**CRAIG:** Hi, I'm Craig Smith and this is Eye on AI.

**CRAIG:** To start off the new year, I wanted to set the stage for some terrific conversations I have coming to you by going back to some earlier conversations that talk about how we got to where we are in deep learning and how those early threads continue to lead innovation.

**CRAIG:** The threads remind us the deep learning began as an effort to create computer models based on how the brain works. And the main motif is a return to unsupervised or self supervised learning after the field's years long preoccupation with supervised learning in which deep learning models are trained on labelled data.

**CRAIG:** while research into supervised learning methods remain strong, the future is clearly pointing to unsupervised learning, which has already brought us such marvels as GPT-3, Dall-E, and now ChatGP T. In this episode, we revisit conversations with Terry Sejnowski about the development of Boltzmann machines whose sleep wake cycle will make a reappearance in Geoff Hinton's forward forward algorithm the subject of a future episode.

**CRAIG:** Terry explains spike time dependent plasticity in the brain, which is also featured in future research. Jeff himself talked about his search for understanding how the brain works, about contrastive learning and the future of unsupervised methods. Yann LeCun talked about his work in self supervised learning. , which continues and will be the subject of a future episode. And finally, Andrew cautioned that large language models. Now all the rage have certainly gotten us closer to the dream of human level intelligent machines, but that we are only at the very beginning of a long journey .

**CRAIG:** Together the four excerpts give a good foundation for this year's episodes, which will focus on where the field is going from here. I hope you enjoy the conversations as much as I did.

**CRAIG:** Terry Sejnowski starts us off talking about the beginnings of deep learning and how it was inspired by the brain, and also gave neuroscientists insight into what happens in the brain.

**TERRY:** Our goal was to try to solve a problem that had been thought to be impossible to solve, which is trying to take a network with multiple layers. You have an input layer, you have an output layer, and then you have layers in between. In the real brain, those would be the cerebral cortex over the surface of the brain that really is representing the world and all of your plans and actions and, it's the highest level of processing in, in your brain.

**TERRY:** There's an outstanding problem, which was how do you learn in a system that has that complexity with all those layers?

**TERRY:** And it was generally thought because of early work that was done in AI in the sixties, that no one would ever find such a learning algorithm because it was just mathematically too difficult. And that's where Geoff and I invented the Boltzmann machine. And so what is the Boltzmann machine?

**TERRY:** The Boltzmann machine is an architecture that's inspired by physics. And what made it different from all the other architectures at the time that were being looked at was that it was probabilistic. See, the models that were available at the time the input comes in, it goes through a series of stages in a deterministic way. But what we tried to do was to say, look, maybe we can make progress if instead of automatically getting the same output, if the unit itself would have a probability to have an output that varied with the amount of input that you're giving it.

**TERRY:** So more input, probability gets higher, it's gonna produce an output, and if the input is low, you know, you'll still produce an output with a very low probability. And it, it introduces a degree of variability. Not only that, but it creates a different class of network, which is generative. What do I mean by that?

**TERRY:** By that I mean in the traditional input output network, you know, no input, no output, basically, there's nothing going on inside. But in the Boltzmann machine, even without an input, the thing is chugging away because there's always some probability that there'll be an, an output from each unit. And therein lies this, the, the secret that we discover to how you are able to learn in a, a very complex network with many layers, which we now call deep learning.

**TERRY:** And that was by giving the network the input and then keeping track of the activity patterns within the network. And for each connection, you kept track of the, the input and the output, the correlation. But then in order to be able to learn, and this is all mathematical analysis that we had done, you have to basically get rid of the inputs and the outputs and let it free run in a sense, put the network to sleep.

**TERRY:** but it's not off cuz it's chugging away. And, and you, you can do the same measurement for every pair of units with a connection. You keep track of the correlation and we call it the sleep phase. And the learning algorithm is very simple. You subtract the sleep phase correlation from the awake learning phase and that's how you change the strength of the weight.

**TERRY:** Either it goes up or it goes down. And we showed that if you do that and you have a big enough data set that you can learn arbitrary mappings.

**TERRY:** And not only that, and this is the generative part and why this is a much more elegant architecture than the traditional back prop that we now have that is used routinely.

**TERRY:** It's that once you've trained it up, you now can look at the output and say that, that you've trained it to, in order to be able to discriminate between, which at the time is what we were doing, handwritten digits on zip codes. So there's 10 output units. And so you, you give it some input, which is a little handwritten number two.

**TERRY:** And then the unit at the very top, which represents two, is gonna be active at the highest level compared to the other units. And so that was how you classified the digits. But now what you can do is clamp, we called it, which means that you fix the output of the two so that it's the only one that's active and the rest are off.

**TERRY:** And now that percolates down because this network had inputs and outputs going up and down. It was literally a very highly recurrent network. and what it would do is start creating inputs , that looked like twos, but they would be constantly changing. You know, the, the, the loop at the top would come and go, and then the loop at the bottom would come and go and it would, they would wander around.

**TERRY:** And so it was basically dreaming. It was a dreaming about two-ness. And, and the network had created an internal representation of what it meant to be a two.

**CRAIG:** And when you say put it to sleep, you mean stop with the inputs?

**TERRY:** That's right. That's right. In other words, you prevent any input from coming in so that the network could express an input that represented this concept at the highest level.

**TERRY:** And so the information now, instead of flowing from the input to the output, is flowing from the output to the input. And that's what's called a generative network . and now we have even more powerful generative networks. The generative adversarial networks, which are amazing because not only can you generate twos, but you can generate pictures of people's faces.

**TERRY:** You have to give it a bunch of examples, right? You just give it a bunch of examples of rooms like the one we're in and it will start generating new rooms that don't exist with, you know, different kinds of tables and chairs and windows. And they all look real photorealistic.

**TERRY:** And that's what's really astonishing, because we can create very high fidelity models of the world, and in a sense that's what the brain does. Because when we fall asleep and we dream, that's exactly what we're seeing. And we're seeing the generated patterns that are based on our experience.

