History of RSB Interview: John J. Hopfield

July 17, 2020, 10:00-11:15am (EDT). Final revision: September 15, 2020.

Interviewers:
Patrick Charbonneau, Duke University, patrick.charbonneau@duke.edu
Francesco Zamponi, ENS-Paris

Location:
Over Zoom, from Prof. Hopfield's house in Princeton, New Jersey, USA.

How to cite:
P. Charbonneau, *History of RSB Interview: John J. Hopfield*, transcript of an oral history conducted 2020 by Patrick Charbonneau and Francesco Zamponi, History of RSB Project, CAPHÉS, École normale supérieure, Paris, 2020, 21 p. https://doi.org/11280/5fd45598

PC: [0:00:00] As I said, the main focus of this interview is going to be on the period where your work and your efforts interfaced most closely with the spin glass work and community. I think we could start on a question about that genesis and about that time. In your 2018 autobiographical essay entitled "Now What?" you describe the various steps that led to the formulation of what we now know as the Hopfield model. Would you mind walking us through some of those steps and highlighting the stages, at which the ideas from the work on spin glasses were most important?

JJH: I'll be happy to. I wandered in the area quite by accident. I had taught the Hodgkin–Huxley equations as part of teaching biophysics (or even general biophysical chemistry sometimes). I never thought I would actually do research on how the brain worked—I thought the brain was simply too complex [for physics]. I didn't see any way of finding me "a problem". And so [I learned] the details of the biophysics of how individual neurons operate for classroom lecture material, and that was all for neurobiology. Then Francis O. Schmitt descended on me one day.

I had come back very recently from Copenhagen where I had been at the Niels Bohr Institute/Nordita. Niels Bohr was himself peripherally interested in biology, and every once in a while he would organize something biological [at his institute]. That family has gone back and forth, or course. Niels Bohr's father was a very famous physiologist. There is a so-called Bohr effect in hemoglobin, which has to do with the pH dependence of oxygen binding. That was his father. Niels Bohr had an intuitive interest in how quantum mechanics and biology might relate. There is a famous lecture of Niels Bohr in 1932 entitled "Light and Life", a talk given to doctors who used light physiologically to cure various medical problems which are

photosensitive. Young Max Delbrück happened to arrive in Copenhagen, from Germany, just in time to attend that seminal lecture. Bohr was trying to describe the parallel between how biology operated and quantum mechanics [operated], and whether there might be things in biology which were so sensitively dependent on quantum mechanics that biology, through evolution, could actually elucidate some laws of physics that you wouldn't be able to measure or discover in conventional physical experiments. An interesting idea, I think. In the long run not particularly sound, but very intriguing, and it provided the impetus which moved Delbrück from conventional theoretical physics to pursuing biology.

[My six month stay at in Copenhagen] in 1977 was part of an occasional—every 10 or 20 years—exploration of what physics has to do with biology. I organized a series of seminars at the Bohr Institute, inviting excellent people to lecture. You might lump the lecturers together as biophysicists or physical biochemists, with marvelous quantitative understandings of biological problems. Each was working on a very particular biological problem. I was [running the lecture series] because it was why I had been invited. On the other hand, I was myself looking for a problem. And I didn't find one.

Of course, no one invited to talk to such a seminar audience says “I'm failing to address the major problems of my enterprise” and try to enlist help. Instead, they try to convince you that they have the major problems all defined, in hand, all well solved. They are [viewed as] great scientists because they've done so. Unsurprisingly I didn't come back to Princeton with any new problem to work on.

The following autumn, Francis O. Schmitt descended on me. Schmitt had a pretty strong background in engineering and in biology, and later in life became interested in neurobiology. He must have been 75 at the time I first met him, but he was maintaining a strong interest in the intellectual enterprise of neurobiology. In fact, he coined the term neuroscience for the combinations of neuro-X, understanding that there were psychologists, neuroendocrinologists, neuroanatomist, neurochemists, and structural biologists who are all looking into particular subsets of details of the science of the brain. There were many diverse, separate scientific fields which he felt would need to be understood together before you would actually understand how the brain works. My formulation [now] of the issue would emphasize that, in the long run, what the brain does is computation. Computation has to be described as a mathematical structure. If you can't describe the brain in mathematical terms you are never going to understand how it operates. Frank was trying to interest me in attending his little meetings, in which he had the diverse subfields trying to talk to each other.

Schmitt was very good at getting good people to go to his meetings. He was posing a challenge 'how does the brain actually work' to a very diverse community of neuroscientists, a challenge none of them individually were capable of meeting, in part because virtually all of them were illiterate in mathematics. I knew no neurobiology, so there was not much I could directly contribute to the grand question either, but at least I understood the relationship between mathematics and quantitative science. Schmitt wanted to get me to attend his meetings because he felt that representation from physics might somehow be useful. [He was right—the field decades later finds the viewpoint from physics useful.]

I went to my first meeting, and from the general tenor [of it], I realized that the set of people present was never going to solve the problem. Suppose like Schmitt, you want to understand how the brain works. If you ask a cardiologist how the heart works, you'll get a good description of a pump, and as an engineer you can understand the cardiologist's model. If, on the other hand, you ask a neurosurgeon how the brain works, e.g. how you recognize your wife, you'll get a story which has nothing to do with reality. That's the enormous gap in science you're trying to fill. I joined Schmitt's little 'club' realizing Frank was giving me the opportunity to get an education, and that if I joined the club I might be able to find a problem.

