AICSCSHistory_Interview005_GeoffreyHinton

Please note that this transcript has been lightly edited relative to the original audio, in order to improve readability.

SPEAKERS
Vassilis Galanos, Geoffrey Hinton

Vassilis Galanos 00:00
Today is the 14th of June 2024. My name is Vassilis Galanos and I am pleased, very pleased and honoured to be with Professor Geoffrey Hinton today alongside my colleagues, Professor Christopher Williams and Xiao Yang for an interview with Geoffrey, on the topic of AI History at the University. So before we delve into the specifics of your notable career and the evolution of AI, Edinburgh, could you please share an introduction about yourself, Geoffrey, highlighting your major contributions to the field and your personal journey within AI?

Geoffrey Hinton 00:47
Okay, I went to Edinburgh as a graduate student in AI in 1972. I have been working on neural networks ever since. For a long time, they didn't work very well. And now they do work very well. My main contributions were, let's see Boltzmann machines, which were a great idea that never worked that well. Distributed representations. So from the early 80s, I was pushing the idea you should, for example, represent words as vectors of neural activities, where the representation is distributed over many different feature detecting units. I was involved in introducing backpropagation, and showing that it learned good representations for words. In the 90s, I got into variational methods. I think I was the first person to introduce variational Bayes, where you're taking complicated models, where you can't compute the posterior distribution exactly. And you're showing that even though you can't compute the posterior distribution of the latent variables, you can still learn effectively by just using approximations to it. Let's see. I had very good students, Chris was one of them, at both Carnegie Mellon, and the University of Toronto. And in 2009, two of my students, George Dahl, and Abdel-rahman Mohamed showed that neural nets were actually better for the acoustic modeling, the front end of a speech recognition system, than existing technology, which had been developed over a long period. They were only slightly better. But that made a big impact on speech recognition. And by 2012, Google had productized it with lots of clever engineering. And it came out on the Android and suddenly the Android caught up to Siri, if not getting slightly better at speech recognition. Later on neural networks took over the whole speech recognition system. But to begin with, it was just the acoustic model. Then another two of my students Ilya Sutskever and Alex Krizhevsky, developed what we now call Alex net, with my help, in 2012. And that had a big impact. That was kind of the way the dam broke. And finally, lots and lots of people switched to doing neural networks. So that's often regarded as the sort of the beginning of the huge wave of neural networks. Ilya then went on to introduce sequence-to-sequence models with other people, which ended up being very important for machine translation. Ilya was the person who fired Sam Altman. Let's see, I worked for Google for 10 years, half time, from 2013 to 2023. I left Google at the end of April of 2023 because I was old. But I used the opportunity to warn about the dangers of AI, particularly the existential threat. I focused on that because people recognize the other

threats, but a lot of people were saying the existential threat was just science fiction, that these things were just stochastic parrots and didn't understand anything. That was the line taken by good old fashioned AI people. And I thought that was deeply wrong. And I wanted to make sure people understood there really was an existential threat, even though it's longer term. And since then I've been talking to the media a lot about the dangers of AI.

Vassilis Galanos 05:06
This is a great summary. I impressed too, as always honoured to be part of this conversation. So let's turn travelled back to 1972. And your arrival at Edinburgh. So the University of Edinburgh has been historically significant for AI research for several years. So in your view, what factors made Edinburgh the sort of pivotal space and place for the development of these fields during the early days, and what made you decide to come to Edinburgh?

Geoffrey Hinton 05:52
Well, I think it was pivotal for AI because it was the only well-funded AI centre in Britain. The Science Research Council, or whatever it was called then, decided to fund one big centre of AI. They wanted to try and keep up with the Americans I guess. And they decided to do it in Edinburgh. And one of the founding members was Christopher Longuet-Higgins, who had been a very eminent chemist. He was Professor of Chemistry at Cambridge in his early 30s. At the age of about 40, he switched to doing neural networks. And he got interested in things like holographic memories. He did some important work with David Willshaw on binary memories. At about the same time as I arrived, he gave up on neural networks. He was convinced by Winograd's thesis, among other things, that neural networks were the wrong approach and symbolic AI was the right approach. He never liked probabilistic things. He liked things to be neat. He was a mathematician and liked clean, neat things. He didn't like probabilities. And so he and I didn't get along very well. I now think retrospectively, the only reason I survived as a graduate student was because my father was a Fellow of the Royal Society. So he was in the same club as Longuet-Higgins. And it would have been embarrassing for him to get rid of someone whose father was in the Royal Society. So I think it was only white privilege that allowed me to survive as a graduate student. Because all of my interactions with him, almost all of my interactions until very near the end of my thesis, were shouting matches, not always shouting matches, but he was always trying to persuade me what I was doing was nonsense. And he would listen, he would understand what I was doing. But he'd say, that's the wrong approach. And so I would keep promising that okay, if I couldn't make it work in six months time, then I'll switch to symbolic AI. And it got to be six months, and I would need another six months. It's lucky, I didn't tell him “well, it's actually going to take 50 years”.