**TERRY:** CHORD BREAK

**TERRY:** Do you think that it really is an analogy for brain learning during sleep?

**TERRY:** We thought so. Geoff and I were completely convinced we had figured out how the brain worked. In other words, is it just a coincidence that in order to be able to learn in the multilayered network, you had to go asleep?

**TERRY:** Humans to go to sleep every night for eight hours. Why do we go to sleep ? And in fact, and this is really fascinating area because one of the areas that now that I've helped to, to pioneer is trying to really understand what goes on in your brain when you fall asleep. Neuroscientists, and people who are doing computational models like me, have really made a tremendous amount of progress on understanding something about how experiences you have had during the day get integrated into your brain at night.

**TERRY:** It's called memory consolidation. And there's an overwhelming amount of evidence now, both on the psychological side, but also recordings, that this is what's happening. There's something called replay that happens between a part of your brain that's important for memories, episodic memories, things that have happened to you, events, unique objects, something happens to you during the day that's never happened before.

**TERRY:** Right. And you remember it. You need the hippocampus for that. And during the night, the hippocampus plays back literally those experiences to the cortex. And the cortex then has to integrate that into the knowledge base, the semantic knowledge that you have about the world. The Boltzmann machine analogy turned out actually to be a really insight into what's really going on during sleep. But now obviously , what's really going on during sleep is orders of magnitude more complex in terms of the numbers of neurons, the patterns of activity, which we have studied in great detail, but we really think that computationally it's, it's actually what's going on.

**TERRY:** There's a convergence going on right now between our knowledge of the brain on the one hand. And our ability to now create these large scale networks in the image of the brain.

**TERRY:** Not precisely, we're not trying to duplicate the brain, but rather take the principles from the brain and try to build up systems that have some of the capabilities of the brain like vision. Like speech recognition, like language processing. And, and this is really going back and forth now because now neuroscientists are watching what's happening with deep learning and getting inspired and coming up with hypotheses and now going back and testing it in the brain. And as we learn more about the brain, how it solves these problems, we can take that and I'll give you some examples, like attention. As We're looking around we're not just trying to process everything that's out there.

**TERRY:** We focus, right? You focus on a particular object you wanna pick up, you focus on reading, you're reading a sentence, and you're, you're looking for something, right? And that means you have to redirect your attention around, well, it turns out. that if you add attention to these deep learning networks, you vastly improve their performance.

**TERRY:** Having the network decide what's important in a scene like this. In other words, salience, what's, what's important? Where, where should you be looking? Or if, if you're trying to do language translation, a word at the beginning of the sentence, may have a strong relationship with a word later in the sentence.

**TERRY:** And so you wanna be able to hold onto that information, attend to it, while the inputs are coming in, in sequence. And now another word shows up. And those two words have to link together, right? So that's why attention is a way of marking and saying, this is important. Keep it in mind. And then after you've linked up all these words in into a semantic, it's now meaningful representation, you then begin to output words in another language. Again, respecting those relationships between the words, how they're ordered and what their clauses look like. And in German, you have to wait till the end of the sentence in order to put the verb right. The network has to understand that it has to keep track of what the verb is, know what the verb is, and know where to put it.

**TERRY:** And, and this is all something we take for granted, right? That's what our brains are really good at. And so as we learn more about the mechanisms that the brain uses for processing words and also, speech, vision and so forth, these will get incorporated and improve the performance of the networks.

**TERRY:** And it's is now, especially with natural language processing, this has really reached a point as, as you probably know from your cell phone, where it's, it's really good. I mean, you know, speech recognition has gotten amazingly good , even in noisy environments to be able to detect people, and, and now even voicemail is getting transcribed on your phone.

**TERRY:** So this is a whole new era.

**TERRY:** There's overt attention, because we have a fovea and can move it around, we can automatically attend to, with high resolution, a particular object in front of us, and then jump to another object , we're not aware of it, but we're, we're jumping around three times a second, taking in in input and combining them across saccadas, they're called - very fast eye movements - but then there's covert attention. In other words, I could be looking at this and attending to you. And that means that I have the ability in my mind's eye to switch information channels around. And, and both of those are important.

**TERRY:** And going from a camera which has a uniform resolution, to a foveal representation, which you have very high res at the very central few degrees, and then falls off in our eyes very, very rapidly. It's still very sensitive to motion and to other things that you need for alerting you. If there's something coming at you from the side, you, you wanna respond to that quickly.

**TERRY:** But you may not be able to detect what it is with lo rez. But then you could look at it and, and it's interesting. This is another case where the model that we have for computer vision is based on the camera, which is frame based. So when you're taking a video, it's really a sequence of images, and your brain then puts them together into a sequence.

**TERRY:** And so you can see motion and recognize things that are moving. There's a whole new generation now of cameras that are based on how your retina works. Your retina is actually a part of the brain, it's a little pouch in your eye back surface, and through several layers of processing, it then converts the image first into electrical signals and then into spikes.

**TERRY:** The information flowing into the brain, has coded information about things having to do like color motion and other properties for example, in time. How are things changing in time and the relative strength, for example, on an edge where you have a change in contrast that's coded in, in spikes.

**TERRY:** So you have all of that information Now in this train of spikes that is asynchronous. What do I mean by that? Unlike a frame where you collect information over 30 or 40 milliseconds, you can send a spike out any time. And that means you can send out spikes as something occurs in the world with microsecond precision.

**TERRY:** And the relative timing of the spikes carries a lot of information about where things are going. Much more information than if you use a frame based camera. So it turns out that a lot of computer vision is simplified if you use the spike based representation. It's called it a dynamic vision sensor.

**TERRY:** And what's nice about them is that they're very low power cuz they're only putting out these spikes. And it's very sparse in the sense that if nothing's moving, you don't, you actually don't get anything. You have to have motion. And it's very lightweight. It's the perfect thing that you can use, for example, if you want a robot, because power in a robot is very, very expensive.

**TERRY:** And so if you can do vision with spikes instead of supercomputers, the GPUs, which is what being used for deep learning now, it's easier to be autonomous. And, and that's where we're headed. That is to say, edge devices like your cell phone and your watches , they're computers and they're soon gonna have chips , which are deep learning chips, which are very power hungry.

