you put that input. The network responds. You take the input away, and it goes right back to

zero. If you put an input in this direction, all of that input is in this direction, and so that input just gets integrated by the network. OK?

What happens if you put an input in this direction? Then it has a little bit of-- it has some projection in this direction and some projection in this direction. The network will respond to the input in this direction. But as soon as that input goes away, that will decay away. This, the projection in this direction, will continue to be integrated as long as the input is there.

So let me show you what that looks like. So I'm going to show you what happens when you put an input vertically. What that means, input in this direction means that we have an input to H1. Input to this neuron is 0, but the input to that neuron is 1. That corresponds to H0 being-- H1 direction being 0, and the H2 direction being 1 that has a projection in this direction and this direction.

And here's what the network does. OK, sorry. I forgot which way it was going. So you can see that the network is responding to the input in the H1 direction. But as soon as that input goes away, the activity of the network in this direction goes away as soon as the input goes away. But it's integrating the projection in this direction.

So you can see it continues to integrate. And than you put an input in the opposite direction, it integrates until the input goes away, and it stops there. OK, let me play that again. Does everyone get a sense for what's going on?

So now we have a input that has a projection in the minus F1 direction. And so it's the network is just integrating that negative number. OK, is that clear? OK, all right, so that's a neural integrator. It's that simple. It has one mode that has an eigenvalue of 1. And all of its other modes have small eigenvalues or a negative.

OK, so notice that no matter where you start, this network evolves. As long as there's no input, that network just relaxes to this line, to a state along that line. So that line is what we call an "attractor" of the network. The state of the network is attracted to that line.

Once the state is sitting on that line, it will stay there. So that kind of attractor is called a "line attractor." That distinguishes it from other kinds of attractors that we'll talk in the next lecture. We'll talk about when there are particular points in the state

space that are attractors. OK, no matter where you start the network around that point, the state evolves toward that one point.

OK, so the line of the line attractor corresponds to the direction of the integrator mode, of the [INAUDIBLE] mode. So we can kind of see this attractor in action. If we record from two neurons in this integrator network of the goldfish during this task, if you will, where the [INAUDIBLE] saccades to different directions, so here's what that looks like.

So again, we've got two neurons recorded simultaneously, and we're following the [INAUDIBLE] rate [INAUDIBLE] versus [INAUDIBLE]. And Marvin the Martian here is indicating which way the goldfish is looking [INAUDIBLE]. OK, any questions about that?

So the hypothesis is that-- so I should mention that there-- I didn't say this before. There are about a couple hundred neurons in that nucleus in area one that connect to each other, that contact each other. What's not really known yet-- it's a little hard to prove, but people are working on it.

What's not known yet is whether the connections between those neurons have the right synaptic strength to actually give you lambda, give you an eigenvalue of 1 in that network. So it's still kind of an open question whether this model is exactly correct in describing how that network works. But Tank and others in the field are working on proving that hypothesis.

You can see that one of the challenges of this model for this persistent activity is that in order for this network to maintain persistent activity, that feedback from these other neurons back to this neuron have to be-- have to exactly match the intrinsic decay of that neuron. And if that feedback is too weak, you can see that lambda is slightly less than 1. What happens is that neural activity will decay away rather than being persistent. And if the feedback is too strong, that neural activity will run away, and it will grow exponentially.

So you can actually see evidence of these two pathological cases in neural integrators. So let's see what that kind of mismatch of the feedback would look like in the behavior. So if you have a perfect integrator, you can see that the-- you'll get

saccades. And then the eye position between saccades will be exactly flat.

The eye position will be constant, which means the derivative of eye position will be zero between the saccades. And it will be zero no matter what eye position the animal is holding its eyes. So we can plot the derivative of eye position as a function of eye position, and that should be zero everywhere if the integrator is perfect.

Now, what happens if the integrator is leaky. Now you can see that, in this case, the eye is constantly rolling going back toward zero. So but if the eye is already at zero, then the derivative should be close to zero. If the eye is far away from zero, then the derivative should be-- if the eye position is very positive, you can see that this leak, this leaky integrator corresponds to the derivative being negative.

So if e is positive, then the derivative is negative. If e is negative, then the derivative is positive. And that corresponds to a situation like this. Positive eye position corresponds to negative derivative. And you can see that the equation for the activity of this mode which then translates into eye position is just e to the minus a constant times t.

If you have an unstable integrator, if this lambda is greater than 1, then positive eye positions will produce a positive derivative, and you get x runaway growth of the eye position, and that corresponds to a situation like this-- positive eye position, positive derivative, negative eye position, negative derivative. And then that equation for that situation is either the plus constant times t.