That's how my trajectory started, from knowing nothing about neurobiology. The year was 1977 so there was a fair interest in spin glasses in the physics community by then. I knew some of the cast of characters. I had an earlier incarnation in Bell Labs with semiconductors and so on beginning in 1958 when I joined its theory group. I have known Phil Anderson since that time. I'd always gone to see Phil when I went back to Bell later as a consultant or part-time employee. And I generally knew the kind of things that he was working on, although usually didn't know any details at all. But I was in a unique situation where ideas about spin glasses and the facts of neurobiology could at least get together in one person. That's where the connection between the two began.

PC: [0:09:55] Would you mind telling us a bit more about, as you just said and you wrote in your essay, Phil Anderson was really influential in you knowing about the existence of spin glasses. So can you tell us a bit more about your relationship with Phil and how did spin glass emerge in these discussions?

JJH: I went to Bell labs, basically as a postdoc in 1958. Phil was at that time still a hidden gem. He had published his papers on localization in 1957, and done some work on gauge invariance in superconductivity. The original BCS equations were not gauge invariant. Phil really viewed himself as the

world's most underrated theoretical physicist. But we got along. He taught me the game of Go. Afterwards, when I moved off to academia and to consulting at Bell with a group more focussed on light and semiconductors, I kept in peripheral contact with the theoretical group. And often when I went as a consultant for a few days, I would first find Phil and spend half an hour with him, hearing what he felt was currently exciting.

The first thing I ever heard about what would now be called a spin glass was as follows. There was a theoretical seminar at Bell in about 1963. You can figure out what the date was by the subject. The theoretical seminar at Bell was an institution where the whole idea was to try to ask such telling questions of the speaker that you made it clear that you knew much more about the speaker's subject than the speaker did. The speaker that day was A. W. Overhauser, my former thesis professor, and of course a great theorist. For no obvious reason, Phil just did not like Al. What Overhauser described were the very peculiar properties of dilute manganese in copper. Overhauser was good at organizing experiments, pointing out one anomaly we don't know how to explain after another, and assembling a set of unexplained results to create a problem which he could then unify. He had done a masterful job of organizing anomalies for these dilute alloys of manganese and copper. A particular detail I remember is the low-temperature specific heat of dilute alloys of manganese in copper. They are metallic, with linear specific heat at low temperatures. But the slope of that specific heat increases linearly with the manganese concentration. The area under the curve is related to the mole-fraction of manganese. I can't remember any more details, but it was before I had heard any public discussion of a system which later would be termed a spin glass. Overhauser was, in the seminar, trying to explain a diverse set of experiments in terms of spin density waves. Phil felt that was morally wrong. Discussion was bloody. The term spin glass I probably didn't hear until much later—1973, I'm guessing. This seminar was a decade earlier. Phil among others was worrying about these dilute magnetic alloys early. Theorists were all floundering, trying to identify what the problem was.

PC: [0:14:25] I see. You said that by around 1977, that's when you made this connection between, or maybe a couple years later, between spin glasses and the brain.

JJH: Because I heard this early lecture in the sixties, when spin glasses came along I would always listen for them when I went to see Phil. And because it was Phil, there were interesting visitors at Bell whom I would also hear. When Phil was in Cambridge I spent half a year at the Cavendish Laboratory with the group that Mott, Phil, and Heine were in. I met some of the very young crew and some of the older crew who were scattered around

England. I met Sherrington at that time, I met a young Richard Palmer who came and did a postdoc with me later. It afforded me the opportunity to learn what was going on in spin glasses at a time when spin glasses had begun to be mathematically developed.

However, when I started on neurobiology I didn't have any idea that spin glasses would have any relevance whatsoever. I was just looking for a problem. I had to seek a problem which was solvable from the point of view of a physicist, because if [I] had to know all the details of general biology to pose a problem I would never succeed. There had to be something that could be mathematized without infinite biological detail. If the problem did not connect with any of the physics I knew, I was probably not be able to make any progress.

The way I generally made progress on biological problems was by already understanding something about how physics works, and finding peculiar ways of mapping my understanding onto biology. So I made up models of multiple neurons interacting with each other and trying to understand what kind of collective behavior the neural activity would have. The one thing which became immediately obvious thinking about neurobiology, in terms of physics, was that we were basically trying to explain a dynamical behavior. Equilibrium is not of interest. You're dealing with the trajectory of a strongly driven system. And although technically there's much about computation which is reversible, when as an engineer you build anything real (other than quantum computers), all the computation is done irreversibly. There are two inputs to a logical gate, which has a single binary output—ahah! information compression, which means irreversibility. So I knew I had to be analyzing a driven dynamical system. My problem there was that I knew nothing about dynamical systems. The only multivariable dynamical systems I understood at all were those for statistical mechanics, where dynamics described the way the systems sought equilibrium.

Coming from the world of physics, what do you want to understand when I ask you 'how you recognize your wife?' The following is NOT an answer. Simulate all the dynamical variables describing 10^{11} neurons interacting via 10^{14} synapses, and this will show you how the decision is made. In the same sense, you don't want to understand aerodynamics by saying let's simulate Newton's Laws of motion for 10^{24} interacting molecules. Instead, you derive the Navier-stokes equation*, in which molecules have disappeared*, and you start from there. That same spirit is needed as you try to think about how the brain works. The really interesting questions are in some sense higher level. What we have to do is see if you can make some progress from a lower level of description to some next higher level. The whole idea of emergence had become popular in the 70s. Neurobiology, I would still

claim, has more levels of emergence than most physical systems you have thought about. So I began looking for emergent behavior in systems of large number of neurons. At the time I was working, 30 was a big number, 100 was a very big number. Computers have gained, roughly speaking, a factor of 2 in computing power every 2 or 3 years. To appreciate the computing power available to me in 1978, that's 15 powers of two. Start with the computing power you have now, go backward by 2^{15} –that is roughly what I had to work with.