**TERRY:** So, you have to have better batteries. Right. But ultimately, if you could replace the digital circuitry with some of these analog VLSI circuits, like the DVS [Dynamic Vision Sensor] camera, that is going to revolutionize the amount of computing you can do on board, in your hand.

**TERRY:** So the neurons in the brain emit these spikes and, and they're all or none. They last about a millisecond. And they're relatively slow compared to digital electronics. In the sense that they, they are analog, ultimately. The difference between a digital circuit, digital chip has a clock, and on every cycle, every transistor is updated, right?

**TERRY:** So you have to have synchrony across the whole chip. Whereas in one of these analog VLSI chips, it's asynchronous. So every single model neuron can admit a spike whenever it wants. and these are then transferred up the road to other chips through a digital line. So it's a hybrid chip, right? It has a analog processing, which is really cheap and not very accurate, by the way, but that's okay. It turns out if you do a lot of parallel computations with a lot of elements and then integrate that information, it turns out that you're, you're better off. But to communicate between chips, just like the way the brain does, you have to use these, these long distance connections.

**TERRY:** And, and in the case of these analog VLSI chips, you can basically convert it into a digital bus, send the information using some protocol.

 MUSIC INTERLUDE

**TERRY:** eighties, and that was a very exciting time. But once we realized that learning was possible in multilayer networks, then a bunch of other learning algorithms were discovered literally within years. And the one that has been the most popular is the back propagation of error, which requires that you take information about how well you're doing by comparing it to a teacher, a labeled input, and then using that error to go backwards and update the weights as you go down. And, and that was a very efficient, stochastic gradient descent. You're always reducing the error and you can do that very efficiently and very quickly. And because it's so efficient, it's now the way that most of these practical problems are attacked.

**TERRY:** With bigger and bigger and bigger networks, and it's reaching the point now where, you know, the brain has 12 layers in the visual cortex. So now people are dealing with networks that have 200 layers or more. And what's we didn't know back then, and this is the key to success, is that these learning algorithms scale very well.

**TERRY:** What do I mean by that? . So a typical algorithm in AI is able to solve small problems where you have just a few variables that you're trying to find an optimal solution for. Traveling salesman is a good example. You give a bunch of cities and say, what's the fastest route between the cities? So you visit them all once .

**TERRY:** Well, that's called NP Complete. And what that means is that as the number of cities goes up, it becomes exponentially more difficult. at some point it doesn't matter how fast your computer is, he's just gonna saturate it. And that's the problem with many of the algorithms that have been used, that are used in a digital computer with a single processor, which is Von Neumann architecture, with the memory separated from the processor. So it has this bottleneck between the two.. Now, fast forward here we are. The beauty of these neural networks that we pioneered in the eighties was that they're massively parallel. That means that they're simple processors. The memory is located on the processor, they're together so you don't have to ship it back and forth.

**TERRY:** In the brain we have a hundred billion neurons that are working together in parallel, so it means that you can just do much, much more computing in real time, and you don't have to worry about buffers or anything. And as you add more and more neurons to your network and more and more layers, the performance gets better and better and better.

**TERRY:** and that means it scales beautifully. In fact, and this is, this is absolutely amazing , if you have parallel hardware, that is, say if you're simulating each unit at the same time and you're passing the information through, the connection weights at the same time, then it's called order of one scaling.

**TERRY:** That means the amount of time it takes is independent of the number of units you've got. It's fixed. And that's how the brain works. The brain is working in order one, in other words, as the cortex evolved and more and more neurons and in primate brains, especially in human brains, it still works in real time.

**TERRY:** It still works with the same amount of time. In order to come to a conclusion, just to recognize an object, it's about a hundred millisecond. And, you can't get better than that. So nature has found a way to scale up computation . Nature was way ahead of us.

**TERRY:** And now we're just finding that out. And now, hardware has become a really big part of machine learning. And the reason is that up until recently, there was memory chips, there were CPU chips and maybe some digital signal processing chips. But now these machine learning algorithms are being put into silicone .

**TERRY:** Google already has a tensor processing unit, TPU, which does deep learning. But there's a ton of other machine learning algorithms that could be put into silicon and it's gonna vastly improve the amount of computing that you can do because these are like supercomputers now, these chips, in fact, there's one Cerebras, they have a chip that is 20 centimeters across, 400 million processing units.

**TERRY:** So that's getting up to real scale. Of course it's a kilowatt , so you have to have a, a power generator there, but it is scaling up, it's all being scaled up and, and it's a completely new type of chip that people just beginning to appreciate. And, and some of the advantages are , first of all, it's asynchronous and that means you don't need a clock on a chip, right.

**TERRY:** You can just let the whole thing go. Number two, you, you could do with low accuracy. You don't need 64 bit accuracy. You can get by with eight bits. So that means vast savings on memory . . And then there's a high degree of connectivity locally. So that means that, you know, the processors that are near to each other have a lot of information that they're exchanging all the time.

**TERRY:** That's how the brain works too. And you load all the data as it's coming in, just the way it is through your senses. It flows through, it's like a pipeline, right? Information is circulating and decisions are being made. It's a dynamical system. It's, it's a incredibly complex dynamical system, ultimately.

**TERRY:** And we're faced now with an interesting problem, which is we can see how the problem was solved by looking at the input and the output. But what we really wanna know is what's going on inside , what has it learned? And the hottest thing right now is, is, is probing the artificial neural network with the same experiments that neuroscientists do on the brain, where you put your electrode onto one of the units, and then you see what it responds to when it responds. Is it firing before the decision or after? And that, that gives you hints and it tells you a little bit about how the information is flowing through the network.

**TERRY:** And so we're doing that now with these artificial networks, and it's really exciting. we're Using these deep learning networks to create what's called word embeddings. And this is a way for language, for example, a string of words coming in as a sentence to be represented in the semantics space.

**TERRY:** And once you've done that, you can use it for answering questions from articles, it's amazing. It'll answer the question, it will figure out the semantics and what's going on, and it'll answer the question. They went in and they said, well, how was the sentence represented?

**TERRY:** And so what they did is they looked at the pattern of activity. It's in a a million dimensional space. It's huge. And then that collapses into a much lower subspace for a single sentence. And then they look at the graph of how the activity for the different words are represented.

