Interview to John Hopfield by Lex Fridman

Transcript by Alfonso R. Reyes

2020-04-27

Introduction by Lex Fridman

The following is a conversation with John Hopfield, professor of Princeton, whose life's work weave beautifully through Biology, Chemistry, Neuroscience and Physics. Most crucially he saw the messy world of Biology through the piercing eyes of a physicist. He's perhaps best known for his work on associative neural networks, now known as Hopfield Networks, that were one of the early ideas that catalyzed the development of the modern field of deep learning. As his 2019 Franklin medal in Physics award states, he applied concepts of theoretical physics to provide new insights and important biological questions in a variety of areas including Genetics and Neuroscience with significant impact on machine learning. And as John says in his 2018 article titled "Now what?", his accomplishments have often come about by asking that very question "now what?", and often responding by a major change of direction.

Prof. Hopfield received the Benjamin Franklin of Physics 2019 for applying concepts of theoretical physics to provide new insights on important biological questions in a variety of areas, including neuroscience and genetics, with significant impact on machine learning, an area of computer science.

In 1982, Hopfield developed a model of neural networks to explain how memories are recalled by the brain. The Hopfield model explains how systems of neurons interact to produce stable memories and, further, how neuronal systems apply simple processes to complete whole memories based on partial information. The contemporary impact of the `Hopfield model` is evident in fields as diverse as physics, biology, and computer science. By constructing an artificial neural network capable of modeling certain functions of the human brain, machines can now use these processes to store "memories."

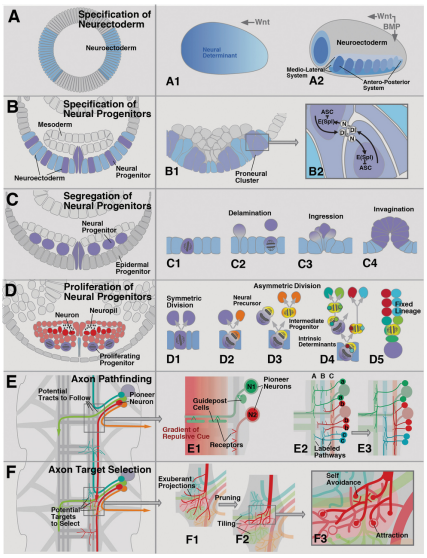


Figure 1: Neural development starts at the embryo phase. The millions of neurons that the human brain requires for thought is reached by a process of proliferation through cell divisions.

Source: Development of the Nervous System of Invertebrates at oxfordhandbooks.com.

Start of Interview

Biological vs Artificial Neural Networks

Lex: What difference between biological neural networks and artificial neural networks is most captivating and profound to you -at the higher philosophical level? Let's not get technical just yet.

John: One of the things very much intrigues me is the fact that neurons have all kinds of components, properties to them. Evolutionary biology, has some little quirks on how a molecule works, of how a cell works, that can be made use of. And evolution will sharpen it up and make it into a useful feature rather than a glitch. So, you expect in neurobiology for evolution to have captured all kinds of possibilities of getting neurons . . . how you get neurons to do things for you. And that aspect has been completely suppressed in artificial neural networks.

[Lex] So, the glitches become features in the biological neural networks?

[John] They can. Let me take one of the things that I used to do research on. If you take things which oscillate, their rhythms, which are sort of close to each other, under some circumstances, these things will have a phase transition, and suddenly because of this rhythm everybody will fall into step. There was a marvelous physical example of that in the Millennium bridge across the Thames River about 2001. Pedestrians walking across. Pedestrians don't walk synchronized, they don't walk in lockstep. But if they're all walking about the same frequency, the bridge could sway at that frequency, and the slight sway

made pedestrians tend a little bit to walk in lockstep. After a while, the bridge was oscillating back and forth and the pedestrians were walking in step to it. You could see it ==[ininteligible]== at the bridge. The engineers made a simple-minded a mistake: they had a feeling that when you walk is step, step, step; it's back and forth motion. But when you walk it's also right foot, left foot, side to side motion. And the side to side motion -for which the bridge was strong enough-, but it wasn't stiff enough. As a result you would feel the motion and you'd fall under step with it. People were very uncomfortable with it. They closed the bridge for two years while adding stiffening for it.

Now, nerve cells produce action potentials. You have a bunch of cells, which are loosely coupled together, producing action potentials at the same rate. There'll be some circumstances under which these things can lock together; other circumstances which they won't. Well, if they fire together you can be sure the other cells are going to notice it. So, you could make a computational feature out of this in an evolving brain. Most artificial neural networks don't even have action potentials, let alone have the possibility for synchronizing them.

[Lex] You mentioned the evolutionary process. The evolutionary process that builds on top of biological systems leverages the weird mess of it, somehow? So, how do you make sense of that ability to leverage all the different kinds of complexities in the biological brain?

[John] Well . . . Look. At the biological molecular level you have a piece of DNA which encodes a particular protein. You could duplicate that piece of DNA, and now one part of it encodes that protein. But the other one could itself change a little bit and then start coding for a molecule which is just slightly different.

Now, if that molecule was just slightly different has a function which helped any old chemical reaction which was important to the cell, you would go ahead and let it try an evolution slowly and improve that function. So, you have the possibility of duplicating, and then having things drift apart; one of them retain the old function, the other one does something new for you. And there's evolutionary pressure to improve. There is in computers too. But improvement has to do with closing some companies, opening some others. The evolutionary process looks a little different.

[Lex] Similar timescale perhaps . . .

[John] Much shorter in times still . . .

[Lex] Companies close, go bankrupt, and are born. Yeah, shorter but not much shorter. Some companies lasts for centuries. You're right. If you think of companies as a single organism that builds and all . . . it's a fascinating dual correspondence there between biological . . .

[John] And companies have difficulty having a new product competing with an old product. When IBM built this first PC, -you probably read the book- they made a little isolated internal unit to make the PC. And for the first time in IBM's history they didn't insist that you build it out of IBM components. But

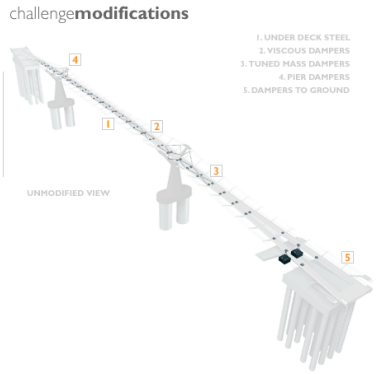


Figure 2: Viscous passive dampers were installed at several locations at The Millenium Bridge in London to limit dynamic excitation by pedestrians walking in synchronization with the bridge movement.