I don't know whether your roots are quantum chemistry—how have you wound up in a chemistry department. What are your roots?

PC: [0:20:42] I am a chemical physicist. My roots are: I worked with David Reichman as a graduate student and with Daan Frenkel in Amsterdam for my postdoc. So I come from the interface between chemistry and physics from the stat mech world.

JJH: So not from the massive chemistry computation side, no.

I made various models. One of the ones which intrigued me greatly for a while was Conway's Game of Life, which you probably know. The kinds of patterns you would see in the dynamics of The Game of Life when you could run sufficiently long or sufficiently fast were really quite intriguing. On the other hand, the rules of the game were very rigid, and if you didn't have exact rules you really couldn't produce any interesting long-term behaviors. I tried for a while to make a Game of Life which was somewhat like The Game of Life rules but also more like neurobiology, and did simulations. I couldn't do very good simulations in the available computer environment. What I produced was fundamentally junk.

In January 1980 I moved from Princeton Physics to Caltech, where I was jointly in Chemistry and Biology. Princeton Physics had by the standards of the day somewhat mediocre computation facilities, and Caltech Chemistry, thanks to people like Bill Goddard, had really first-rate computer facilities. So when I went to Caltech, I could suddenly do, easily, simulations of things I had wanted to do for a year, and inside three weeks I convinced myself that a modified Game of Life was an absolutely useless direction to pursue.

The reason was that the rules of the Game of Life are simple logical rules. If the Game rules were described in terms of many-body interactions, the effective interactions were very short-ranged, and involved more than two bodies. That wasn't going to be much like neurobiology, which I knew was intrinsically one neuron talking to on the scale of a thousand others. It was

also unlike any physics spin-system I had ever seen. I did simulations, but Life has no energy function behind it, and thus no statistical mechanics. I could only do dynamics by simulation. To do dynamics by simulation requires knowing all the neuron-neuron interactions. I wasn't going to get very far because I didn't have any general theory of dynamical systems. I had to look for a dynamical system whose dynamics describe an approach to equilibrium. An 'Aha' moment came when I realized that by working with a system whose dynamics could be described as going downhill on an energy function, I had a link to statistical mechanics. Maybe I can solve this statistical mechanics in some special cases. I can therefore do simulations in which the state of the system approaches equilibrium, and talk about that elementary kind of dynamics, approach to equilibrium, as a simple form of computation. The simplest case to consider would be at zero temperature.

What's the simplest computation you can imagine, a computation that you could alternatively describe as a dynamical system going toward equilibrium? I hit upon idea of recovering all of a particular memory from partial information about that memory. Once I formulated memory and recall in that way, things started to fall together. It was easy to make a 'spin' model of activity of a large group of neurons, spin up meaning a neuron is active, spin down, inactive. Given a state that you want for a memory, the coupling parameters that will make this state a low energy state are trivial to write down. You want to have more than one memory, of course. That leads to summing sets of exchange parameters. The results don’t represent anything perfectly, there is *frustration* as it were. I could immediately see there’s going to be an interesting connection between a neural memory and frustrated spin systems.

I never used any of the spin glass mathematics myself, because for useful memory you needed to be in an intermediate-range, where you have many understandable almost spin glass states, but you don’t yet have the limiting case. A spin glass doesn't remember anything you put into it. It averages over too much, and doesn't know anything in particular. Associative memory does know about particular stored things. That's what I found, finally doing very simple mathematics and simulations. If you put in a large number of memories you obtain an infinite-range spin glass. If you have only one memory, the system is essentially a ferromagnet. In between? I didn't try to do any elegant mathematics of the in-between regime. I only did zero temperature simulations. That is the case in which you could actually see the system usefully computing something, namely reconstructing a big memory from partial information about that particular memory.

The most astounding thing to me now was what happened after I had written it down and send it off to PNAS. PNAS is today a quality journal with very wide-ranging interests. At the time PNAS was almost solely a biological Journal. It had occasional articles on physics—very, very few indeed, and nothing at all in computer science. But it was in some physics libraries, and as the NAS journal it ought to promote such interdisciplinary material. I had the advantage of having been elected to the NAS for my work in condensed matter physics. At the time, if you were a member of the Academy, you could publish a short paper in PNAS with essentially no reviewing process. So I took advantage of my situation and published in PNAS. An easy thing for me to do, but the likelihood of being found by relevant scientists seemed small.

By virtue of the fact that I had *not* done any of the interesting mathematics of this intermediate-range, or studied the system at finite temperature, there were obvious further directions to explore. When the paper came out, theorists from spin glasses and statistical physics said, "that's something of which I could actually put my talents". The fact that I had not exhausted the field, but had only opened the lid of the box made it possible for people to join in. It wasn't like the set of lectures I organized in Copenhagen, where the lecturers were chiefly representing themselves as having already understood the fundamentals of their fields so thoroughly that outsiders could not hope to contribute much.

PC: [0:29:00] So picking up on that last point, as you said you were doing mostly the computations in that regime. This is what gave you the confidence in writing that paper and publishing it. And many people picked up on those ideas. Did you follow their work after that? How much were you aware of, let's say, the efforts of Amit, Gutfreund, and Sompolinsky?

JJH: I had met Amit in my condensed matter physics days. I had met Sompolinsky because he had been a frequent visitor of the theoretical group in Bell labs. I knew who Gutfreund was from solid state physics days. But in fact I did not know they were working on following up my PNAS paper until they published, and someone called my attention to their published paper[1].