Before I went to Edinburgh, I'd been working as a research assistant, on a project studying child language development in three to five year olds, to test out Chomsky's theory that it was all innate: language was just kind of unfolding in a maturational way, but you really knew it all already, which I always thought was a rubbish theory. And so we collected a lot of utterances from children in their homes, by bugging them in their homes. It was the same time as Watergate. And we had children in their homes wearing little radio microphones and there was a receiver in the home. But you could actually sit outside the house with a radio, the right kind of radio, and listen to what was going on in the house. And so we collected natural language from children. And then the idea was to analyze it, but nobody had a clue how to analyze it. But we did collect one sixth of a million utterances. So at that time, that was big data. But we didn't know what to do with it. Anyway, that was what I'd been studying before I went to Edinburgh.

So my plan at Edinburgh was to make a system called LUDWIG. Partly in honour of Wittgenstein, I didn't know about Boltzmann then, but mainly because it stood for Language Understanding Device Without Innate Grammar. And that's what I was going to do for my thesis, which as you can see doesn't fit very nicely with Winograd's thesis. And the first thing to do was get neural networks working. And so one of the very first things I read in Edinburgh I think, was a big fat book by Rosenblatt called Principles of Neurodynamics.

Rosenblatt has been unfairly criticized. He understood a lot about the limitations of things that only had one layer of adaptive weights. He put a random layer in front of it, which doesn't really make much difference to what it can do. And so he had this perceptron convergence procedure that was guaranteed to find a set of weights that would solve the problem *if such a set existed*. That was the problem.

So Minsky and Papert in 1969, three years before I arrived in Edinburgh, had shown there were strong limitations on what you could compute that way. For example, you couldn't compute whether two inputs were the same. Unfortunately, they expressed this as you couldn't compute XOR that is whether the two binary inputs are different. I think if they'd done it the other way around and said you couldn't compute if the two inputs were the same, it might have drawn people's attention to the idea that it would be nice to have a primitive operation that compared things and looked at their similarity. In fact, that's what transformers have. Transformers have as a sort of primitive operation of how similar are these two vectors, and that makes a big difference to the power of neural networks. Anyway, I'm rambling.

I came to Edinburgh with that dream. I spent a long time trying to go beyond perceptrons. Everybody at Edinburgh with the exception of David Willshaw, said why are you doing this, Minsky and Papert have shown it's nonsense. Look at Winograd's amazing thesis. He has things that can actually understand language. There is an interesting aside there, which is that AI people back then were convinced that if you could say “Take the block that's on the green block, and put it in the blue box” and in a simulation, a robot could do that, that proved it understood. Now, of course, people who want to say neural networks don't understand anything, won't accept that as evidence it understands. They'll somehow argue it's just statistical correlations it's using, as if statistical correlations aren't understanding.

I had many years of arguing with Christopher Longuet-Higgins, which was fairly good discipline for me, I guess. Then he left and went to Sussex, after I'd been in Edinburgh for two years. And his group, Mark Steedman and me and Steve Isard who was in effect a postdoc with him, all went down to Sussex. While I was a graduate student at Edinburgh, Christopher and I got along rather badly, as I've mentioned, and Steve Isard, basically took over as my advisor. He was like a co-advisor. And I got along with him much better. And he suggested that if I really wanted to see if neural networks worked, maybe I should choose a simpler problem, like recognizing handwritten characters. And I spent many, many years doing that. But that was the suggestion of Steve Isard's when I was a struggling graduate student. He was he was very, very helpful.

So then we all went down to Sussex. I remember how much nicer it was in Brighton than in Edinburgh. It was warm. And there weren't any people talking about Sassenachs. Because when I was in Edinburgh it was the height of the emergence of Scottish nationalism. England was busy stealing their oil. And Scotland would have been rich if it had been able to set up a sort of sovereign fund with the oil, like Norway did. But the English stole it. So I was quite sympathetic to Scottish nationalism, but I was happier in England, particularly since it was warm.