Source: Archived web page of the Millenium Bridge project. Video of oscillating bridge here.

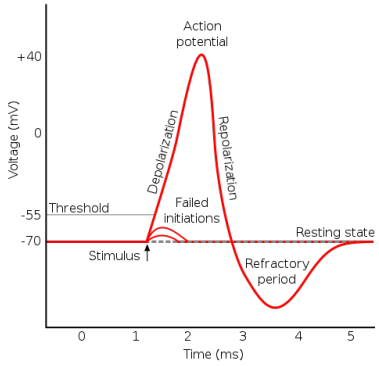


Figure 3: Action potential, the brief (about one-thousandth of a second) reversal of electric polarization of the membrane of a nerve cell (neuron) or muscle cell.

Source: Generation of Action Potentials at teachmephysiology.com, and Action potential at britannica.com.

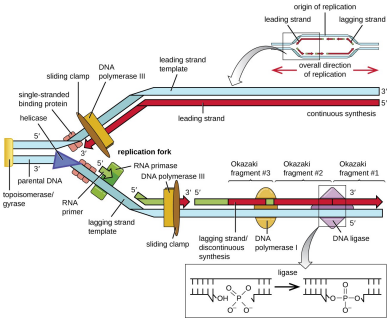


Figure 4: Construction of new DNA molecules involves two replication forks formed by the opening of the double-stranded DNA at the origin, the parental DNA.

Source: DNA Replication at lumenlearning.com.

they understood that they could get into this market which was a very different thing by completely changing their culture. And biology finds other markets in a more adaptive way.

[Lex] It's better at it. It's better at that kind of integration. So, maybe you've already said it, but what would be the most beautiful aspect or mechanism of the human mind? Is it the adaptive . . . , the ability to adapt, as you've described? Or there's some other little quirk that you particularly like?

[John] Adaptation is everything when you get down to it.

But there are differences between adaptation, where your learning goes on generations over generations in evolutionary time, or your learning goes on at the timescale of one individual who must learn from the environment during that individual's lifetime. And biology has both kinds of learning in it. The thing which makes neurobiology hard is that a mathematical system that were built on this other kind of evolutionary system.

10:02

[Lex] What do you mean by mathematical system? Where's the math in the biology?

[John] Well, when you talk to a computer scientist about neural networks it's all math. The fact that biology actually came about from evolution, and the fact that biology is about a system which you can build in three dimensions.

If you look at computer chips, computer chips are basically two dimensional structures; a 2.1 dimensions. They really have difficulty doing three-dimensional wiring. Biology in the neocortex is actually also sheet-like and it sits on top of the white matter, which is about ten times the volume of the gray matter, that contains all what you might call the wires. But there's is a huge . . . the effect of computer structure on what is easy and what is hard is immense. And biology does . . . makes some things easy that are very difficult to understand how to do computationally. On the other hand, you can't do simple floating-point arithmetic because it's awfully stupid.

[Lex] You're saying this kind of three dimensional complicated structure mix . . . it's still math; it's still doing math? The kind of math is doing enables you to solve problems of a very different kind?

[John] That's right, that's right.

[Lex] So, you mentioned two kinds of adaptation: the evolutionary adaptation and the adaptation in learning at the scale of a single human life. Which is particularly beautiful to you, and interesting? From a research and from just a human perspective? And which is more powerful?

[John] I find things most interesting that I begin to see how to get into the edges of them and tease them apart a little bit, see how they work. And since I can't see the evolutionary process going on, I am in awe of it but I find it just a black hole as far as trying to understand what to do. And so in a certain sense I'm in awe of it but I couldn't be interested in working on it.

[Lex] The human life timescale is however a thing you can tease apart and study?

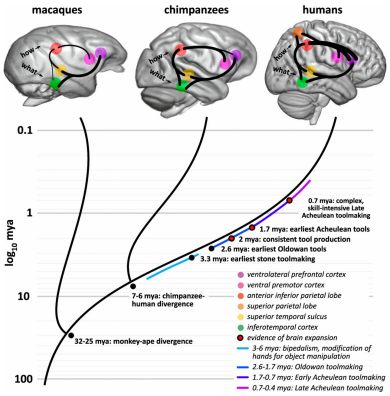


Figure 5: Biological evolution of greater intelligence in human individuals has promoted innovation and allowed mastery of more complex concepts and skills.

Source: Evolutionary neuroscience of cumulative culture at Proceedings of the National Academy of Sciences (PNAS).

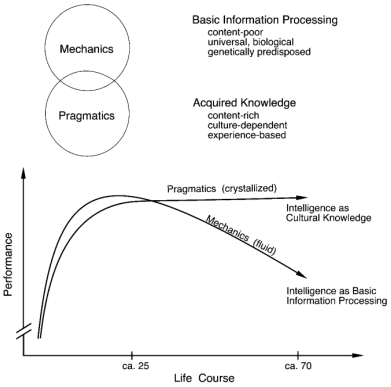


Figure 6: Lifespan theories of cognitive development posit two-component models of cognition. The top section defines the categories, the bottom section illustrates postulated lifespan trajectories.

Source: Cognitive Development at sciencedirect.com.

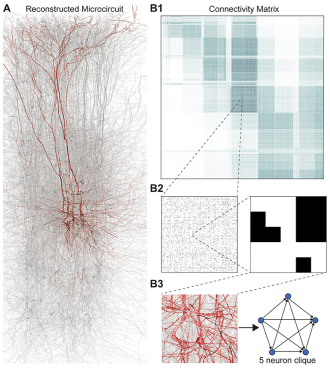


Figure 7: digital reconstructions of rat neocortical microcircuitry that closely resemble the biological tissue in terms of the numbers, types, and densities of neurons and their synaptic connectivity.

Source: Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function at frontiersin.org.

[John] Yes, you can do it. There's developmental neurobiology which understands all of these connections. Now, the structure evolves from a combination of what the genetics is like and the real; the fact is you're building a system in three dimensions.

[Lex] In just days and months, those early days of human life are really interesting . . .

[John] They are. And of course there are times of immense cell multiplication. There are also times of the craziest cell death in the brain; it's during infancy.

[Lex] Turnover . . .

[John] What is not effective, what is not wired well enough to use at the moment. is thrown out.

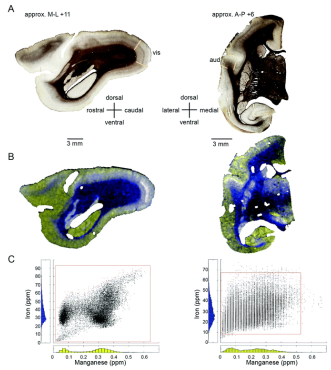


Figure 8: Stained brain sections in marmosets indicating the primary visual cortex and the primary auditory cortex.