PC: [0:30:02] Ok. Following your work, as you said, there's a lot of people in the spin glass community that got interested, including those three, and

[1] Daniel J. Amit, Hanoch Gutfreund and H. Sompolinsky, "Storing Infinite Numbers of Patterns in a Spin-Glass Model of Neural Networks," *Phys. Rev. Lett.* **55**, 1530 (1985). https://doi.org/10.1103/PhysRevLett.55.1530

I'm aware of at least one meeting in 1985 at Les Houches[2], where you went. It was a NATO workshop on disordered systems and biological organization, where there were some people from the spin glass community.

JJH: I would listen to them and they would listen to me. And yet we had separate enough interests and abilities that I never actually directly worked with these people, or the mathematical problems that they were dealing with.

PC: [0:30:48] Were there many such meetings? Or is this 1985 [meeting] the only one?

JJH: There were not many such meetings. What sprung up instead was more meetings in which those who did mathematical theory, and engineers interested in computing hardware got together with neurobiologists to generate the more detailed mathematics that real neurobiology would necessitate. So the interface with the statistical physics community did not directly bear much fruit in neurobiology, except through enabling many people to move from physics into computational biology or neural networks. A substantial set of physicists entered the field by reading my paper, Amit, Sompolinsky, Gutfreund, Larry Abbott, David Kleinfeld, Sara Solla, Larry Jackal, Leo van Hemmen, for example. Some tried to describe experimental results by modifications of the original associative memory paper. Daniel Amit went from the original model towards biology because he could see that it would be possible to use the style of mathematics used in disordered systems, the kind of mathematics that physicists had done before, except now you had to apply it to non-equilibrium and non-random systems.

I hadn't even heard the name Lyapunov when I began working in these directions. It wasn't until I had written the draft of that paper, and had gone somewhere to give a seminar on it, when someone said "Isn't that a form of Lyapunov function". And I replied, "Who is Lyapunov?". General dynamical systems were usually not part of the curriculum of physics education. What little I knew of such systems came from an engineering course in electronic (vacuum tube!) circuits.

PC: I see.

[2] Cf. The proceedings of the NATO Advanced Research Workshop on Disordered Systems and Biological Organization held at Les Houches, February 25-March 8, 1985. *Disordered Systems and Biological Organization*, E. Bienenstock, F. Fogelman Soulié, and G. Weisbuch eds. (Berlin: Springer-Verlag, 1986).

JJH: Lyapunov functions were simply not a part of physics education. Hydrodynamics was seldom part of physics education in the US. Physics education here emphasized quantum mechanics, equilibrium statistical mechanics, classical mechanics, electricity and magnetism, the atom, the nucleus, and the solid state. General theories of strongly non-equilibrium systems (with the possible exception of the work of Prigogine), were seen as very suspect and quite possibly wrong.

PC: [0:33:42] If you allow me, I'd like to push a bit further in this direction, because following the work of Amit and co-workers there was a number of other papers that revived the perceptron model, which in some ways was antecedent to neural networks, and had been left fallow since the late sixties. In particular, the work of Elizabeth Gardner and Bernard Derrida, which explores artificial networks in which the couplings are not given by a prescribed rule, such as the Hebb rule. So did you follow this? This work was not necessarily directly bridging between physics and biology, but was more about expanding the physics interface. Were you aware of it?

JJH: Meetings such as those in Santa Barbara (1985) and Snowbird (1986), Neural Networks for Computing became the home of this expanding new community. There were very nice papers showing things about the capacity of the perceptron for random problems. (I remember Elizabeth Gardner's name attached to that.) The perceptron is an interesting phenomenon, because a perceptron with one layer of weights has well-defined capabilities for pattern recognition, and they are very limited. They cannot solve most problems. Marvin Minsky wrote the book *The perceptron*[3], which is chiefly concerned with the one layer of weights perceptron.

The book also contains little bit of hand-waving about what happens if there are more layers of weights. Minsky did not understand how to train a network with more layers of weights. He even surmised that it would turn out to be no more powerful than the single layer of weights perception, so the generalization to more layers of weights was quite possibly useless. Because Minsky had done such defining mathematics of the simple perceptron, this surmise had a lot of impact, and almost killed the neural network field for a while[4]. People who didn't really know about Minsky, on the other hand, pursued the multi-layer analog neural networks. Min-

[3] M. Minsky and S. Papert, *Perceptrons: an introduction to computational geometry*, (Cambridge: MIT Press, 1969). https://en.wikipedia.org/wiki/Perceptrons_(book)

[4] See also, Mikel Olazaran, "A Sociological Study of the Official History of the Perceptrons Controversy," *Social Studies of Science* **26**, 611-659 (1996). https://doi.org/10.1177/030631296026003005; "A Sociological History of the Neural Network Controversy," *Advances in Computers* **37**, 335-425 (1993). https://doi.org/10.1016/S0065-2458(08)60408-8

sky was basically stuck by thinking of computing elements as logical devices. The real progress was made by saying that let's presume that computing elements are continuous devices. Then you can use the power of continuous mathematics. Again you can see that the effect of available computer power was very important. The multi-layer perceptron was first appropriately described as a learning rule by differentiation through multiple layers of weights in about 1973 by Werbos, but in 1973 you couldn't do the simulations necessary to demonstrate that the learning rule would actually be useful. By 1980, you could do the simulations—it was costly but you could—and by 1985, anybody could do them. The development of computer power had a major impact on how and when 'neural networks for computing and neurobiology' and 'physicists in neurobiology and neural networks' came about. Feed-forward neural networks do not have any interesting dynamics. My 1982 paper described a feedback networks with useful computing dynamics, and thus interesting physics. This explains why it, rather than the perceptron, was seminal to interesting physicists in neural networks.