**TERRY:** And it basically, it parses the sentence. It knows what the noun is, it knows where the phrase is. In other words, it, it has learned the structure of, of how syntax is organized in sentences. And it it did that on its own by seeing a lot of sentences. So in the very same network, we have both knowledge of syntax and semantics, just like we have in the brain. The brain doesn't have a semantics box. In a syntax box, right? It's, it's integrating that information . it's all giving you hints about meaning. Which is ultimately what you need as the output.

**TERRY:** You need to be able to answer a question, right? You need to understand what's going on. You, you can't just look at the word order, which is what linguists were doing in the last century. You, you really need to know what the words mean .

**CRAIG:** On modeling the algorithms of the brain. Boltzmann machines, as you described, seem to come close. The field shifted to back propagation because it was so much more efficient, but s pretty clear that the brain doesn't do back propagation

**TERRY:** well, you know, it's doing something that may be equivalent. And now you see this gives us a real strong hypothesis. How could the brain do it?

**TERRY:** You're right, it's not gonna do the same algorithm, but there is information, there are feedback connections, there are more feedback connections than feed forward connections in, in this hierarchy,

**CRAIG:** but

**TERRY:** nobody knows what the information is being carried.

**TERRY:** It's a mystery. And so now it gives us a hypothesis. Let's go in. Maybe that information is giving the earlier layer information about error , how to change the weights. It may not be the backdrop way of doing it, but there may be an equivalent way of doing it.

**CRAIG:** But isn't that happening in the Boltzmann machine as well?

**CRAIG:** You were saying that the information during the sleep period is

**TERRY:** Right. Well, that's an example. Okay. That's an example. Now the Boltzmann machine has another assumption that we make, which is that every pair of units has reciprocal connections with equal strength, which is a pretty strong assumption.

**TERRY:** It's approximately true within the cortex, but it's not exactly true. Because of that it, it is doing the equivalent of back propagation, right? But it's doing it locally. It doesn't need to have the information flowing down . it's all being done at the same time over the whole network.

**TERRY:** And so it may be that the brain is somewhere between a Boltzmann machine and a back prop net, right? And this actually leads to a really exciting new area of research, which is, of all possible computing systems that are parallel, that have this ability to learn and the ability to take in lots of data and be able to classify or predict, we're just scratching the surface here.

**TERRY:** deep learning Is able to do things that are unaccountable. We don't understand how it does so well. Like this language example I was giving you, nobody would've been able to predict that even back in the eighties . if you had asked me, I would've said, well, that, that's unlikely.

**TERRY:** It's too difficult. The language is too difficult. No, it wasn't. And now we have to figure out why.

**CRAIG:** CHORD BREAK

**CRAIG:** Rich Sutton 's temporal difference learning is recognized as the algorithm that the brain is using in reinforcement learning.

**TERRY:** and my lab was the place where that came together.

**TERRY:** So Rich engineered reinforcement learning control theory. I had two postdocs in my lab, Peter Diane, Reid Monague, and we came up with a hypothesis that the dopamine neurons in, in the basal ganglia were computing what's called reward prediction error.

**TERRY:** That's the key thing to temple difference. And, and that has subsequently been tested in monkeys and , with humans, with functional imaging, huge research projects. And now it actually created a field called neuroeconomics, which is how humans make decisions based on the reward prediction error that is coming out of the domain neurons.

**TERRY:** So, he was absolutely right. This is really had , again, feedback from AI directly to neuroscience. And this was back in the nineties when, when we did. Yeah, so bio-inspired, this is what's happening now over and over again . as we understand a little bit more about what works in machine learning, we take that and we go back to the brain and look for it, look for something, not, not the, the details necessarily like the back prop, but we look for the principles.

**TERRY:** The principle is that you have to have information about error somewhere in your system, and you have to get it to the right place at the right time. That, that's the principle now that we're working with in, in neuroscience.

**CRAIG:** Uh, excuse me just a minute. Yes.

**CRAIG:** I'm sorry. Beauregard is new. Um, where was I? Are these algorithms restricted to certain classes of neurons or do algorithms in the brain function across groups of neurons? And could those same groups of neurons be executing other algorithms?

**TERRY:** We know enormous about plasticity.

**TERRY:** Plasticity is the change in the , strength of connection between a neuron at synapses. It's the change and the excitability of neurons. It's the change in thresholds. So all of those variables in neurons are under constant shifts to maximize information flow in order to be able to keep track of information, content coming in that you wanna hold onto.

**TERRY:** And so we know at least 20 algorithms for synaptic plasticity that are used in different parts of the brain for different purposes. So the short answer is that nature has taken advantage of many, many mechanisms. And, and I'll give you, just to give one example. We've been talking about deeper learning.

**TERRY:** Well, that's a model for the cerebral cortex, which is the top of the brain. Well, right under that is the basal ganglia where I was telling you about the dopamine neurons. That's where they live, and that's where temporal difference learning works. So here's two different learning algorithms. They're both in the brain.

**TERRY:** There's a deep learning algorithm of some sort in the cortex, and then there's a temporal difference learning in the basal ganglia. And the basal ganglia talks back to the cortex. There's a big loop there. And we know the basal ganglia is there to learn sequences of actions that take you to a goal or a reward.

**TERRY:** And that could be fire in the future. And that's what temporal difference does for you. And you put these two together and where do you get AlphaGO, right. AlphaGO depended on having a really rich representation of the board. That's deep learning at the same time that it was making decisions about what move to make.

**TERRY:** And that's temporal differences . And it's the talk between those, the crosstalk back and forth that produces magic.

 MUSIC INTERLUDE

**CRAIG:** Geoff Hinton talked about how back propagation of error, which he helped introduced and which has become the standard algorithm for supervised learning is probably not what's happening in the brain and how spike time dependent plasticity, and contrastive learning may be the direction that deep learning researchers need to explore.

**GEOFF:** neuroscientists have been very skeptical about whether the brain can do anything like back propagation. And one of the big problems has been how does the brain communicate gradients? Because in back propagation, you need to change your weight in proportional to the gradient of the error with respect to that weight, whatever your error function is.