Source: Whole-brain metallomic analysis of the common marmoset (Callithrix jacchus) at pubs.rsc.org.

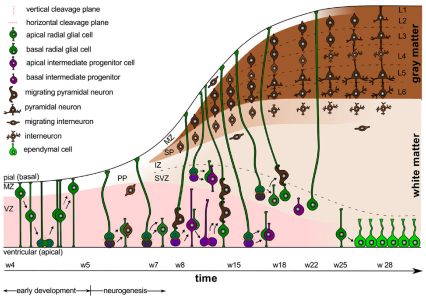


Figure 9: Timeline of early development and neurogenesis including cell division and cell migration in the human brain.

Source: Physical biology of human brain development at frontiersin.org.

Neural Networks lack Understanding

[Lex] It's a mysterious process. Let me ask: from what field do you think the biggest breakthroughs in understanding the mind will come in the next decades? Is it neuroscience, computer science, neurobiology, psychology, physics, maybe math, maybe literature? [Laughter]

[John] Well, of course, I see the world always through a lens of physics. I grew up in physics and the way I pick problems is very characteristic of physics and of an intellectual background which is not psychology, which is not chemistry, and so on, and so on.

[Lex] Both of your parents were physicists . . .

[John] Both of my parents were physicists. And the real thing I gathered was a feeling that the world is an understandable place, and if you do enough experiments, and think about what they mean, and structure things so that you can do the mathematics of the relevant, of the experiments, you also be able to understand how things work.

[Lex] But that was a few years ago. Did you change your mind at all? Through many decades of trying to understand the mind? Of studying in different kinds of . . . not even the minds, just biological systems. You still have hope the physics that you can understand?

[John] There's the question of what do you mean by "understand"?

[Lex] Of course . . .

[John] When I taught freshman physics I used to say "I wanted them to understand the subject, to understand Newton laws." I didn't want them simply to memorize a set of examples to which they knew the equations to write down, to generate the answers. I had this nebulous idea of understanding. So, if you looked at a situation you can say "oh, I expect the ball to make that trajectory. All right, I expect." So, I'm into a notion of understanding. I don't know how to express that very well; I've never known how to express it well, and you run smack up against it. Look at these simple neural nets, feed-forward neural nets, which do amazing things, and yet you know contain nothing of the

essence of what I would have felt was understanding. Understanding is more than just an enormous lookup table.

[Lex] Let's linger on that. How sure you are of that? What if the table gets really big? So, asked in another way, these feed-forward neural networks, do you think they'll ever understand?

[John] I will answer that in two ways. I think if you look at real systems, feedback is an essential aspect of how these real systems compute. On the other hand, if I have a mathematical system with feedback, I know I can unlay this, undo it in parts of it. But I have an exponential expansion and the amount of stuff I have to build so I could resolve the problem that way.

[Lex] So, feedback is essential. We can talk even about recurrent neural nets. Do you think all the pieces are there to achieve understanding? Through these simple mechanisms like, back to our original question, what is the fundamental . . . is there a fundamental difference between artificial neural networks and biological? Or is it just a bunch of surface stuff?

[John] Suppose you ask a neurosurgeon when does somebody's dead. He'll probably go back to saying "well, I can look at the brain rhythms and tell you this is a brain which never could have functioned again. This is another but this other one is one which if we treat it well is still recoverable." And then just do that by so many electrodes looking at simple electrical patterns. Just don't look in any detail at all or what individual neurons are doing. These rhythms are utterly absent from anything which goes on in Google.

[Lex] But the rhythms?

[John] But the rhythms, what?

[Lex] It's like . . . you're comparing the greatest classical musician in the world to a child first learning to play. The question I'm at - but they're still both playing the piano - I'm asking, will it ever go on at Google? Do you have a hope? Because you're one of the seminal figures in both launching both disciplines, both sides of the river . . .

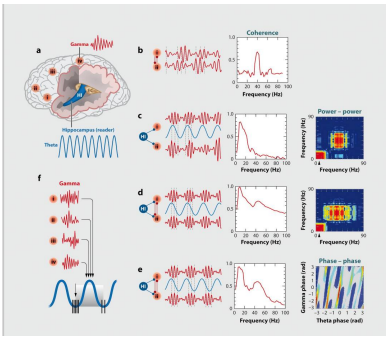


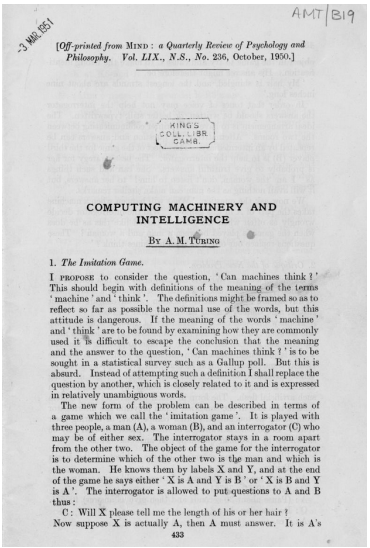
Figure 10: Brain rythms and transient neuronal oscillations at different regions.

Source: Brain rhythms and neural syntax: implications for efficient coding of cognitive content and neuropsychiatric disease at ncbi.nlm.nih.gov.

A Timeline for Artificial Intelligence

[John] I think it's going to go on generation after generation the way it has, where you might call the AI computer science community, and says: "let's take the following: this is our model of neurobiology at the moment. Let's pretend it's good enough and do everything we can with it". And it does interesting things, and after a while sort of grinds into the sand, and you say "Oh something else is needed from neurobiology". And some other grand thing comes in and enables you to go a lot further. What was going on ==[ininteligible]== It can be generations of this evolution. I don't know how many of them. And each one is going to get you further into what a brain does.

In some sense, passes the Turing test longer and in more broad aspects. And how many of these are good there are going to have to be before you say "I've made something, I've made a human", I don't know.


AMT/B19

[*Off-printed from* MIND : *a Quarterly Review of Psychology and Philosophy. Vol. LIX., N.S., No.* 236, October, 1950.]

KING'S COLL. LIBR. CAMB.

COMPUTING MACHINERY AND INTELLIGENCE

By A. M. Turing

1. *The Imitation Game.*

I propose to consider the question, 'Can machines think ?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think ?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The new form of the problem can be described in terms of a game which we call the 'imitation game'. It is played with three people, a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either 'X is A and Y is B' or 'X is B and Y is A'. The interrogator is allowed to put questions to A and B thus :

C : Will X please tell me the length of his or her hair ?