All right, so you can actually produce a leaky integrator in the circuit by injecting a little bit of local anesthetic into part of that nucleus. And so what would that do? You can see that if you inject lidocaine or some other inactivator of neurons into part of that network, it would reduce the feedback connections onto the remaining neurons. And so lambda becomes less than 1, and that produces a leaky integrator when you do that manipulation. So this experiment is consistent with the idea that feedback within that network is required to produce that stable, persistent activity.

Now, you can actually find cases where there are deficits in the ocular motor system that are associated with unstable integration. And this is called congenital nystagmus. So this is a human patient with this condition. And the person is being told to try to fixate at a particular position.

But you can see that what happens is their eyes sort of run away to the edges, to the extremes of eye position. So they can fixate briefly. The integrator kind of runs away, and their eyes run to the edges, to the extremes of the range of eye position. And it's thought that that one hypothesis for what's going on there is that the ocular motor integrator is actually in an unstable configuration, that feedback is too strong.

So exactly how precisely do you need to set that feedback in order to produce a perfect integrator? So you can see that the getting a perfect integrator requires that lambda minus 1 is equal to 0. So lambda is equal to 1. But if lambda is slightly different from 1, we can actually estimate what the time constant of the integrator would be.

So you can see that the time constant is really tau over lambda minus 1, tau over lambda minus 1. So given the intrinsic time constant tau n, you can actually estimate how close lambda has to be to 1 to get a 30-second time constant, OK? And that turns out to be extremely close to 1. In order to go from a 100-millisecond time constant to a 30-second time constant, you need a factor of 300 or, if the neural time constant is even shorter, maybe even 3,000 precision in setting lambda equal to 1.

So this is actually one of the major criticisms of this model, that it can be hard to imagine how you would actually set the feedback in a recurrent network so precisely to get a lambda equal to give you time constants on the order of 30 seconds. Does anybody have any ideas how you might actually do that? What would happen?

Let's imagine what would happen if we were-- we make saccades constantly. We make several saccades per second, not including the little microsaccades that we make all the time. But when we make a saccade, what happens to the image on the retina?

AUDIENCE: [INAUDIBLE]

MICHALE FEE: Yeah, so and if we make a saccade this way, the image on the retina looks like the world is going whoosh, like this. And as soon as it stops and our eyes-- if our

integrator is perfect when the saccade ends, our eyes are at a certain position. What happens to the image on the retina? If our eyes make a saccade, and stop, and stay at a certain position and the velocity is zero, then what happens to the image on the retina? It becomes stationary.

So but if we had a problem with our integrator-- let's say that our integrator was unstable. So we make a saccade in this direction, but our integrator's unstable, so the eyes keep going. Then what would the image on the retina look like if we would have a motion of the image across the retina during the saccade? And then if our eyes kept drifting, the image would keep going. If we had a leaky integrator and we make a saccade, the image of the world could go whoosh, and then it would start relaxing back as the eyes drift back to zero.

So the idea is that when we're walking around making saccades, we have immediately feedback about whether our integrator is working or not. And so, OK, I'm going to skip this. So the idea is that we can use that sensory feedback that's called "retinal slip," image slip, to give feedback about whether the integrator is leaky or unstable and use that feedback to change lambda.

So if we make a saccade this way, the image is going to go like this. And now if that image starts slipping back, what does that mean we want to do? What do we need to do to our integrator, our synapses in our recurrent network if after we make a saccade, the image starts slipping back in the direction that it came from?

We need to strengthen it. That means we have a leaky integrator. We would need to strengthen or make those connections within the integrator network more excitatory. And if we make a saccade this way, the world goes like this and then the image continues to move, it would mean our integrator is unstable. The excitatory connections are too strong. And so we would have a measurement of image slip that would tell us to weaken those connections.

A lot of evidence that this kind of circuitry exists in the brain and that it involves the cerebellum. David Tank and his colleagues set out to test whether this kind of image slip actually controls the recurrent connections or controls the state of the integrator, whether you can use image slip to control whether the integrator network is unstable or leaky, whether that feedback actually controls it. Rebecca.

AUDIENCE: [INAUDIBLE] is the [INAUDIBLE] between slip and overcompensation with [INAUDIBLE] versus unstable integrator, the direction of [INAUDIBLE]

MICHALE FEE: Yes, exactly. So if we make a saccade this way, the world on-- the image on the retina is going to, whoosh, suddenly go this way. But then if the image goes-- OK, in the unstable case, the eyes will keep going, which means the image will keep going this way. So you'll have-- I don't know what sign you want to call that, but here, it's they did a sign flip.