**GEOFF:** And the idea is that you represent an error by the rate of change of neural activity. And that's nice cuz it can have both signs that as neural activity can be going up or it can be going down. So you can represent both signs of error. And it also implies that the learning rule , which uses a gradient, is gonna be something called spike time dependent plasticity.

**GEOFF:** That is, When you change a synapse strength, you're gonna change it in proportion to the error derivative, and that means you're gonna want to change it in proportion to the rate of change of the postsynaptic activity. It's gonna be the presynaptic activity times the rate of change of the post activity, and that's called spike time dependent plasticity, which they found in the brain.

**GEOFF:** And in fact, I've been suggesting for a long time that we use activity differences. I had a paper with Jamie McClelland in 1987 suggesting that temporal differences activity be used as error derivatives, and that was actually before spike time dependent plasticity had been discovered. About 2005, I got interested in activity differences again and much more recently people have managed to make that work quite well.

**GEOFF:** I'm still somewhat skeptical. I think the brain could do back prop if it wanted to do back prop. It's a little clumsy and I'm now skeptical cuz I think back prop is too good an algorithm for the brain. So the brain is actually dealing with a very different problem from what most neural nets are dealing with.

**GEOFF:** Most neural nets want to get a lot of knowledge represented in a modest number of parameters, like only a billion parameters, for example. For a brain, that's a tiny number of parameters. That's the number of parameters you have in a cubic millimeter of brain, roughly. So we have trillions and trillions of parameters, but we don't have many training examples.

**GEOFF:** We only live for like a billion seconds or 2 billion seconds, and so we, we don't get much experience and we've got a huge number of parameters and neural nets mostly are in the other regime. They get lots of training and they don't have many parameters. Now if you've got lots and lots of parameters and not much training data, what you want to do is somewhat different from back population, I think.

**GEOFF:** So I got very interested in the idea that that is one way of making this activity difference method work nicely of trying to generate agreement between a top down representation and a bottom up representation. So the idea is you have say some hierarchy of parts. You look at an image, you instantiate parts at different levels, and then from the high level parts, you top down predict the low level parts.

**GEOFF:** And what you'd like to see is agreement between the top down prediction, which depends on a larger context, and the bottom up extraction of a part, which depends on a smaller context. So from some local region of the image you extract apart from many of those parts, you predict a whole from the whole of you now Top down predict the individual parts. But those predictions of the parts have used more information cuz they're based on the whole that got to see more. And what you want is agreement between the top down prediction and the bottom up extraction of a part representation and you want it to be significant agreement.

**GEOFF:** So what you really want is on the same image, they agree, but on different images they disagree. So if you take the parts from one image and the top down, predictions from another image, they should disagree and. That's contrastive learning , but it also suggests a learning algorithm for the brain that is somewhat different from back prop, and I got very excited.

**GEOFF:** It's not quite as efficient as back prop, but it's much easier to put into a brain because you don't need to go backwards through many layers. You just need to compare top down prediction with a bottle up prediction. I call it back relaxation. and over many time steps. It will get information backwards, but it won't get information backwards in one trial.

**GEOFF:** And back propagation sends information all the way backwards through a multilayer net on a single presentation of an image. And back relaxation just gets it back one layer each time and it needs multiple presentations of the same image to get it back all the way. So I got really interested in back relaxation and whether that might explain how the brain was doing this learning in multilayer net.

**GEOFF:** but then I discovered that sort of pure, greedy, bottom up learning did just about as well. I hadn't done the controls carefully enough. And that was a huge disappointment to me. I still want to go back and see if I can make back relaxation work better than greedy bottom up.

**CRAIG:** Is the assumption that, that the brain is so efficient that even if greedy bottom up can do it on its own, that there wouldn't be this top-down function or is it possible that that top-down function exists as an optimizer or, or something?

**GEOFF:** Well, you'd like this top down prediction on making it agree with the bottom up extraction, you'd like that to be better than just training a stack of autoencoders one layer at a time. Otherwise, it's not worth doing.

**GEOFF:** And training a stack of autoencoders is one hidden layer at a time, turns out to be pretty good. Deep learning really got going in about 2006 when we discovered that if you train stacks of auto encoders or restricted Boltzmann machines, one, one hidden layer at a time, and then you fine tune it, it works very well, and that got neural nets going again.

**GEOFF:** People then did things like speech, and vision on ImageNet, where they said, you don't need the pre-training, you don't need to train these stacks of auto codes. You can just train the whole thing supervised. And that was fine for a while, but then when they got even bigger data sets and even bigger networks, people have gone back to this unsupervised pre-training.

**GEOFF:** So that's what Bert is doing. Bert is unsupervised pre-training, and GPT-3 uses unsupervised pre-training and that is important now. So there was this on again, off again, where there was supervised learning, and then I introduced unsupervised pre-training. And then people said, oh, but we don't need that.

**GEOFF:** We just use supervised learning. But now they're back to saying, oh, but we do need some unsupervised learning. Right. But the unsupervised learning algorithms are now getting more sophisticated.

**GEOFF:** CHORD BREAK

**CRAIG:** I had a long conversation with Rich Sutton about temporal difference learning, and there is a view that that algorithm describes what's happening in the the lower brain function and what you're talking about is cortex learning. Are they completely different systems?

**GEOFF:** Different systems? Yes. Big successes of computational neuroscience has been taking the work that Rich Sutton and others did on temporal differences and relating it to experimental studies on the brain and dopamine. Peter Diane in particular was very important in showing the relationship between this theoretical learning algorithm and what's actually going on in the brain.

**GEOFF:** But that's, that's for reinforcement learning, and I think reinforcement learning is kind of the icing on the cake. Most of the learning is gonna be unsupervised learning. You have to learn how the world works and you don't wanna learn how the world works by using reinforcement signals. You don't wanna learn vision by stubbing your toe all the time.

**GEOFF:** You wanna learn to do vision some other way.

**GEOFF:** My main goal in life has been to understand how the brain works and all of this technology that's come out of attempts to understand how the brain works that aren't really how the brain works.