Now suppose X is actually A, then A must answer. It is A's

433

Figure 11: Alan Turing's paper 'Computing Machinery and Intelligence'.

Source: The Turing Digital Archive at turingarchive.org.

20:15

[Lex] But your sense is it might be a couple . . .

[John] My sense is might be a couple more. And going back to my brain waves of the word. From the AI point of view, they would say "ah, maybe these are epi-phenomena and not important at all." The first car I had ==[ininteligible]== a 1936 Dodge. It could go 45 miles an hour and the wheels were shimmering. Good speedometer. Now, ==[ininteligible]== design the cars that way, the cars now ==[ininteligible]== function to have that. But in biology, it would be useful to know when are you going more than 45 miles an hour, you just capture that and you wouldn't worry about where it came from. It'll be a long time before that kind of thing which can take place in large complex networks of things is actually used in the computation. Look, how many transistors are there in a laptop these days?

[Lex] Actually, I don't know the number . . .

[John] it's on a scale of 10 to the 10 [1], I can't remember the number either. All the transistors are somewhat similar and most physical systems with that many parts, all of which are similar, have collective properties.

Sound waves in air; earthquakes, what have you, have collective properties. Weather. There are no collective properties used in artificial neural networks, in AI. If biology uses them it's gonna take us some more generations ==[ininteligible]== before people actually dig in and see how they are used, what they mean.

[Lex] You're very right. It might have to return several times to neurobiology and try to make our transistors more messy . . .

[John] Yeah, yeah. At the same time the simple ones will conquer big aspects, and I think one of the most . . . , biggest surprises to me was how well learning systems, which are manifesting non-biological, how important they can be actually, and how important, how useful they can be in AI.

[1] In fact, a CPU in a laptop has above 10^9 transisistors

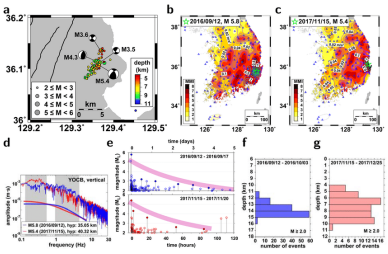


Figure 12: Earthquake collective properties.

Source: Time-advanced occurrence of moderate-size earthquakes... at researchgate.net.

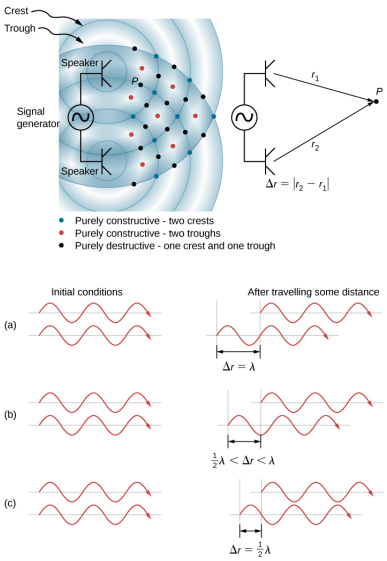


Figure 13: Sound waves produced by a speaker showing collective properties such as phase, frequency, interferance, wavelentgh, and amplitude.

Source: Normal Modes of a Standing Sound Wave at courses.lumenlearning.com.

Hopfield Networks and Associative Memory

[Lex] If we can just take a stroll to some of your work, that is incredibly surprising that it works as well as it does that launched a lot of the recent work with neural networks, if we go to what are now called Hopfield Networks, can you tell me what is associative memory in the mind for the human side? Let's explore memory for a bit.

[John] What you mean by associative memory is: "oh, you have a memory of each of your friends. your friend has all kinds of properties from what they look like, whether voice sounds like, where they went to college, where you met them, go on and on, what science papers they've written." If I start talking about a five foot ten weary cognitive scientist that's got a very bad back. It doesn't take very long for you to say "are you talking about Geoff Hinton". I never mentioned the name, or anything very particular, but somehow a few facts are associated with this, with a particular person, enables you to get a

hold of the rest of the facts, or of another subset of of them. It's this the ability to link things together, link experiences together, which goes under the general name of associative memory, and a large part of intelligent behavior is actually just large associative memories at work, as far as I can see.

[Lex] What do you think is the mechanism of how it works in the mind? Is it is it a mystery to you still? Do you have inklings of how this essential thing for cognition works?

[John] What I made 35 years ago was, of course, a crude physics model to show the kind . . . actually, enable you to understand my old sense of understanding as a physicist because you could say "ah, I understand why this goes to stable state. It's like things going down downhill." And that gives you something with which to think in physical terms rather than only in mathematical terms.

[Lex] So, you've created these associative artificial . . .

[John] That's right. And now, if you if you look at what I did, I didn't at all describe a system which gracefully learns. I described it as a system in which you could understand how things, how learning could link things together, how very crudely it might learn. One of the things which intrigues me, as I reinvestigate that system no,w to some extent is "Look, I see you every second for the next hour, or what have you. Each look at you is a little bit different. I don't store all those second-by-second images, I don't store 3,000 images. I somehow compact this information. So, now I have a view of you which I can use. It doesn't slavishly remember anything in particular but it could pack the information in useful chunks which are . . . Somehow, it's these chunks, which are not just activities of neurons; bigger things than that which are the real ==[ininteligible]== which are useful to you.

[Lex] Useful to you to describe, to compress this information?

[John] I just compressed it in such a way that if I get . . . the information comes in just like this again, I don't bother about how to rewrite it. Or efforts to rewrite it simply do not yield anything because those things are already written. And that needs to be not . . . look this up, it has started somewhere already. It has to be something which is much more automatic in the machine hardware.

[Lex] Right. So, in the human mind how complicated is that process, do you think? You created . . .- it feels weird to be sitting with John Hopfield calling them Hopfield Networks . . .

[John] It is weird . . .

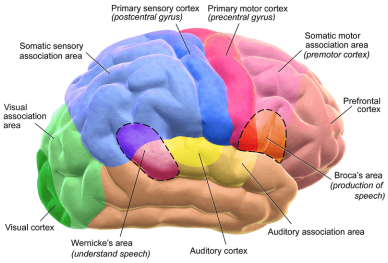


Figure 14: Associative memories are what allow individuals to make certain connections and inferences even when they're not clearly explained or spelled out.

Source: An Overview Of Associative Memory at betterhelp.com.

Neural Networks aren't Dynamical Systems

[Lex] Yeah. But nevertheless that's what everyone calls them. So, here we are. So, that's a simplification. That's what a physicists would do. You and Richard Feynman sat down and talked about associative memory. Now, if you look at the mind . . . you can't quite simplify it so perfectly . . .