Geoffrey Hinton 15:05
So, where do you want to go next? Oh, when I was in Edinburgh, apart from David Wilshaw, everybody said neural networks was a waste of time. All the other graduate students said, Why are you doing this? Minsky and Papert proved it was nonsense. Now Minsky and Papert never proved it was nonsense. They showed that the simple kinds of networks with a one layer of adaptive weights for which you had a guaranteed learning algorithm, had all sorts of things they couldn't do. They didn't show that you couldn't do those in deeper networks. And they didn't show there couldn't be a learning algorithm with deeper networks. They just didn't know what it was. And they sort of implied you'd never find one. And indeed, we never did find one in the sense that they wanted. So what they assumed was a learning algorithm, I'm fairly sure about this, I'd have to go back and look at the book, a learning algorithm would be something where it could guarantee to find the answer, there'd be some sort of guarantee like there is with algorithms like A-star. And we never did find that, we just found an algorithm that would find pretty good solutions. Part of this assumption that you had to have a guaranteed algorithm was, well, if you just find pretty good solutions, you get trapped in local optima. And people believed that for many, many years, never actually doing empirical tests of whether you did get trapped in local optima. They would never go and take the solution, look at the curvature in all directions and check it was a local optimum, rather than a plateau or a saddle point, for example. They just made this assumption, because it's so plausible that they will get trapped in local optima. It's a nice example of how a myth survived in symbolic AI for years and years and years, with no evidence, other than the fact it's obvious. But nobody actually looked. And one thing I learned about neural networks was, every time you make a new way of displaying what's going on, you discover that your beliefs about what was happening were wrong. So when you display the weights, you suddenly discover weird things happening, when you display the activities, other weird things are happening. Okay, ask me another question.

Vassilis Galanos 17:19
Very interesting. No, and you have been covering many of my questions in different ways. I really appreciate that you also offered the political context behind that was surrounding these technical debates. This is extremely useful, especially the parts about Scottish nationalism and oil. This is something I think that is not sufficiently highlighted when we read histories of AI. I recently found out that Rosenblatt and McCarthy were high school classmates. So essentially, their rivalry extended to, to do what you have been exposed to. So these are all fascinating aspects. I want to ask you more about the research environment, besides the debates you had about disbelievers and critics of neural networks. How was the research environment in Edinburgh from other recollections, we have, you know, people talk about a very vibrant and interdisciplinary community. I wonder if you encountered that or maybe because of your, for the time heretical approach, maybe you didn't really see that. What was your impression about the research culture at Edinburgh University,

Geoffrey Hinton 18:36
It was it was a good research culture. People collaborated other people. We had a great computer. It had, I think, to begin with, when I got there, it had 64 kilobytes of memory. And the 64 kilobytes of memory was shared by about 40 researchers. So you got more than a kilobyte each. It didn't really matter because we had virtual memory. We had this language called POP-2, which was like a kind of Lisp with Pascal syntax. And so not all those brackets. And it didn't matter that there was only 64k of memory, because we had a disk with virtual memory. And the virtual memory was practically unlimited. I think there were two megabytes of it. So it's all the memory you could possibly want. And I was trying to make neural networks on that.

I remember the students used to operate the machine at night. During the day there was an operator, and at night students used to go in and operate it. There was a sort of mat you had to wipe your feet on, because you didn't want any dust and stuff in there. And there was a speaker that made a click every time the program counter decreased. So if you're in a loop, every time you went back, there's a click. And so if you're in a tight loop that made a tone. And so normally what the computer would do is, it would be sort of humming, because it was in a tight loop waiting for input. And then if you gave it input, so it would go "whoom". And then you give an input again, "whoom-whoom-whoom" as it went round other loops. I remember the first time I gave it a neural net, when I was in the computer room listening to the speaker. So it's going "whoom". And I gave it this neural network and it went "ah-ah-ah-ah-ah", it really didn't like it. And it went like that for hours because I was trying to learn something, and it was very slow.

Then, in my second year, I think, probably so probably 1973, we got new memory. So the 64k of memory was actually little magnetic rings, with coloured wires going through them, that was all threaded by hand. And so that's 64k. They were probably 16 bit words, I don't remember, possibly 32 bit. Actually, it's an old computer, so it's possibly 29 bit or something who knows. So that was 64k of those, then we got new memory. And I think we got another 128k. So the total was now 192k. And I remember going into the computer room, and after the new memory had arrived, expecting to see dozens of filing cabinet size things. So the 64k was a whole bunch of filing cabinet size things. And I couldn't see where the new memory was. And then someone opened the cupboard doors underneath the control terminal. And there was a big board. My memory is it was about a foot square, it's probably less than that. And that was the new memory, just this one board. It was amazing. It was twice as much memory as before, and it was all on a single board. Then we had of course heaps of memory. So yeah, that's what I computed on to begin with. And on that computer, I think probably once I had the 192k of memory. Um, one of my first projects after I got some the perceptron convergence procedure working and the least mean squares procedure working, which are both for one layer of adaptive weights. Everybody was bugging me, all the other students were bugging me. And some of the faculty were bugging me saying, look, neural networks are no good because they can't do recursion. And because everybody used Lisp or POP-2, a Lisp like language, they thought recursion was the essence of intelligence. And so neural networks are never going to be intelligent unless you can do recursion.