Here the case of decay. So dE/dt is less than zero. That means that the eyes are going back, which means that after you make a saccade, the image goes this way, and then it starts sliding back.

AUDIENCE: So it'll return to--

MICHALE FEE: Return, yeah. So if dE/dt is negative, that means it's leaky. The image slip will be positive. And then you use that positive image slip to increase the weight of the synapses. So you change the synaptic weights in your network by an amount that's proportional to the negative of the derivative of eye position, which is read out as image slip. OK, is that clear?

OK, so they actually did this experiment. So they took a goldfish, head fixed it, put it in this arena. They made a little-- you put a little coil on the fish's eye. So this is a standard procedure for measuring eye position in primates, for example. So you can put a little coil on the eye that measures-- you measure-- OK, so you put a little coil on the eye, and you surround the fish with oscillating magnetic fields.

So you have a big coil outside the fish on this side, another coil on this side, a coil on the top and bottom, and a coil on front and back. And now you run AC current through those coils. And now by measuring how much voltage fluctuation you get in this coil, you can tell what the orientation of that coil is. Does that makes sense?

So now you can read that out here and get a very accurate measurement of eye position. And so now when the fish makes a saccade, you can read out which direction the saccade was. And immediately after the saccade, you can make this spot, so there's a like a disco ball up there that's on a motor that produces spots on the inside of the planetarium.

Notice the fish makes a saccade in this direction. What you do is you make the spots drift back, drift in the direction as though the eyes were sliding back, as though the integrator were leaky. Does that make sense? So you can fool the fish's ocular motor system into thinking that its integrator is leaky.

And what do you think happens? After about 10 minutes of that, you then turn all the lights off. And now the fish's integrator is unstable. So here's what that looks like. There's the spots on the inside. There's the disco ball. That's a overview picture showing the search coils for the eye position measurement system.

And here's the control. That's what the fish-- the eye position looks like as a function of time. So you have saccade, fixation, saccade, fixation. That right there, anybody know what that is? That's the fish blinking. So it blinks.

Give feedback-- OK, here they did it the other way. So they give their feedback as if the network is unstable, and you can make the network leaky. If you give feedback as if the network is leaky, so it makes a saccade, and now you drift the spots in the direction as if the eye were sliding back to neutral position, and now you can make the network unstable. So it makes a saccade, and the eyes continue to move in the direction of the saccade. Saccade, and it runs away. Any questions about that?

So that learning circuit, that circuit that implements that change in the synaptic weights of the integrator circuit, actually involves the cerebellum. There's a whole cerebellar circuit that's involved in learning various parameters of the ocular motor control system that produces these plastic changes. OK, so that's-- are there any questions? Because that's it.

So I'll give you a little summary. So the goldfishes do integrals. There's an integrator network in the brain that takes burst inputs that drive saccades. And the integrator integrates those bursts and produces persistent changes in the activity of these integrator neurons that then drive the eyes to different positions and maintain that eye position. So we've described a neural mechanism, which is this recurrent network, a recurrent network has one eigenvalue that's 1 that produces an integrating mode, and all the other eigenvalues are close to-- are less than 1 or negative.

The model is not very robust if you have to somehow hand-tune all of those [INAUDIBLE] to get a lambda of 1. But there is a mechanism that uses retinal slip to tell whether that eigenvalue is set correctly in the brain and feeds back to adjust that eigenvalue to produce the upper lambda, the proper eigenvalue in that circuit so that it functions as an integrator and using visual feedback.

And I just want to mention again, so I actually got most of these slides from Mark Goldman when he and I actually used to teach an early version of this course. We used to give lectures in each other's courses, and this was his lecture. He later moved to-- he was at Wellesley. So we would go back and forth and give these lectures.

But he moved to Davis. So now I'm giving his lecture myself. And the theoretical work was done by Sebastian Seung and Mark Goldman. The experimental work was done in David Tank's lab in collaboration with Bob Baker at NYU.

OK, and so next time, we're going to-- so today we talked about short-term memory using neural networks as integrators to accumulate information and to perform-- to generate line attractors that can produce a short-term memory of continuously graded variables like eye position. Next time, we're going to talk about using recurrent networks that have eigenvalues greater than 1 as a way of storing short-term discrete memories. And those kinds of networks are called Hopfield networks, and that's what we're going to talk about next time. OK, thank you.