[John] Let me backtrack just a little bit. Biology is about dynamical sys-

tems. Computers are dynamical systems. If you want to ==[ininteligible]== neurobiology. What is the time scale? there's a dynamical system in which you have a fairly fast timescale in which . . . the synapses don't change much during this computation, so I'll think of the synapses are fixed, and just do the dynamics of the activity. Or, you can say the synapses are changing fast enough that I have to have the synaptic dynamics working at the same time as the system dynamics in order to understand the biology. If you look at the feed-forward of artificial neural nets they're all done as learning. First of all. I spent some time learning and not performing, then I turned off learning and I perform . . .

[Lex] Right

[John] That's not biology. As I look more deeply at neurobiology, even as associative memory, I've got to face the fact that the dynamics of a synapse change is going on all the time, and I can't just get by by saying "I'll do the dynamics of the activity with a fixed synapses."

[Lex] So, the dynamics of the synapses, is actually fundamental to the whole system?

30:02

[John] Yes. And there's nothing necessarily separating the time scales. The time scales can be separate. And it's neat for the physicists - of the mathematicians point of view-, but it's not necessarily true in neurobiology.

Hopfield Networks and Learning

[Lex] You're kind of dancing beautifully between showing a lot of respect to physics, and then also saying that physics cannot quite reach the complexity of biology. So, where do you land, or do you continuously dance between the two?

[John] I continuously dance between them because my whole notion of understanding is that you can describe to somebody else how something works in ways which are honest and believable, and still not describing all the nuts and bolts in detail. Weather. I can describe weather as 10 to the 32 molecules (10^32) colliding in the atmosphere, I can simulate weather that way, or have a big enough machine to simulate it accurately. It's no good for understanding but I just want to understand things. I want to understand things in terms of wind patterns, hurricanes, pressure differentials, and so on; all things as were collective. And the physicist in me always hopes that biology will have some things which can be said about it which was both true and for which you don't need all the molecular details of the molecules colliding. That's what I mean from the roots of physics by understanding.

[Lex] How did Hopfield Networks help you understand what insight to give us about memory, about learning?

[John] They didn't give insights about learning. They gave insights about how things having learned could be expressed. How having learned a picture -of a picture of you- reminds me of your name. That didn't describe a reason-

able way of actually doing the learning. Only says if it had previously learned the connections of this kind of pattern would now be able to behave in a physical way ==[ininteligible]== part of the pattern in here, the other part of the pattern will complete over here. I can understand that physics if the right learning stuff had already been put in, and you couldn't understand why then putting in a picture of somebody else would generate something else over here. But it did not have a reasonable description of the learning process.

[Lex] But even . . . forget learning, I mean that's just a powerful concept that sort of forming representations that are useful to be robust, you know, for error correction kind of thing. So, this is kind of what the biology does we're talking about?

[John] What my paper did was simply enable you. There are lots of ways of being robust. If you think of it a dynamical system here, you think of a system where a path is going on and in time, and if you think for a computer is a computational path which is going out in a huge dimensional space of ones and zeros. And an error-correcting system is a system which if you get a little bit off that trajectory will push you back onto that trajectory again till you get to the same answer in spite of the fact that there were things though that the computation wasn't being ideally done all the way along a line. There are lots of models for error correction but one of the models for error correction is to say: there's a valley that you're following flowing down, and if you push a little bit off the valley, it's just like water being pushed a little bit by a rock gets back and follows the course of the river, and then basically the analog in the physical system just enables you to say "oh, yes error free computation and an associative memory are very much like things that I can understand from the point of view of a physical system." The physical system can be under some circumstances an accurate metaphor. It's not the only metaphor. There are error correction schemes which don't have a valley and energy behind them but those are correction schemes such a mathematician may be able to understand but I don't.

[Lex] So, there's a the physical metaphor that seems to work here?

[John] That's right

Boltzmann Machines

[Lex] So, these kinds of networks actually led to a lot of the work that is going on now in neural networks, artificial neural networks. So, the follow-on work with restrictive Boltzmann Machines and Deep Belief Nets, followed on from the from these ideas of the Hopfield network. What do you think about this continued progress of that work towards now reinvigorated exploration of feed-forward neural networks and recurrent neural networks, convolutional neural networks, and kinds of networks that are helping solve image recognition, natural language processing, all that kind of stuff?

[John] it's always intrigued me one of the most long-lived of the learning

systems is the Boltzmann Machine, which is intrinsically a feedback network. And was the brilliance of Hinton and Sejnowski to understand how to do learning in that. It's still a useful way to understand learning, and the learning that you understand has something to do with the way that feed-forward systems work. But it's not always exactly simple to express that intuition. It always amuses me, as Geoff Hinton keeps going back to that, well yet again, on a form of the Boltzmann machine because really . . . which has feedback and interesting probabilities in it. This is a lovely encapsulation of something in computational.

[Lex] Something computational?

[John] Something both computational and physical. Computational in that very much related to feed-forward networks. Physical in that Boltzmann machine learning is really learning a set of parameters for physics Hamiltonian or Energy function.

[Lex] What do you think about learning in this whole domain. Do you think the aforementioned Geoff Hinton all the work there with backpropagation, all the kind of learning that goes on in these networks . . . How do you . . . if we compared to learning in the brain, for example, is there echoes of the same kind of power that backpropagation reveals about these kinds of recurrent networks? Or is it something fundamentally different going on in the brain?

[John] I don't think the brain is as deep as the deepest networks go; the deepest computer science networks. I do wonder whether they're part of that depth, of the computer science networks, is necessitated by the fact that the only learning is easily done on a machine is feed-forward. So, there's the question of to what extent has the biology, which has some feed-forward and some feedback, been captured by something which got many more neurons but much more depth to the neurons in it.

[Lex] So part of you wonders if the feedback is actually more essential than the number of neurons or the depth -the dynamics of the feedback?

[John] The dynamics of the feedback . . . if you don't have feedback it's a little bit like a building a big computer and running it up through one clock cycle, and then you can't do anything until you reload something coming in. How do you use the fact that there are multiple clocks? How do I use the fact that you can close your eyes, stop listening to me, and think about a chessboard for two minutes without any input whatsoever?

On Consciousness

40:00

[Lex] Yeah, that memory thing. That's fundamentally a feedback kind of mechanism. You're going back to something. Yes, it's hard to understand; hard to introspect. Let alone consciousness . . .

[John] Oh, let alone consciousness . . .

[Lex] Yes. Because that's tied up in there too. You can't just put that on

another shelf . . .