I sort of agreed that well, there must be a way for neural networks do recursion. And the point about a neural network doing recursion is, if you want it to be true recursion, you have to be using the same neurons and the same weights for the recursive call, as you are using for the thing that made the recursive call. And that raises a memory problem, which is what happens to all the knowledge about the thing that made the recursive call when you reuse those neurons for doing the recursive call? You don't want to have copies of neurons. That's crazy. That's biologically crazy. So it seemed to me you have to have some kind of memory that will remember what you're doing before the recursive call, and will reinstate it after the recursive call. It's only working memory, but you don't want to do it in neural activities because you don't want to copy neurons. If you want to do five deep, you'd have to have five copies of everything. And that seems crazy. So I decided what you needed was an associative memory that would act as a stack. So the idea is your all of your variables are the states of groups of neurons. And when you make a recursive call you need to replace those states by the states of neurons in the recursive call, and then you need to be able to get them back again.

So I implemented a system that had a short term associative memory, where it associated together, states of different groups that are active at the same time. So it could do pattern completion. And it had a level counter, which was at what level you are at in your recursive calls. And it will automatically, as it ran, just be associating states together. And it will be also be associating them with the level marker, which said what level you're at in the recursion. And when you want to do a pop, when you want to return from the recursive call, what you'd do is you'd subtract one from the level counter. So instead of being at level four, I'm now at level three, you will put everything else into a neutral state that said "I don't know". And then you give it a few iterations, so that the level counter could then propagate and try and fill other things in. And they would also help fill each other in. And you'd return to the state you were in before, except that you'd have incremented the program counter by one for the high level call, because you've already done that recursive call. And now you need to do whatever the next thing was. And I made all that work.

I made it all work for a simple problem where you were drawing capital letters. To begin with it was non-recursive. So you'd say at the top level, draw me an H. And it would know that to draw, and you'd also say, what size and in what position and what orientation. There are only four orientations. And the question is, could it now output a sequence of instructions to draw the individual strokes, which involves knowing the size and orientation and position, and I got that working. And then what I did was took a capital letter I. And instead of defining an I as a stroke, up the middle, a stroke across the top and a stroke across the bottom, I defined a capital letter I as a capital letter T with a stroke across the bottom. So now, the first thing you do when you want to draw an I, is forget about the I and draw a T. But then when you finish drawing the T, you have to remember what it was you were doing. And so you have to pop back, get the parameters of the I, realize you've done the first thing and ask what's the second thing I need to draw now? And I actually got all that working.

The first talk I gave to the research group, in 1973, was about that, it was about how to do recursion in a neural net by using a short term associative memory. And I don't think anybody understood it. The experience I got was, why is he talking about this nonsense with neural nets? I wasn't very good at giving talks. So we have no idea what he's doing. It's something to do with recursion in neural nets. But why bother? I mean, if you want to do recursion, why not just do recursion on a computer? This is crazy.

Geoffrey Hinton 28:24
About 50 years later, it's now publishable I think. And a couple of years ago, I discovered the printout from the teletype where it was actually doing the drawing and what all the groups of neurons were doing. You had to tell it what every group should do at every time. So it's in that sense fully visible in order to train it because we didn't know how to train hidden units then. And so I would spell out for it what every group should be doing in every instance in order to do this, but I made it work. And I keep thinking about, now that it's 50 years old, 50 years old last year, I thought maybe I know that it is antique, I should actually finally publish it. But to do that, I've only got the printout. So what I want to do is reimplement it in MATLAB, so I've got a working system. Anyway, that's what I started off doing in Edinburgh. And there was no audience for it.

Vassilis Galanos 29:26
Fascinating, and my interpretation of it is that you have essentially invented back propagation in this way.