PC: [0:37:50] Just to summarize: If I understand correctly, you know of this work on the perceptron that came out theoretically, but you didn't follow it at the time much. Is that fair to say?

JJH: That's right. The thing is, the single-layer perceptron had been beautifully and exhaustively analyzed by Minsky. There was an elegant statistical theorem that Minsky proved about the capacity of it, and made it clear that it was not a useful basis for talking about the complex tasks solved by biological neural networks.

PC: [0:38:30] Fair enough. Talking about computation: if I understood correctly in the early 80s, 1981 to 1983, you were teaching a class on the physics of computation at Caltech, or co-teaching it with Feynman, in particular. Is it correct to say that you were teaching a module on neural networks in that class or that's for a later class. I am a bit confused.

JJH: Ok. Let's get a little bit of Caltech history[5]. Shortly after I got to Caltech, Carver Mead, a computer science/electrical engineering, and Feynman got together over supper. Carver liked to be a sort of buddy to Feynman occasionally, and I think they both enjoyed the interaction. They thought that it would be interesting to try to do a course involving physics, computation—covering all possible subjects of computation—and how the brain works. Each of them saw describing how the brain works as a problem in

[5] See also: John Hopfield, "Feynman and Computation", In: *Feynman and Computation*, Anthony J. C. Hey ed. (Reading, MA: Perseus Books Publishing, 1999).

understanding how a physical system produces computation. They wanted to have a course on the physics of computation, but neither of them had a strong suit in biology. Biology was a subject, as Tom Lehrer said[6], I [they] knew from nothing. [Why do I think of a Tom Lehrer lyric here? Because he wrote a comic song, nominally about Lobachevsky, which I knew from graduate student days.] They needed a biologist as part of this enterprise. They thought 'Hopfield is sort of a biologist, but more a physicist. Let's make it the three of us, and try to organize a course around the *physics of computation*'. It was to go from the physics of how neurons and synapses work to the physics of transistors, and questions of the role of quantum mechanics and reversibility in computation. All would merge together as one enormous beautiful course. A grand idea. At Caltech you could get away with trying it.

Shortly after the notion of such a course had been duly approved by the faculty, Feynman had one of his horrible but heroically fought bouts of cancer. So that first one-year course did not have Feynman participation at all—he was unavailable the entire year, and not a participant in the detailed planning. Carver and I did the best we could, but lacked Feynman's genius. That was a major weakness. We tried to make up for this lack by getting our stellar friends to come and give lectures. Carver and I didn't have to do much lecturing ourselves. We assumed that the students were brilliant, and therefore you didn't really have to manufacture problem sets to help in their education. You had only to expose the students to the style of thought of the lectures described and that would be sufficient.

The number of people in attending class died with an exponential of about 0.3 years time-constant. A brave few held on to the end, one of them being a current Caltech faculty member, Markus Meister. (Have you ever run across Markus Meister? No). He is a German physicist who had come to Caltech on a one year visiting graduate fellowship. He got hooked by physics and biology, and had a trajectory which ultimately brought him back to Caltech biology, where he does very interesting, totally physics-based, experiments. He is one of the two people I know that attended to the end, the other being David Beratan, a theoretical chemistry student of mine. I myself profited immensely from attending *all* the lectures. It was as if the subjects and level of the lectures had been designed for someone of exactly my background, which was of course true.

The following year I ran into Feynman at lunch. He said: "Whatever happened to this course that me and you and Mead were going to give." I told him that it took place, and that it had been a disaster without Feynman,

[6] From the song *Lobachevsky* by Tom Lehrer: https://en.wikipedia.org/wiki/Lobachevsky_(song)

and that Carver and I had decided to never again try such a thing. “But I would have liked to have done it. Couldn't we do something?” Carver was not comfortable with trying again, nor was I. But I finally agreed to participate with Feynman and jointly produce a one term course, on the condition that Feynman himself would be primarily responsible for planning and inviting guest lecturers, which led us to a series of fascinating Athenaeum lunches.

Every week contained one lecture on some topic which Feynman had picked. For each topic he found a friend or contact who was an expert to the field to lecture on the subject. The other lecture of the week was most often given by Feynman, either on what the lecturer would have said if he'd actually understood his field, or on how to understand and organize the material presented in lecture from the viewpoint of physics. And it was brilliant Feynman. I learned a lot from that course. It was through this kind of thing I learned—Oh, take error codes. I knew about error correcting codes, but had never thought about them as part of physics. Feynman was always interested in the limits that physics placed on possible technology. As lecturer, Feynman got one of the world’s experts on error-correcting codes from JPL, which was concerned with error-correcting codes because of needing to deal with the weak signals coming from interplanetary space probes.

The first lecture was Marvin Minsky talking about Minsky’s view of the world of computers, computation, and AI. Near the end of term, Feynman talked about his views on the limitations from quantum mechanics that were put on the classical devices currently used in computer hardware. I gave a couple of lectures on my view of neural networks. Feynman knew about my 1982 paper, for his son had a summer job... The story is too long and convoluted to describe here, but involves Minsky, Danny Hillis, and a bizarre computer with 64,000 one-bit processors being built by a startup called Thinking Machines Corporation, and Feynman finding an efficient way to program associative memory dynamics onto a Connection Machine. There was a large audience when finally Feynman lectured on quantum computation.