[John] Every once in a while like I get interested in consciousness and then I go -and I've done that for years-, and ask one of my betters what's their view on consciousness. It's interesting collecting them.

[Lex] What's consciousness? Let's try to take a brief step into that room . . .

[John] Well, that's Marvin Minsky; his view on consciousness. And Marvin said "consciousness is basically overrated". It may be an epi-phenomenon, after all, all the things your brain does . . . they're actually hard computations, you do not do consciously. And there's so much evidence that even the things . . . , the simple things you do, you can make decisions, you can make committed decisions about them. The neurobiologist can say "He's now committed, he's going to move the hand left", before you know it.

[Lex] So his view that consciousness is not . . . that's just like little icing on the cake. The real cake is in the subconscious?

[John] Yes, yes. Subconscious non-conscious.

[Lex] That's the better word there

[John] it's only the Freud captured the other word

[Lex] Yeah. It's that's a confusing word subconscious

[John] Nicholas Chater wrote an interesting book, I think its title is "The mind is flat". Flat, in a neural net sense, flat is something which is of very broad neural net without anything other than the layers in depth, or as a deep brain would be many layers and not so broad. In the same sense that if you push Minsky hard enough he would talk to you and say "consciousness is your effort to explain to yourself that what you have already done."

[Lex] Yeah, it's the weaving of the narrative around the things that already been computed for you . . .

[John] That's right. And then, so much of what we do for our memories of events. For example, if there's some traumatic event you witness, you will have a few facts about it correctly done. If somebody asks you about it you will weave a narrative which is actually much more rich in detail than that. Based on some anchor points you have of correct things and pulling together general knowledge on the other but you will have a narrative and once you generate that narrative you are very likely to repeat that narrative and claim that all the things you have hidden are actually the correct things. There was a marvelous example of that in the Watergate / impeachment era of John Dean. John Dean -you're too young to know-, had been the personal lawyer of Nixon. And John Dean was involved in the cover-up. John Dean ultimately realized the only way to keep himself out of jail for a long time was actually to tell some of the truths about Nixon. John Dean was a tremendous witness; he would remember these conversations in great detail, very convincingly detail. Long afterward some of the tapes, the secret cases from where John Dean was recalling these conversations, were published. And one found out that John Dean had a good but not exceptional memory. What he had was an ability to paint vividly, and in some sense accurately, the tone of what was going on.

[Lex] By the way, that's a beautiful description of consciousness. Where do you stand in today -perhaps has changed day to day-, but where do you stand on the importance of consciousness in our whole big mess of cognition? Is it just a little narrative maker, or is it actually fundamental to intelligence?

[John] That's a very hard one. I asked Francis Crick about consciousness. He launched forward a long monologue about handling the peas, and how Mendel knew that there was something, and how biologists understood there was something in inheritance which was just very, very different, and that the effect that inherited traits didn't just wash out into a grave where this or this propagated. That was absolutely fundamental in biology and it took generations of biologists to understand that there was genetics. And it took another generation or two to understand that genetics came from DNA. But very shortly after Mendel, thinking biologists did realize that there was a deep problem about inheritance. And Francis, in all likelihood, would have said "that's why I'm we're working on consciousness." But of course he didn't have any smoking gun in the sense of Mendel. And that's the weakness of his position. If you read this book which he wrote with Koch, I think . . .

[Lex] Yeah, Christof Koch

[John] I find it unconvincing for this first smoking gun reason. Start going on and collecting views without actually having taken a very strong one myself because I haven't seen the entry point. Not seeing the smoking gun and the point of view of physics, I don't see the entry point. Whereas the neurobiologist, once they understood the idea of a collective and evolutional dynamics, which could be described as a collective phenomenon, I thought "Ah, there's a point where I know about physics", it is so different from any neurobiologist that I have something that I might be able to contribute.

[Lex] Right now there's no way to grasp at consciousness from a physics perspective?

[John] From my point of view that's correct. And of course people . . . this is like everybody else. You think very but broadly about things you have. The closest related question is about free will. Your belief you have free will. Physicists will give an offhand answer and then backtrack, backtrack, backtrack, when they realize that the answer they gave must fundamentally contradict the laws of physics.

[Lex] Answering questions of freewill and consciousness naturally lead to contradictions from a physics perspective. It eventually ends up with quantum mechanics, and then you get into that whole mess of trying to understand how much from a physics perspective, how much is determined -already predetermined-, much is already deterministic about our universe, there's lots of difference . . .

[John] And if you don't push quite that far you can say "essentially all of neurobiology, which is relevant, it can be captured by classical equations of motion." Because in my view of the mysteries of the brain are not the mysteries of quantum mechanics but the mysteries of what can happen when you have

a dynamical system, a driven system, with 10 to the 14 (10^14) parts. The bare complexity is something which is . . . , the physical complex system is at least as badly understood as the physics of phase coherence and quantum mechanics.

Attractor Networks

[Lex] Can we go there for a second? You've talked about attractor networks, and just maybe you could say what are attractor networks? And, more broadly, what are interesting network dynamics that emerge in these or other complex systems?

[John] You have to be willing to think in a huge number of dimensions. In a huge number of dimensions the behavior of a system can be thought of as just the motion of the point over time in those huge number of dimensions. An attractor network is simply a network where there is a line and other lines converge on it in time. That's the essence of an attractor network that's how you . . .

[Lex] In a highly dimensional space . . .

[John] And the easiest way to get that is to do it in a high dimensional space where some of these dimensions provide the dissipation ==[ininteligible]==. In a physical system, trajectories can contract everywhere - they have to contract in some places and expand in others. There was a fundamental classical theorem in statistical mechanics which goes under the name of Liouville's theorem which says "you can't contract everywhere; if you contract somewhere, you have to expand somewhere else." In interesting physical systems you get driven systems where you have a small subsystem which is the interesting part, and the rest of the contraction of an expansion, the physicists say it's the entropy flow in this other part of the system. But basically attractor networks are dynamics funneling down . . . if you start somewhere in the dynamical system you will soon find yourself on a pretty well determined pathway which goes somewhere; if you start somewhere else, you'll wind up on a different pathway. I don't have just all possible things, you have some defined pathways which are allowed and under which you will converge. And that's the way you make a stable computer, and that's the way you make a stable behavior.

51:06

[Lex] So, in general, looking at the physics of the emergent stability in networks, what are some interesting characteristics that .., what are some interesting insights from studying the dynamics of such high dimensional systems?

[John] Most dynamical systems, driven dynamical systems, where driven means they're are coupled to an energy source, their dynamics keeps going because of its coupling to the energy source. In most of them, it's very difficult to understand at all what the dynamical behavior is going to be . . .