Geoffrey Hinton 29:34
No, it wasn't back propagation. I wasn't training with back propagation. What I was using was an associative memory with fast weights in order to do short term memory so you could do recursion, or to do the stack. But I trained it by showing it exactly what states things should go through at every time for every group of neurons. Instead of having individual neurons I had groups of neurons that had multiple alternative states. So it's what physicists call a Potts glass. And the first thing I did was showed the perception convergence theorem applies to multi state things too. I got very excited by that, that was going to be my first paper. And so I wrote the paper. And then people pointed out, you should have references in a paper. So I decided I'd better put references in it. And so I went and looked at some of the standard books on machine learning. And I found a book by Nils Nilsson who was really a symbolic AI guy. And unfortunately, it had a proof of the multi-state version of perceptrons. So that was the end of that paper. But it did me good to reinvent it myself.

Actually, one useful thing Christopher taught me was not to read the literature. So Christopher's view was, he had a saying, reading rots the mind. His approach was, come across a problem, figure out how you'd solve it. And after you figured out how you'd solve it, go and read the literature to see if that was already known. This isn't really advisable as a research technique, but it fits naturally with my personality. I don't like reading instructions. So it'll work if you've got good friends who do read the literature. So in neuroscience, for example, I seldom read the original literature. But I talk to Terry Sejnowski, who's read everything. And that's fine. But I think it helps a lot to think how you would solve a problem before you figure out, before you've read how the problem is normally solved. If you like puzzle solving, it's great. So my approach to sort of research problems is no, no, no, don't tell me I want to figure it out myself. It works for some things, not for other things.

Vassilis Galanos 32:05
That's what my high school teacher in mathematics also told me. Fascinating. So when you arrived, what kind of initiatives were people working on? What kinds of research questions

Geoffrey Hinton 32:20
It was nearly all symbolic AI. The were some lectures I went to, there were no exams. So as far as I know, I've never done an examination in AI or computer science. There were lectures you went to that were kind of voluntary. And I remember going to lectures about logical approaches to AI. There was something weird called skolemization, which I never really understood. But skolemization was very big back then. Some of the people there like Bob Kowalski were people very much involved in things like Prolog later on. I had a very good friend then called Mark Steedman, who was another student. He was just finishing up with Longuet-Higgins doing music, but symbolically.

Geoffrey Hinton 33:18
Mark Steedman moved to Sussex with us. He was one of that group of people who moved to Sussex, when Christopher left Edinburgh. Yeah, after Mark had left Sussex, Christopher was still working on music. He was a very able musician. I remember, it must have been about 1975. So after I've been his student for more than two years, I had a little office in Sussex. And he came into my office one day and said, Geoffrey, I've been working on this program that's parsing music. And I'm having difficulty writing this program. Maybe you could help. And so he described what he was doing. He just wanted somebody to talk it through with. I'm surprised he chose me, but there you go. Maybe there was nobody else around.

So he explained the problem to me. And I said, what seemed obvious to me, which was, well, you need a function that will return two results, and then it'd be easy to write it recursively. And he looked at me with amazement, and said, Geoffrey, that's really rather clever. How come you can do this and you can't do anything else? And then he said, Oh, I'm sorry. I didn't mean that. But he clearly meant it. He was clearly very surprised that I'd got a clever solution he hadn't seen, not very clever, fairly obvious if you program, to how you wrote this parsing program. But he revealed his opinion, which was I just couldn't make anything work, which is another piece of evidence that I think I only survived because of white privilege.

Vassilis Galanos 35:13
So I know Christopher is looking forward to the barriers and the obstacles question. So I will leave my intended question for less notable figures for later. So 1972 And what is 1973? We know it has been tumultuous period for AI in the UK but also globally. But so I'm obviously pointing to the Lighthill report here. But this is also an open question about obstacles and barriers in the history of artificial intelligence, and I'm tossing of the ball to you to talk about any kind of obstacles and barriers you think are relevant in the history.

Geoffrey Hinton 35:56
So I think you can't understand the history of AI without understanding that there's a public school, that is a private school, called Winchester which is much more intellectual than Eton. Very upper class. It's the kind of Yale to Eton's Harvard, if you see what I mean. It's not quite as good, not quite as expensive, but more intellectual. And there was a class there. Now, this is all verbal information probably from David Willshaw. So you need to check it, but there was a class there, in which there were three students. There was Christopher Lounguet-Higgins, James Lighthill, and Freeman Dyson. And the problem was Lighthill took mathematics, and Dyson took physics. So Christopher had to do chemistry. But that meant that Lighthill and Christopher, they're both in the same club -- the Royal Society, had known each other since high school days. And I think they respected each other. And so one interesting thing about the Lighthill report is there's a little out in it, which basically says AI is rubbish, unless you study how the brain works, in which case it's okay. And that sort of let Christopher out. Now, I basically agree with a Lighthill report. This won't make me any friends among symbolic AI people. So what was happening was symbolic AI people like McCarthy were saying, look, anything that can be computed can be computed symbolically. So there's nothing we can't do, which completely ignores the issue of complexity. People weren't thinking much about complexity then. I think all this complexity theory came along later. And so McCarthy said, you know, we can do everything this way. And Lighthill, who was a control theorist, that's one of the things he was eminent for, said, you can't, because you can't do it efficiently. And if there was an assumption that both of them made, I think, which was a plausible assumption, which was computers won't get a million times faster. Back then it would have been plausible computers might get 100 times faster, even 1000 times faster, but not a million or a billion times faster. And under that assumption, I think Lighthill was completely right and McCarthy was completely wrong.