By the next year, things had further separated. I gave an early version of what was to evolve into a neural net/computational neurobiology course. Feynman was truly doing the physics of computation. Carver Mead had gone back to doing the physics of VLSI[7] and large-scale VLSI using analog systems, where I also gave a week of lectures.

[7] VLSI: Very large-scale integration.

PC: [0:45:56] So by the time it got to be your course in a full semester, or full quarter, of neural networks, former students have said that there was actually a significant fraction of the class about spin glasses. What were you teaching about spin glasses?

JJH: I taught the elementary statistical mechanics of simple model spin systems in particular simple cases and limits, including my 1982 paper and its 1984 extension to continuous variables and a continuous dynamical system. The idea of frustration, in general and as expressed in spin systems. In 1986, the fact that the idea of spin Hamiltonians led to a programming language. I remember hearing Scott Kirkpatrick give a talk at Bell Labs on simulated annealing for 'solving' some hard problems in computation. A very clever idea. In thinking about what this implied more broadly, I realized that because there is an energy function behind a spin glassy system, if you can design that energy function to be the function you want to minimize, you can get the spin approach to low temperature equilibrium perform the computation for you. And this would be true not only for spin systems, but (using the 1984 paper) for continuous variables. This makes a spin Hamiltonian into a programming language there. The papers David Tank and I wrote on these ideas were simple illustrations of such programming, and amusing to theorists because a spin Hamiltonian could be written for a variety of hard problems, including the Travelling Salesman Problem.

PC: [0:47:35] That makes complete sense. I have a question related to that same time and because earlier you've mentioned Richard Palmer, who is my former colleague at Duke. From what I can tell, you wrote one paper with him, and that is the only paper you wrote with someone from the "spin glass community." Can you tell us a bit more about the genesis of that paper, and the relationship you maintained with Richard after he left.

JJH: I had met Richard in Cambridge when he was an undergraduate. When his thesis mentor Phil Anderson asked whether he might do a postdoctoral at Princeton, I seized the opportunity. He came in part because of my interest in biology, but this was before my interest in neurobiology. Then he went to Duke and we lost contact. I went off to Caltech, carried out the research for my 1982 paper and began describing it to the spin physics theorists. I had a little bit of money for visitors, and Richard expressed interest in coming out. This happened at the time that I had become interested in unlearning. When too many memories are written into a network, random correlations result in the formation of 'too dominant', or 'too deep' attractors, which dominate the system. These are forerunners of true spin glass states. My idea of how to get rid of these too dominant attractors was to fall into them because they're too deep (and thus have too large basins of attraction), and then to modify the exchange interactions (synapses) to

make those too-deep states less attractive by unlearning these states, using the usual 'Hebbian learning' procedure but with a minus sign. Iterating this procedure should plausibly even out the performance. All I had when Richard arrived was a 'Just So Story'[8], no mathematics, and a bright physics graduate student David Feinstein who was good at simulations, when Richard arrived to visit I told Richard the 'Just So Story.' He took over most of the research for seeing whether or not my 'Just So Story' was true.

I had known Francis Crick because of my earlier interest in molecular biology and my work on kinetic proofreading in the accurate biosynthesis of proteins and nucleic acids. Crick had moved from molecular biology into neurobiology a few years earlier, and was now at the Salk Institute. These vague connections were the way that I learned of Crick's interest in dreams. His idea was that the function of a dream is not to remember the form of a dream itself. A dream itself is caused by the 'state' of remembering the dream being too easy to get into, even from meaningless inputs. Thus one should 'unlearn' when you are dreaming, and make that state less favorable. What Crick and I were suggesting was the same thing. But whereas Crick and coworker Graeme Mitchison had only a 'just so' description of what they thought could happen, we had mathematics and simulations of the basic phenomenon. I talked with Francis. He agreed that what we at Caltech were working on was very close to what they were thinking about. Crick and Mitchison already had an article on this subject accepted for publication by *Nature*. Crick had great influence on one of the editors, and by this means it was ultimately arranged that a letter on our research would be published in the same issue of Nature with their longer article[9].

PC: [0:51:16] But you never kept on… Sorry, what were you going to say?

JJH: It was the fortunate happenstance of knowing Richard, inviting him out as a visitor, and knowing Francis, and the whole thing just went together beautifully. Richard was a joy to work with; I always liked him. We just didn't naturally crossed paths very often. He was sufficiently from the statistical physics community, and I was sufficiently more neuro/engineering that we just didn't get together to discuss common problems.

PC: [0:52:13] But did you stay in touch with him at all, after this work?

---

[8] ''Just so stories" is a children's book by Rudyard Kipling containing delightfully plausible explanations of how animals got their characteristic forms—all delightfully spurious. https://en.wikipedia.org/wiki/Just_So_Stories

[9] J. J. Hopfield, D. Feinstein and R. Palmer, 'Unlearning' has a stabilizing effect in collective memories," *Nature* **304**, 158–159 (1983). https://doi.org/10.1038/304158a0; F. Crick and G. Mitchison "The function of dream sleep," *Nature* **304**, 111–114 (1983). https://doi.org/10.1038/304111a0

JJH: Only very casually. I knew, of course, when he had a stroke. That tragically separated him from the active of the research community.

PC: Because with such a high-profile paper I could have imagined the birth of a long collaboration. That's why I was curious why it didn't.

JJH: I had only a few long collaborations. There are students I've kept up with for a long time. But my students / my postdocs--I've always wanted to encourage them to have their own problems. And I've never been very good at splitting a problem with people, really sharing the work load and the invention. I did truly share some problems with David Tank; I did so for a while with Carlos Brody. The easiest joint research I ever had was with David Thomas at Bell Labs, because he was an experimental chemist, and I was fundamentally a theoretical condensed matter physicist. We could join hands on a problem and everybody understood who was contributing what. Yet both of us were absolutely necessary for continued progress. When the two of us were as different as that, it was really very profitable to have a continuing intense interaction of ideas.