[Lex] You have to run it . . .

[John] You have to run it. There's this subset of systems which has a clean tone known to mathematicians as the Lyapunov functions. And those systems you can understand convergent dynamics by saying you're going downhill on something or other. And that's what I found without ever knowing what the Lyapunov functions were in the simple model I made in the early 80s; it was an energy function so you could understand how you get this channeling under pathways without having to follow the dynamics in an infinite detail. You started rolling a ball as off of a mountain that's gonna wind up at the bottom of a valley you know that it's true without actually watching the ball fall roll down

[Lex] There's certain properties of the system that when you can know that?

[John] That's right and not all systems behave that way . . .

[Lex] Most don't

[John] Most don't. But it provides you with the metaphor for thinking about systems which are stable enough to have these attractors behave. Even if you can't find the Lyapunov function behind them, or an energy function behind them. That gives you a metaphor for thought.

Neural Networks not a Biological System

[Lex] Speaking of thought, if I had a glint in my eye with excitement and said: "you know I'm really excited about this something called deep learning and neural networks, and I would like to create an intelligent system." And came to you as an adviser, what would you recommend? Is it a hopeless pursuit these neural networks, what kind of mechanism should we explore what kind of ideas should we explore?

[John] Well, you look at this as the simple network for ==[ininteligible]== networks. They don't support multiple hypotheses very well. As I have tried to work with very simple systems which do something which you might consider to be thinking. Thought has to do with the ability to do mental exploration before you make it, take a physical action.

[Lex] Almost they're like we were mentioning playing chess visualizing, simulating inside your head different outcomes . . .

[John] Yeah. And you could do that as a feed-forward network because you've pre-calculated all kinds of things. But I think the way neurobiology does it hasn't pre-calculated everything; it actually has parts of a dynamical system in which you're doing exploration in a way which is . . .

[Lex] There's a creative element . . .

[John] There's a creative element. And in a simple-minded neural net you have a constellation of instances from which you've learned. And if you are within that space, if there is a new question and the question is within this space you can actually rely on that system pretty well. Come up with a good suggestion for what to do. If on the other hand, the query comes from outside the space, you have no way of knowing how the system is going to behave; there are no limitations on what could happen. With the artificial neural net-

work is always very much . . . I have a a population of examples; the test set must be drawn from the equivalent population. If the test set has examples which are from a population which is completely different, there's no way that you could expect to get the answer right.

[Lex] What they call outside the distribution?

[John] That's right, that's right. And if you see a ball rolling across the streets at dusk, if that wasn't in your training set, the idea that a child may be coming close behind that is not going to occur with the neural net.

[Lex] There's something in your biology that allows that . . .

[John] There's something in the way of what it means to be outside of the population, of the training set, the ==[ininteligible]== is that the training set isn't just ==[ininteligible]== set of examples. There's more to it than that. It gets back to my own question of where's is it to understand something.

Physics and Data

[Lex] You know is in a small tangent, you've talked about the value of thinking of deductive reasoning in science versus large data collection. So, sort of thinking about the problem but I suppose it's the physics side of you of going back to first principles and thinking, but what do you think is the value of deductive reasoning in in a scientific process?

[John] There obviously scientific questions in which the route to the answer to it come through the analysis of a hell of a lot of data

[Lex] Right. Cosmology and that kind of stuff . . .

[John] ==[ininteligible]== never written the kind of problem in which I've had any particular insight. Though I would say if you look at cosmology, it is was one of those. If you look at the actual things that Jim Peebles, one of this year's Nobel Prize, ==[ininteligible]== the kinds of things he's done. He's never crunched large data never, never, never. He's used the encapsulation of the work of others in this regard.

[Lex] But ultimately boils down to thinking through the problem, like what are the principles under which a particular phenomena operates?

[John] Look. Physics is always going to look for ways in which you can describe the system, in which rises above the details and ==[ininteligible]==. Biology works because of the details. And physics, to the physicists, we want an explanation, which is right in spite of the details. And they will leave questions which we cannot answer as physicists because the answer cannot be found that way.

Brain-Machine Interfaces

60:00

[Lex] If you're familiar with the entire field of brain-computer interfaces. It has become more and more intensely researched and developed recently.

Especially with companies like NeuraLink with Elon Musk

[John] I know they've always been the interested both in things like getting the eyes to be able to control things, or getting the thought patterns to be able to move what had been a connected limb which is now connected through a computer.

[Lex] That's right. So, in the case of NeuraLink they're doing thousand-plus connections where they're able to do two-way: activate and read spikes, neural spikes. Do you have hope for that kind of computer-brain interaction in the near -or maybe even- far future? Of being able to expand the ability of the mind of cognition, or understand the mind?

[John] This is as watching things go. When I first became interested in neurobiology most of the practitioners thought you would be able to understand neurobiology by techniques which allowed you to record only one cell at a time. People like David Hubel, very strongly reflected that point of view. And that's been taken over by a generation, a couple of generations later, by a set of people who says: "Not until we can record from 10 to the 4 (10^4), or 10 to the 5 (10^5) at a time, where we actually be able to understand how the brain actually works." And in a general sense, I think that's right. You have to look you have to begin to be able to look for the collective modes, collective operations of things. It doesn't rely on this action potential or death of cells; it relies on the collective properties of this set of cells connected to this kind of patterns, and so on. And you're not going to see did the thing what those collective activities are without recording many cells at once.

[Lex] And the question is how many at once what's the threshold?

[John] The motor cortex does something which is complex and yet with the problem you're trying to address is very simple. Now, neurobiology does it in ways that's different from the way an engineer would do it. An engineer would put in six highly accurate stepping motors controlling a limb rather than 100,000 muscle fibers, each of which has to be individually controlled. So, understanding how to do things, in a way which is much more forgiving and much more neural, I think would benefit the engineering world. ==The engineering world: touch. That's where their pressure sensor or to let vary them== an array of of a gazillion pressure sensors none of which are accurate, all of which are perpetually recalibrating themselves.

[Lex] You're saying your hope, your advice for the engineers of the future is to the embrace the large chaos of a messy error-prone system like those of the biological systems? Like that's probably the way to solve some of these challenges?

[John] I think you'll be able to make better computations towards robotics that way than by trying to force things into a robotics, where joint motors are powerful and stepping motors are accurate.

Equations as the Confluence of Biology and Physics

[Lex] But then the physicist in you will be lost forever in such systems because there's no simple fundamentals to exploring systems that are so large . . .