Now, Lighthill was wrong about one thing, which is he attributed people's desire to build robots to the fact men couldn't have babies. That was kind of stupid politics, of the kind you might expect from a nerd. It was kind of irrelevant and weakened his case, but he was right about the brain having this intricate neuropil, I think he called it, but lots of connections. And we've no idea how it works, but it's capable of immense amounts of computation. And so he was sympathetic to neural networks. I don't think he thought they'd ever be able to learn everything, but I think he was basically right. And it pretty much destroyed AI in Britain. Not totally destroyed, but it meant there weren't any faculty jobs in AI. In fact, there was one faculty job in AI, which Harry Barrow got. Steve Isard applied for it. Harry Barrow got it. And then I think later Alan Bundy got it. But when I when I was writing my thesis, there were no AI jobs. Mark Steedman, for example, couldn't get an AI job, he had to get a job in psychology. And that was the result of the Lighthill report. So I think Lighthill was kind of 50 years ahead of his time, or 30 years ahead of his time in understanding that symbolic AI wasn't the way to go.

Vassilis Galanos 40:26
This is fascinating. The combinatorial explosion reference was also very useful here.

Geoffrey Hinton 40:34
Yeah, so later on people like Michie understood that there is combinatorial explosion, they just didn't know what to do about it. They didn't understand that you get combinatorial explosions when you don't have distributed representations. Once you have distributed representations, you can get around a lot of combinatorial explosions by basically sharing features.

Vassilis Galanos 40:59
I guess a follow up question. I have on that. And again, within the context of barriers and obstacles, something that has been recorded in various types of official AI histories that after the Lighthill report and the so called AI winter, AI researchers have been rebranding what they did. They gave it different names. And I want to ask your own experience of that with respect to what you were doing the neural nets approach. Back then, did you consider neural nets and machine learning and connectionism as part of AI? Or did you always think that it was something different? And how did you navigate that environment?

Geoffrey Hinton 41:44
Yeah, that's a very good point. So back then, in the 70s, and to some extent, in the 80s, there were two different approaches to intelligence. And AI meant symbolic AI, AI meant the kind of thing that Newell and Simon or McCarthy did, or that Minsky did later. And neural nets meant learning connection strengths in a neural network, which was what Minsky tried to do for his thesis, I think for his thesis, using reinforcement learning early on. And so neural nets and AI were opposing approaches. And it kind of irritated me that when neural nets really started working, people started calling them AI. I tried to get people not to call them AI to call them neural nets, and say it wasn't AI it was something different. That was hopeless, the politicians and the funders and so on, and industry wants you to call it AI. And now most people think AI is neural nets. But that's not what I meant in the previous century. One reason it irritates me is because now governments give large amounts of money for AI, and the good old fashioned AI people say hey, that's money for AI we should get some.

Geoffrey Hinton 43:09
That's not what the money was for. This, you can see leads to a lot of politics.

Vassilis Galanos 43:15
This is very interesting. It escapes a bit the scope of Edinburgh, but I know that after you left Edinburgh, you have had a brief period of collaboration with Alan Newell. So I wonder how the connections between Edinburgh and Alan Newell were at the moment how did you manage to

Geoffrey Hinton 43:43

I don't think there was much connection between Edinburgh and Alan Newell. So after Edinburgh, I went to Sussex, and after Sussex, I went to the University of California San Diego as a postdoc. And then I briefly went back to Cambridge, to the Applied Psychology Unit, which is the kind of place you had to go if you wanted to do AI. And then I went to Carnegie Mellon, where I worked with Alan Newell, a bit.