I've always found when the two of you are intellectually too close together it's difficult. There tends to be too much feeling for whose ideas drive the work forward, and it is easy for resentment to get in the way of collaboration. One happy end result is to collaborate for a while, then go separate ways as interests or insights diverge.

I have never been able to just hand things off to somebody say: "This is your problem, work on it and tell me about it when you have finished." I learn so much by looking into all the details myself, finding out why a simulation does not in fact work, and learning both science and programming from the process, and locating myself where new interesting problems are to be found in the failure mode of my current point of view. The paper on unlearning was an exception to my usual mode of operation, in that the idea was mine, and the details were Palmer and Feinstein. By not doing more math or simulations myself, working on this paper had not led me to a next important question. There were a few obvious directions to explore, but no smoking gun. Richard might easily have pursued one, but never discussed with me where to go next. Simple geometry took its toll—Palmer ended his short visit to Caltech and returned to Duke. So further explorations from the "unlearning" paper were left to others, who brought additional viewpoints to the subject.

Only very recently have I landed in a true collaboration, with a young former Russian, Dmitry Krotov. While both of us are nominally physics theorists, I am essentially amathematical compared to him. Much that he does

"jointly" I could never have done. However, I bring 40 years of neural networks and neuroscience lore to the discussions. It makes for vigorous interactions, and a creative choice of questions to work on.

PC: [0:55:33] If we go back a bit to the reception to the 1982 PNAS. From what I understand the response was quite immediate. You received enthusiastic both scientific and institutional responses. You received the MacArthur Fellowship the following year for that work essentially, and I heard that both Caltech and AT&T Bell Labs were organizing *Hopfest* just the following year[10], which they kept on afterwards. How do you explain that the response was so immediate and instantaneous to the paper?

JJH: The response was hugely different in different communities. In 1981 I gave my first truly public talk on this subject at an August gathering in Paris called the Institute de la Vie. Other speakers on aspects of neurobiology included physics Nobelists Leon Cooper and Donald Glaser. No one has ever referred to that talk. The first *Hopfest* was a complete surprise. I lunched at the faculty club with a friend, and following lunch he said there was a seminar I might be interested, and took me to the seminar room where the meeting was just beginning, with a very diverse program of Techies and JPL scientists. The paper connected—loosely—a wide swath of science and technology. I was astonished.

However, I was the Dickerson Professor of Chemistry and Biology in 1982. In the following 10 years, I was *never* asked to talk in the Caltech physics department. I happened to know Max Cowan, the prominent editor of a major neurobiology journal, and asked him where I might publish the (future) 1982 paper. His answer was 'I can't see that it is related to neurons or the brain—certainly not in my journal'. Many institutions and scientists ignored or rejected the 1982-84 work.

I *can* talk to condensed matter physicists and electrical engineers in their own language. Communication of the research opportunity to this community was very important. In some circles the seed had fallen on fertile ground and grew rapidly. The opportunity to use your research skills in physical science/engineering to try to understand the brain was hard to resist. I once asked an electrical engineer why he began working on neural networks. He replied that my papers presented the most interesting analog circuit problem he had seen in years. At Caltech in engineering, at Bell Labs, at JPL, meetings, seminars and hardware development projects were rapidly initiated by the 1982 paper.

[10] J. Miller and J. M. Bower, "Introduction: Origins and History of the CNS Meetings". In: J. Bower, ed. *20 Years of Computational Neuroscience* (New York: Springer, 2013).

At the same time, there were other intellectual threads that were closely related. Work on the 'backprop' learning algorithm was just beginning. The 'Boltzmann Machine', based on my 1982 paper generalized to finite temperature was closely related. One of the things that the 1982 paper failed to do was to find an interesting learning algorithm. You could only instruct a network to learn a specific memory in one big step. My instruction step was very like the behavior of Hebbian synapses in biology. Back prop and the Boltzmann machine both developed the idea that integrated incremental synapse changes could construct networks for solving general problems beyond associative memory.

What the Boltzmann machine did involving temperature was trivial, and contained in an early draft of the paper which I wrote for PNAS. The draft was too long so it got taken out. The fact that there was an incremental learning rule for the Boltzmann machine was a major intellectual breakthrough. That advance brought a whole AI learning community in from the sidelines. For the original Boltzmann machine paper by Hinton, Sejnowski and Ackley—Ackley was, I believe, a graduate student who did the programming—Terry Sejnowski had to teach Hinton (intellectual roots in AI) the rudiments of statistical mechanics, and Hinton then could see how to do incremental machine learning. True synergy of disciplines.

PC: [0:59:26] So the world was ripe, in a sense, for that work.

JJH: The world was ripe, that's right. And it didn't hurt that world… Terry Senjowski was an interesting case. Terry Senjnowski was a… Do you know him?

PC: No I don't.

JJH: Do you know of him in any sense?

PC: I wish I could say I do. I will know of him shortly after our interview, but I don't know him at this moment.

JJH: He was the co-discoverer/creator of the Boltzmann Machine in machine learning, a very important force for getting statistical physics and learning together. He started off as a graduate student of general relativist John Wheeler, but there was a falling-out. I ran into Terry when he was auditing a course I was giving in physical biochemistry at Princeton. (How far the biochemistry department at Princeton had fallen. I, who had never had more than elementary chemistry course, was teaching graduate physical biochemistry.) Terry was auditing, and would chat with me after class. He showed me a couple papers he had published while he was a dropout.