[John] Well. . ., there's a lot of physics. The Navier-Stokes equations: the equations of non-linear hydrodynamics; huge amount of physics in them. All the physics of atoms and molecules has been lost but they have been replaced by this other set of equations which is just as true as the equations that ==[ininteligible]== them. Those equations are going to be harder to find in neural biology but the physicist in me says there are probably some equations of that sort.

[Lex] They're out there . . .

[John] They're out there. And if the physics is going to contribute anything it may contribute to trying to find out what those equations are and how to capture them from biology

[Lex] Would you say that's one of the main open problems of our age is to discover those equations?

[John] Yeah. If you look at as molecules and psychological behavior. These two are somehow related. There are layers of detail, there are layers of collectiveness. And to capture that at some vague way, several stages on the way up to see how these things that can actually be linked together.

65:00

[Lex] So, it seems in our universe there's a lot of elegant equations that can describe the fundamental way that things behave -which is a surprise. I mean it's compressible into equations: it's simple and beautiful. But there is still an open question whether that link is equally between molecules and the brain is equally compressible into elegant equations. But you're both a physicist and a dreamer. You have a sense that . . .

[John] Yes. But I can only dream physics dreams. There was an interesting book called "Einstein's Dreams", which alternates between chapters on his life and descriptions of the way time might have been but isn't. As linking between these things, of course, ideas that Einstein might have had to think about the essence of time as he was thinking about time.

A Digital Version of Immortality

[Lex] So, speaking of the essence of time in neurobiology, you're one human, famous impactful human, but just one human with a brain, living the human condition. But you're ultimately mortal like all of us. Has studying the mind as a mechanism changes the way you think about your own mortality?

[John] It has, really. Because as particularly as you get older in the body comes apart in various ways, I became much more aware of the fact that what if somebody is contained in the brain and not in the body that you worry about burying. And it is to a certain extent true that for people who write things down:

equations, dreams, notepads, diaries. Fractions of their thought does continue to live after they're dead and gone -after their body is dead and gone. And there's a sea change in there going on in my lifetime between - when my father died - when, except for the things that were actually written by him, there were very few facts about him that had been recorded. And the number of facts which are recorded about each and every one of us, forever, -now, as far as I can see-, in the digital world. And so the whole question of "what is death?" It may be different for people a generation ago and in a generation ahead.

[Lex] Maybe we have become immortal under some definition?

[John] Yeah, yeah.

On the meaning of life

[Lex] Last easy question: what is the meaning of life? Looking back, you studied the mind, as weird descendants of apes? What's the meaning of our existence on this little Earth?

[John] Oh. The word "meaning" is as slippery as the word "understand" . . .

[Lex] Interconnected somehow perhaps. Is there -it's slippery-, but is there something you, despite being slippery, can hold long enough to express?

[John] I've been amazed at how hard it is to define things in a living system. In the sense that one hydrogen atom is pretty much like another. One bacterium is not so much like another bacterium, even of the same nominal species. In fact, the whole notion of what is a species gets a little bit fuzzy. And species exists in the absence of certain classes of environments. And pretty soon one winds up with the biology, which the whole thing is living. What if there's actually any element of it, which by itself would be said to be living? It becomes a little bit vague in my mind.

70:00

[Lex] So, in a sense the idea of meaning is something that's possessed by an individual, like a conscious creature. And you're saying that it's all interconnected in some kind of way that there might not even be an individual. We're all kind of this complicated mess of biological systems at all different levels. Where the human starts and when the human ends is unclear?

[John] Yeah. As in neurobiology where you say the neocortex does the thinking but there's lots of things that are done in the spinal cord. And so we say, where's the essence of thought? It's just going to be the neocortex? It can't be, it can't be.

[Lex] Yeah, maybe to understand and to build thought you have to build the universe along with the neocortex; it's all interlinked through the spinal cord. John is a huge honor talking today. Thank you so much for your time, I really appreciate it.

[John] Well thank you for the challenge of talking with you and the interesting to see whether you can ==[ininteligible]== with 5 minutes out of coherent sense to anywhere.

[Lex] Beautiful!

End of Interview

Thanks for listening to this conversation with John Hopfield. And thank you to our presenting sponsor CashApp. Download it, used LexPodcast you'll get ten dollars, and ten dollars will go to FIRST, an organization that inspires and educates young minds to become science and technology innovators of tomorrow. if you enjoyed this podcast subscribe in YouTube, give it five stars in Apple podcast, support on Patreon, or simply connect with me on Twitter at @LexFridman.

Epilogue

And now let me leave you with some words of wisdom from John Hopfield in his article titled "Now What".

> "Choosing problems is the primary determinant of what one accomplishes in science. I have generally had a relatively short attention span on science problems, thus, i have always been on the lookout for more interesting questions either as my present ones get worked out, or as it get classified by me as intractable given my particular talents.""

He then goes on to say:

> "What I have done in science relies entirely on experimental and theoretical studies by experts. I have a great respect for them. Especially for those who are willing to attempt communication with someone who is not an expert in the field. I would only add that experts are good at answering questions. If you're brash enough, ask your own. Don't worry too much about how you found them.""

[Lex] Thank you for listening and hope to see you next time.

References

Links

- Podcast: https://lexfridman.com/john-hopfield/
- Video: https://www.youtube.com/watch?v=DKyzcbNr8WE&t=12s
- Transcript in GitHub: https://github.com/f0nzie/transcript_interview_john_hopfield_by_lex_fridman
- Transcript in LinkedIn: https://www.linkedin.com/pulse/transcript-interview-john-hopfield-lex-fridman-alfons
- Tweet: https://twitter.com/OilGains/status/1249787240076775431?s=20

Include anything in the margin.

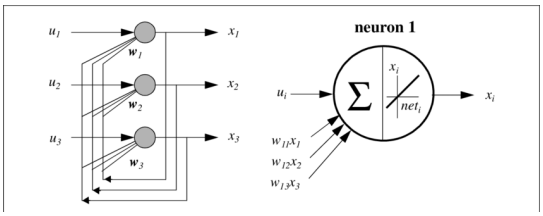


Figure 15: Topology of Hopfield networks with 3 neurons.

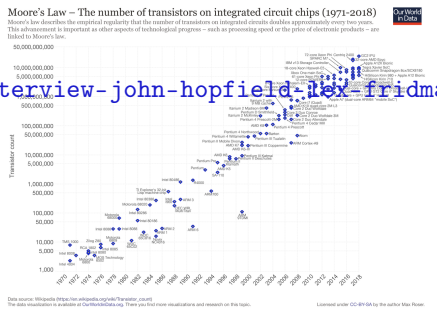


Figure 16: Moore's law and number of transistors.

Source: Number of transistors for microprocessors at en.wikipedia.org.

Keywords: neurobiology evolution of the nervour system