**GEOFF:** It's useful spinoff, but it's not what I was really after .

**GEOFF:** If we can get the idea of top-down predictions and bottom-up predictions agreeing in a contrastive sense. That is, they agree well for the same image and they're very different for different images that may explain how the brain can learn multilayer nets.

**GEOFF:** In computer graphics you represent a house with a particular coordinate frame, and then relatively that coordinate frame, you know where the windows in the door are. Again, that's the kind of representation we need to get into neural nets if neural nets are gonna get more like us at representing objects. A present deep neural nets are very good at doing classification, but they do it a completely different way from people. So they're relying much more on things like texture and they can see all sorts of complex textures that we aren't sensitive to.

**GEOFF:** And that's why you get these adversarial examples where two things look totally different to us, but very similar to a neural net and vice versa.

**GEOFF:** This general trend of extracts and pieces and then get them to interact so you get clearer about what their pieces are, which is what Transformers do. That seems to be a very good way to go about building layers of representation.

**GEOFF:** My main interest has always been unsupervised learning cuz I think that's what most human learning is is.

**GEOFF:** Yann and I share a lot of intuitions. We worked together for a while. Our goals very similar. And the methods are quite similar. So this idea of contrastive representation, learning seems to be very powerful and Yann's exploiting it. Ting Chen made it work really well for static images and we are now trying to. extend that to video. But we're trying to extend you using attention, which is gonna be very important for video cuz you can't possibly process everything in a video at a high resolution.

**GEOFF:** We learn sort of common sense physics, so we learn, you know, if you throw something after it comes down again and if we are good, we learn how to throw a basketball so it goes through the hoop. And that's a very impressive skill cuz you're throwing it from like 20 feet away and you have to get it right to a few inches.

**GEOFF:** That's an amazing thing to be able to do and we. We don't learn that by being told how to do it. We don't learn that using language at all. We learn it from trial and error. But we're understanding how the world works just by observing the world, also by acting in the world. So, just passively observing the world will allow you to understand it, but it's not nearly as good as acting in the world.

**GEOFF:** And in fact, if you think about perception for robots that wander around and act in the world, it changes your view of how perception should work. So if you're just taking images or videos and just passively processing them, it doesn't make you think about attention. , but as soon as you have a robot that's moving around in the world, it's gotta decide what to look at.

**GEOFF:** And the sort of primary question and vision is, where should I look next? And that's been widely ignored by people who just processed static images. Attention is crucial and it's sort of central to how human vision works.

**GEOFF:** CHORD BREAK

**GEOFF:** Let me just say something about supervised learning versus unsupervised learning, because it sounds like a very simple distinction, but actually it's very confusing.

**GEOFF:** So when a kid's mother says that's a cow, we tend to think of it in a machine learning as the mother supplied a label. But what's really happening is this, the child has some sensory input and the child is getting a correlation between the visual sensory input and the auditory sensory input.

**GEOFF:** Now at the top level, the auditory thing gives you the word 'cow.' The visual thing gives you whatever your visual is. Then you learn, they go together, but actually supervision, when you actually get it in reality, it's just another correlation. So it's all about complex correlations in the sensory input, both supervised and unsupervised learning.

**GEOFF:** And then there's correlations with payoffs, and that's reinforcement learning. But I think the correlations with payoffs don't have enough structure in them for you to do most of the learning. So most of the learning's unsupervised.

MUSIC INTERLUDE

**CRAIG:** Yann LeCun talked about the development of convolution neural nets, and how unsupervised learning is the missing link in getting us to higher forms of intelligent machines, explaining his version of self supervised learning.

**YANN:** There was a need for being able to build multilayer neural nets, they just didn't figure out how to do it. And it's probably mostly because they had the wrong neurons. The neurons people were using for neural nets at the time were binary neurons.

**YANN:** And that's incompatible with things like back prop. And so the idea just didn't come up. Even though the basic idea of doing back prop actually existed in the context of optimal control since the sixties.

**YANN:** I started thinking about how can we train multilayer networks? And kind of stumbled on an idea which was very close to back prop. So this must have been 1983 or so. Which was the idea of using the weights that are used in the neural net forward and use them backwards.

**YANN:** I wasn't using them to back propagate gradient. I was using them to back propagate targets. So basically do compute virtual targets for every neuron. My neurons were still binary. And the reason was the computers we had access to at the time were very slow at computing multiplication.

**YANN:** And so if you have binary neurons, you don't need to do multiplication. And then I talked to a friend of mine who was doing a PhD in control in robotics who told me about those methods in optimal control that people come up with in the sixties.

**YANN:** And they said that that looks very much like the stuff I'm working on. And so that's when I came up with backdrop. But then a couple months later, I met Geoff Hinton who came to a meeting in France. And I really wanted to meet him because he had written the paper on Boltzmann machines, which was the first paper I saw that basically allowed to train neural nets that had hidden units.

**YANN:** So I wanna talk to him, either him or Terry Sejnowski I actually met Terry a couple months earlier. And so I meet Geoff, at this meeting in France and we start talking and I tell him what I'm working on. I had a paper in the proceedings of the conference he came tothat talked about this target prop idea and he read it and said, that's really close to back prop. And so we talked together and I told him what I was working on, which was back prop. And then he told me what he was working on, which was also back prop. And and he said, I'm writing a paper and I'm gonna cite you a paper in min e.

**YANN:** I was absolutely delighted by it.

**YANN:** This is mid 1985, so nobody knows about Backdrop yet, but Boltzmann machines have been around for a couple years and it was clear that there's gonna be a lot of people trying to restart working on neural nets.

**YANN:** So, in 1987, I graduated. The backdrop paper had been published a year before.

**YANN:** All of a sudden, the people in France who'd been basically ignoring me for years started talking to me because, I was the local expert on multilayer nets. , so I finished my PhD. Geoff was actually on my thesis committee and I started postdoc with him in Toronto.

**YANN:** So he moved to Toronto from CMU right in the summer, 1987, and I arrived in Toronto two weeks after him.

**YANN:** So the idea of multilayer neural nets is that you can think of the first few layers as extracting features for the following layers to use.