These were very simple neural computational things, not very important in hindsight. But significantly, he had done the work entirely on his own. I said: "Here is the thesis deal. Package these papers with an introduction and a conclusion, and I will attest to it as a physics thesis." Thus Terry wound up being a student of mine, and I've been close to him ever since. He would have known about my associative memory effort well before it was published.

Somebody like that with very good physical understanding and knowledge of what physics can do, who then learns some neurobiology, can contribute a lot. I can't think of a major impact spin glasses as such had on computational neurobiology, other than through the attitudes of the scientists who were changing the field they were invading. Part of it is just the physicists' confidence: I know more than anybody else; I'm smarter than anybody else; if you tell me the facts, I will construct a better explanation or theory than traditionalists in the field.

PC: It's part of the trade, right?

JJH: Yes.

PC: [1:02:40] Before we close, is there anything else you would like to share with us about that era that we missed, that we should be exploring?

JJH: It's interesting. Physicists homed in on the spin glass because simple models containing glassy randomness had very interesting statistical dynamics with a degree of universality. Spin glasses are an emergent behavior of large collective systems characterized by randomness. This partial understanding could lead to a more general understanding of emergence.

If you want to understand how the laws of psychology arise from neurobiology, you will probably need to fight your way through several layers of emergence. Emergence is a subject which was not much discussed except in orderly systems[11]. The general idea of emergence is more difficult with randomness, and much more difficult when there is some structure and some randomness, and in addition multiple physical scales and time scales[12]. Going back to associative memory for a moment, note that the physics viewpoint uses random memories for simplicity. Real biological memories are expected to be highly correlated. To do a physics-based

[11] See, however, P. W. Anderson, "More is Different," *Science* **177**, 393-396 (1972). https://doi.org/10.1126/science.177.4047.393

[12] See J. J. Hopfield, "Physics, Computation, and Why Biology Looks so Different," *J. Theor. Biol.* **171**. 53-60 (1994). https://doi.org/10.1006/jtbi.1994.1211

analysis of psychology, the facts of the correlations will be necessary, as well as of the facts of brain structure, just as the facts of geography and atmospheric composition are essential to doing weather prediction. Weather on the sun will be easier to predict. You can see why, when I re-visit neural network memory these days, I focus on correlated information and finding useful meaning in correlated data.

PC: [1:04:45] It's a good point, a good idea. I want to give the chance to my colleague and collaborator, Francesco, in case he has questions that I might have missed. I would appreciate.

FZ: Maybe to summarize, it would be useful to have your opinion on the extent, to which spin glass ideas have influenced the development of neural networks and vice versa. It seems to me that during the interview you said that most of the mathematics of spin glasses was not influential for the developments in neural network. Did I understand correctly?

PC: [1:06:01] The core of the question I understood: what's your general impression of the cross influence between spin glasses and neural networks in their development? But I missed the second point as well.

FZ: I had the impression during our discussion that you had the feeling that in the end, most of the mathematics that has been developed in the spin glass field was not so influential for the development of neuroscience or neural networks. Did I understand correctly?

JJH: That is my view. Look, there's beauty in mathematics you can do in the large *N* limit[13]. Biology just doesn't operate in the large *N* limit. It operates somewhat noisily, with large but quite finite connectivity, and connectivity which is spatially long-range but limited. On the other end, the intuitions you can get in the large *N* limit are considerable, and it's important to understand that limit as a source of intuitions about what might happen for finite *N*. But being able to do the mathematics of infinite *N* is not of very direct relevance, particularly when most of the large *N* theory is based on averaging the 'spiking' behavior of neurons prior to taking the large *N* limit.

FZ: [1:07:41] Yes, okay.

PC: Thank you. The last question I have is more of a technical one. I presume you still have notes, papers, correspondence from those times. Do you intend to deposit them in some university archives, or has it already been done?

[13] *N* refers to the system size.

JJH: I don't have as much as you might hope. One thing in existence is, I understand, a pretty good set of notes on the lectures which took place in some of the physics of computation courses at Caltech. The person that I would ask about that is Markus Meister, professor of probably biology at Caltech. He was a physics graduate student at the time of the first lectures. The other person that you could inquire of in this direction is Pietro Perona, professor at Caltech in engineering, and who was the long-time head of the Computation and Neural Systems effort at Caltech after I returned to Princeton. My remaining notes for that one-year Physics of Computation course are astonishingly bad. I would learn so much more than was written in my notes. I took notes only because the process of taking notes focusses my attention. I might have notes on the few lectures that I gave, and some of the notes on the course I developed for many years, but which no longer emphasized the physics of computation.

PC: And did your correspondence just disappear when you moved back to Princeton?

JJH: Since 1982 I have changed offices eight times. Correspondence tended to disappear, and there's the transition from paper to computer, and then from computer A to computer B to computer C with conflicting ways of storage. An awful lot of my primary documentation of this transition period has been lost.

PC: It's unfortunate, but I understand. I've seen part of it happen myself. Thank you so much for your time and for your engagement. It's been really a genuine pleasure to get to exchange with you, and to hear more about your viewpoint and that very special time, I think, in the history of modern physics.

JJH: I'm always happy to help. I hope you find something useful. Thank you for pursuing and trying to get it into an archive. I only wish we could have been face-to-face.

PC: Absolutely. Francesco, did you want to say [something]?

FZ: Thank you very much for pointing out this paper on dreams. It looks very interesting and I will read it because I was thinking about related things recently. So I'm very curious to read in more details what you did, and the relation with other papers. Thank you!