Alan Newell was an interesting character. He did not believe in neural nets at all. But he believed in being eclectic. So he thought Carnegie Mellon, which was a big AI place, it was good that they had one person doing neural nets. Because who knows that crazy stuff might eventually work. So he and I actually got along quite well, I think, even though we were very, very different characters. He was a very upright Christian. I remember going to dinner at his house, and he would, in his own house, he would go and pull the chair out for his wife so his wife could sit down. I'm pretty sure they went to church on Sundays. We were very, very different in politically. But he was intellectually very honest. And although he thought neural nets was nonsense, he wasn't certain they were nonsense. And he thought it was good to have somebody doing it. We actually had a lot of good conversations. We once taught a graduate course together. And we had a lot of conversations about analogical reasoning.
There were two big problems for symbolic AI that he recognized. One was that in speech recognition they'd had this big Blackboard system, which was basically symbolic AI with a big working memory, and everything was done in propositions and production rules. And it didn't work as well as what he thought of as a really dumb system, which was a hidden Markov model. And the point is a hidden Markov model had a learning algorithm. And even though it was a fairly weak form of representation of sequences, the fact that it had a learning algorithm made it better for speech recognition than these great big Blackboard systems. And Alan Newell, instead of dismissing that, took it seriously. He said, you know, this is something we need to worry about. These aren't his words, but how come this dumb little thing with a learning algorithm beats our great big sophisticated AI systems? He took that seriously. He also took seriously analogical reasoning. And AI had difficulty doing analogical reasoning.

I have a nice example that I use, to the popular press about analogical reasoning. Here's something logic can't do, but people can. And it turns out there's a difference between men and women, surprisingly. So suppose I told you that you have a choice between two alternatives, neither of which is very biological. Alternative one is that all cats are male, and all dogs are female. And alternative two is that all cats are female, or all dogs are male. Now, if you ask a man, which of those is more plausible? A man will say it's obvious that cats are female, and dogs are male. Presumably, because dogs are big and noisy and chase cats. Women, it turns out, aren't so sure. There are some women I've given that problem to who don't see it as natural that dogs are male and cats are female, which surprised me a lot. Anyway, the point is, you can't get that by logic. The whole thing is illogical in the first place. But you can get that if you use big vectors to represent things, because it's obvious, at least to males in our culture, that the big feature vector for a cat is more like the big feature vector for a woman than it is like a big feature vector for a man. That shows up in a lot of the language we use to. So yeah, that's an example of analogical reasoning. And Newell was very puzzled about analogical reasoning. It didn't really fit with his view of the world.

I also talked to Herb Simon at CMU who really believed in sequential computation. And I remember having conversation with him. I was trying to convince him look, there's some things you have to do in parallel, like recognizing a letter. So letters have features and we use all the features in parallel to recognize the letter. And I remember him arguing that well, how do I know you didn't do it sequentially? Which seems bizarre now. But there you go. So yeah, I got along very well with Newell. He had a lot of integrity. I always contrasted him with Minsky. I got along well with Minsky, too, I think. It's hard to know with Minsky. Minsky was always a child -- he never was a grown up. He was always fascinated by things. He was very bright, but not very thoughtful about other people. Whereas Newell was very responsible. Very, very different characters

Vassilis Galanos 49:25
Very interesting. And I know that in the sort of birth years of artificial intelligence, Newell was always suggesting that the endeavour should be called complex information processing, as more exact and responsible. Now these are these are all fascinating. I want to go back to the Edinburgh context a bit. We've spoken about differences between men and women as well as white privilege. And and I'd like to ask you, whether you remember from your time in Edinburgh, other figures who might have contributed to the field or have the potential to contribute to the field that you might have met or remembered, but they haven't received the widespread recognition other people have have received. People that we should look for, for future historians, and so on.

Geoffrey Hinton 50:19
So one is clearly David Willshaw. So David Willshaw believed in neural nets and made this very interesting memory called well, I think it's called a Willshaw net. It's got binary units for the memory, binary connection strengths that are either on or off. What's interesting about the Willshaw net is, if the Hopfield net had been invented first, so Willshaw did that work in probably around 1970. I think his thesis probably around that time, maybe 69, maybe 70-71, that kind of period, whereas Hopfield was 10 years later. But Willshaw's work could easily have been seen as an advance upon Hopfield, on Hopfield nets. And I don't think Hopfield referenced Willshaw at all. I'm not sure about that. So he always believed in neural nets. And Willshaw also understood the problem, that Longuet-Higgins and Buneman and Willshaw called the four point, maybe the three point or four point problem. And it was a binary version of hidden units. So they were thinking about if you have binary variables. They didn't like probabilities. They wanted things to be kind of all or none. So if you have binary variables, and you want there to be structure, you mustn't have all four combinations of the binary variables, you know, 00,10, 01, and 11. And I think that was what they called maybe the three point condition that at most three of those had to be possible for there to be structure. The way we would say now is, you've got to have correlations for there to be structure. But they were kind of dealing with the idea of correlation in a binary world where there are no probabilities. And there's not going to be any correlations if all four conditions occur. And so, when I went to Edinburgh, either Peter Buneman, or David Willshaw explained this to me. And that was sort of the hidden unit problem. That was their funny binary version of the hidden unit problem. They would say you had to put it in by hand, I can't remember the details of it. But I remember, that's when I started thinking about hidden units. Because of this 3 point condition.