**YANN:** And if you can train the entire thing end to end, that means the system learns its own features. You don't have to engineer the features anymore. They just emerge from the learning. So that, that's what was really appealing to me. And, and the idea that there's necessarily some sort of, hierarchical structure in those features.

**YANN:** And the reason why you need some sort of hierarchy is because the perceptual world, natural data is compositional in the sense that pixels assemble to form edges, for example. Edges assemble to form motifs like corners and crosses and things like this.

**YANN:** And then those motifs assemble to form more complex shapes like circles and squares, and those assemble to form parts of objects and those assemble to form objects, et cetera. So you have this natural compositional hierarchy. And it's the same in speech. You have raw signal and then phones, phonemes, words, sentences, et cetera. You have the same in text, in any natural language, you have this composition or hierarchy because the world is compositional. .

**CRAIG:** And the visual cortex also has these layers.

**YANN:** Yeah. There's anatomical layers on these functional layers.

**YANN:** So here we're talking about the functional layers, . So the visual signal goes from your retina to little piece of the brain at the bottom called lgn, and then it goes to the back of the brain, V one V two, V four, it, and it, the inferior temporal cortex is where object categories are encoded.

**YANN:** And some neurons will fire when you look at a chair, regardless of what chair it is, if it's occluded or not, what type of orientation, what color, it doesn't matter, right? So that's called invariant representation.

**CRAIG:** But there is a hierarchy of components, right?

**CRAIG:** Yeah. There are neurons that fire when it, when it sees a cross or when it sees an edge, right? And then that gradually is built up. .

**YANN:** So this is called the ventral pathway hierarchy, right? V1 V two V four. It, these are the four big, visual cortex areas that are used for recognition of objects in the visual field.

**YANN:** That's only five layers. There's more kind of internal layers if you want. And so the the next question, I asked myself very early on, before even I got to Toronto when I was finishing my PhD, is can we build a network whose architecture would be somewhat inspired by what we know of the visual cortex.

**YANN:** And, a very natural idea, which people already had in the sixties was connecting neurons to a small area in the visual field. So they detect local features and things like this, right?

**YANN:** And I built neural net s like this before even I came up with backdrop that tried to reproduce this kind of architecture with the crude software tools I had available. So what I set out to do when I got in Toronto was I started an ambitious project of writing a neural net simulator .

**YANN:** Basically build one of those kind of visual hierarchical model . That was convolutional nets. So this started working in the spring of 1988 when I was still in Toronto. Did some early experiment there, and then I moved to Bell Labs.

**YANN:** And at Bell Labs they had a big data set of 9,000 training samples of zip code digits. And I tried it, my code was ready. I just tried it on the data set, and within two months I had, better results than everybody else.

**YANN:** So because of this multilayer structure, and in the case of convolutional nets, it exploits this compositional nature of natural signals .

**YANN:** A lot of it was intuition. Some of it was a little bit of biological inspiration. Of course the whole idea of convolution are very classical in signal processing.

**YANN:** CHORD BREAK

**YANN:** There is a limit to what you can apply deep learning to today due to the fact that you need a lot of labelled data. And so it's only economically feasible when you can collect that data and you can actually label it properly.

**YANN:** And that's only true for relatively small number of applications. So that's one mode of training, right? Supervision learning. It works great for categorizing objects in images, for translating from one language to another. If you have lots of parallel text, it works great for speech recognition if you have collected enough data. But it doesn't work for all kinds of stuff.

**YANN:** There is a lot of situations where collecting data is just not the right thing or it is not sufficient. For example, if you wanna train a system to hold a dialogue with someone, you cannot just collect a training set and train the system to hold a dialogue.

**YANN:** You actually have to train it with people talking with people. If you want to train a system to interact with an environment, you have to have an environment in which he can train himself to interact. So that's one problem. The second problem is there is a second type of learning called reinforcement learning, It's a weaker form of learning in a sense that instead of telling the system, here is the correct answer. You only tell the system you are right or you are wrong, or you give it a number that corresponds to how right or wrong you think it is.

**YANN:** And that number can be generated automatically by the environment, so for example, you want to learn to ride a bike if you fall, that's a negative reinforcement. If you keep riding the bike for another second, that's a small, positive, reward so by trying to figure out the sequence of action that maximizes the reward, then you'll learn to ride a bike.

**YANN:** Here's the problem though. Almost any human is capable of learning to drive a car in about 30 hours of training with hardly any supervision. If you were to use reinforcement learning at least in its current form, to, get a car to drive itself, it would have to crash thousands of times.

**YANN:** It would've to drive hundreds of thousands of hours, if not million. Crash thousands of times. Kill mini pedestrians, destroy itself multiple times, runoff cliffs multiple times before it figures out how not to do it. So what that tells you is that we're missing something really essential in human and animal learning that is not reflected in the type of reinforcement learning or supervised learning that our machines can do.

**YANN:** A kid can figure out what an elephant is with's just two pictures, right? We can do this to some extent with learning today using transfer learning, you pre-train the machine with lots of images and then you can retrain it to recognize new objects with various samples.

**YANN:** But there is something we're missing. And one hypothesis that I have, and Geoff has been saying for 30 years or more, is that that thing should be unsupervised learning, which means just learn how the world works. Just learn the dependencies, the structure, the regularity of the world by observing it.

**YANN:** So I have a form of it called self supervised learning, which is a ver y natural idea. You give the machine a piece of input. Let's imagine it's a video clip, for example. You mask a piece of the video clip and you ask the machine, try to predict what is masked from what you're seeing.

**YANN:** So predict the future of this video clip. What's gonna happen in that video? From what you can see from the past? Or here is an image, I'm gonna block a piece of it can you reconstruct that piece? In the context of text, you give it a window of a dozen words and you take out 20% of the words and you ask the system, can you predict what words are missing?

**YANN:** And so when the machine train itself to do this kind of filling in the blanks, it has to develop some representation of the data so it can do this job. So to be able to predict what's gonna happen in the video, you kinda have to understand that there are objects that move independently of backgrounds and there are objects that are animate, others that are inanimate. The inanimate objects have predictable trajectories.

**YANN:** The other ones don't. Things like that. So presumably by training a system to predict or filling in the blanks, it's gonna have to understand a lot about the structure of the world. And so the idea is that you would train a system in this self supervised manner with tons and tons of data.

**YANN:** There's no limit to how many YouTube videos you can make the machine watch. It will distill some representation of the world out of this. And then what you would do is when, whenever a particular task comes in, like learning to drive a car or recognizing particular objects, you use that representation as input to train that classifier supervised.

**YANN:** And in fact, around 2003, 2004, the idea was to use unsupervised learning to pre-train a network and then fine tune it using supervised learning.

**YANN:** Cuz we had this idea that it was very difficult and perhaps hopeless to train a very, very large, very deep network using backprop. It wouldn't work. So the idea was we would pre-train it using unsupervised methods. Until we realized that in fact you could train very deep, very large neural nets, with backdrop from scratch if you had GPUs . So take the, the example of video prediction, you give the machine a video clip. And you ask it, what's gonna happen next? And it cannot possibly predict exactly what's gonna happen next

**YANN:** it will predict the average of all the possible futures. And that ends up being a blurry image.

**YANN:** So, one way to get around this problem is you have an extra variable that you draw randomly.

**YANN:** It's called latent variable. You draw it randomly. And depending on the value that you draw, it's not a single variable, it's gonna be a vector, right? So depending on which values you draw, the prediction is gonna change. And now the name of the game is to train that machine to make predictions so that as you draw different values of the certain variable, the predictions basically go through all the possible futures in the video.

**YANN:** And the problem with this is how can you tell the machine whether it's prediction is good or not? To do that, you have to train a second neural net, and that neural net is trained to tell the difference between a good prediction and a bad prediction. That's called a discriminator or critic, and adversarial generative neural networks is the idea of training those two networks together.

**YANN:** CHORD BREAK

**YANN:** Can we build machines at some point that will be as intelligent as humans in all the tasks that humans are intelligent in?

**YANN:** The answer is of course, there's no question. It's a matter of time. And it's very important to make progress in that direction because we'd like to have machines that have some level of common sense. Because we'd like to be able to build virtual assistants that help people in their daily lives, can answer any question you, you have, can manage your interaction with the digital world and with each other.

**YANN:** We'd like image recognition systems that don't get easily fooled. We'd like, self-driving cars that are very robust and, understand how the world works enough that, they make the right decisions when they see unusual situations..

**YANN:** The question is, how is it that the best of our AI systems have less common sense than a house cat? There is some learning process that animals have access to, to acquire all the knowledge they have about the world that we don't have in our machines.

**YANN:** So one hypothesis is self supervised learning, but there might be other approaches for machines to learn by observation, run without requiring too many labelled samples. Accumulate enough background knowledge by observation that some sort of common sense will emerge.

 MUSIC INTERLUDE

**CRAIG:** Andrew brings a bit of perspective to where deep learning is today and how far we are from the goal human level intelligent machines. That's it for this episode. Please stay tuned for new episodes coming up with Terry, Jeff, and Jann. I'll soon have open AI's Ilia on the podcast and we will follow development of hardware to enno of hardware to enable evermore powerful models.

**YANN:** Can you talk the future of supervised versus unsupervised learning?

**ANDREW:** Yeah, so I think a lot of things go on different time skills. In the near term for the next few years, supervisor Learning will continue to create the bulk of economic value of deep learning.

**ANDREW:** Even though supervised learning is creating the vast majority of the economic value of AI today, I feel like there are many other tools like generative synthesis algorithms like GANs, even reinforcement learning algorithms, and then also the various forms of unsupervised or self supervised learning algorithms.

**ANDREW:** When I started Google Brain many years ago, I actually initially started off betting on unsupervised learning and I was motivated at the time actually by some conversations with Geoff Hinton that made me realize that a lot of human learning is much closer to the unsupervised end of the spectrum.

**ANDREW:** There's a case to be made that a lot of human learning is not supervised learning with parents pointing out every single little thing to you.

**ANDREW:** Back then, and even to this day, I still have that AGI dream of building machines that, maybe a few decades or a few centuries from now, we'll finally get them to be as intelligent as a human.

**ANDREW:** I dunno how long it'll take, but I, I actually do believe that unsupervised learning will play a big role in that. So I think it's a great research topic.

**ANDREW:** I started Google Brain thinking we scale up unsupervised learning algorithms, but then what happened was as we built really big neural networks at Google, we found it was so useful for supervised learning that then kind of distracted us all right, to just focus on supervised learning, but I think unsupervised learning will be, it is a key tool already with word embeddings and large language model and things in computer vision. But I hope that there'll be more research on that.

**CRAIG:** It appeared to a lot of people that the large language models are approaching the kind of general multimodal model that will at some point, approximate human intelligence.

**ANDREW:** While I'm excited by the great research in large language models, I don't think there's a clear shot from large language models to AGI. I feel like while building skyscrapers is very impressive and being at the top of a skyscraper gets you closer to the moon.

**ANDREW:** And the ability to build really tall structures was a key part of how we eventually got to the moon. I think building tall structures was also not the only thing that was needed to get to the moon.

**ANDREW:** Today's large language models have consumed way more text than any human will ever consume in their lifetime, and they are still far less capable than a typical human.

**ANDREW:** Maybe if there was a way to generate even vastly more text than all of humanity has ever generated, and a lot of other things have not even discovered, frankly, and that none of us really know what they are, then in a few decades with new technology that doesn't exist today yet, we'll get to AGI. But that's how I feel about the challenge of the difficulty ahead of us.

**ANDREW:** Once we thought chess was the ultimate problem, and surely once we beat the human world chess champion, isn't that the highest expression of intelligence. Well, that didn't work out.

**ANDREW:** I think that maybe with large language models we're many hundreds of meters closer to the moon now, which is fantastic progress, but the remaining path is still very long.

**CRAIG:** As always, you can find a transcript of this show on our website, e y e hyphen O N a I. And feel free to email us your thoughts. Comments from listeners are always welcome. If you have time, rate and review us on Apple Podcasts, or whichever platform you listen on. The ratings help surface the podcast for others to follow.

**CRAIG:** And remember, the singularity may not be near, but AI is about to change your world, so pay attention.