David Willshaw then got more interested in real neuroscience and did interesting work on how you develop connections between the retina and the tectum in a frog. And did nice work on that. But I don't think David Willshaw got as much credit as he deserved.

Vassilis Galanos 53:34
Thankfully, I interviewed him, he was the first person to interview for this project. So it's great. Have you got any other people that come to mind? Of course, if not, it's absolutely fine, because

Geoffrey Hinton 53:48

There were a whole bunch of other students there. I never got to know any of them that well. One of the students at the same time as me was called Austin Tate, who did well in symbolic AI and I met him again when he was a full professor in Bristol many years later. And a whole bunch of other students who I met in odd places many years later, but I never knew any of them that well. I knew Mark Steedman very well. And David Willshaw. And Richard Power, I knew Richard Power quite well. He did do symbolic AI. And he I think he was probably the same year as me. And there was always a big contrast between him me. So I was this obstinate person who insisted on doing neural nets. And Richard Power was kind of the model of what Christopher wanted. He was clever. He was a good programmer. And he worked on symbolic AI and natural language understanding using symbolic AI, and did nice work on that.

Vassilis Galanos 54:55
Fascinating. Thank you very much.

Geoffrey Hinton 54:57
Years and years later on. In Sussex, I had a girlfriend for a while called Felix. And she later married Richard Power. And then I met her years later when I was in the Applied Psychology Unit in Cambridge in the 80s. And she had a daughter with Richard Power. And when she introduced me to her daughter, and said, my name was Geoffrey, her daughter burst out laughing. And I asked her why her daughter was laughing? And she said, well, they had a large and stupid dog, and it was called Geoffrey.

Vassilis Galanos 55:52
Well, that's, that's a great anecdote. About the history of symbolic AI and neural networks. Fascinating. Well, thank you very much. We are nearing the hour of this conversation. My last kind of very big and vague and abstract question, but I think very important, given the latest phase of your research that you outlined at the beginning of this conversation, is sharing your visions for future researchers of artificial intelligence. Based on your experiences in Edinburgh, but also beyond, what are the lessons you think that contemporary AI researchers should keep in their mind, both in terms of visions and innovation, but also in terms of barriers, and, and obstacles, and so on.

Geoffrey Hinton 56:46
So I like to look for problems where I think everybody else is doing it wrong, or where it's very ill defined, and someone needs to get in and sorted out. I think at present, there's a huge problem to do with what it means to be a being. So for example, Yann LeCun thinks developing super intelligent AI isn't going to be a problem. Since we built them, we're going to be able to make them do what we want. I don't think that. But a lot of it hinges on what it is to be a being, I think they're going to be beings. And there's a question of, could you have a being that's super intelligent, but only does what a less intelligent being wants? Or will it be the case that once you have these AI agents, they'll start developing wants of their own? So a good example would be if suppose I made a real robot that got out there and was intelligent, you'd have to build into it some sense of self preservation. It wouldn't want to injure itself, it would want to preserve itself. That has to be built in. That's the beginning of a slippery path to it having his own desires. Also, I think you'll have to give the ability to create sub-goals. And it'll pretty soon develop the sub-goal of getting more control, because that makes it easier to do everything else. I think the most urgent problem in AI at present is, can you develop super intelligent beings and keep them under control? And nobody knows how to do that. I think it's improbable, but maybe possible. But that seems to be the most urgent problem in AI. Now, we may never get there because some of the other bad uses of AI may wipe us out first.

Vassilis Galanos 56:50
Yeah, fascinating. I think when I took a look at your PhD thesis, in the acknowledgement section, I saw the name of Aaron Sloman. Yes, some of his writings, he's been saying that in the case, we are able to create highly intelligent artificial beings and discriminate against them that would be specialist.

Geoffrey Hinton 58:46
I should have mentioned Aaron because when I was in Edinburgh, he was a visiting fellow there. And he was one of the very few people who took me seriously. I had good conversations with Aaron there. And then later on, when I finished my or was finishing my thesis with Christopher, I got a postdoc job with Aaron in Sussex. And he was always very nice to me, although we had very different views about how intelligence was going to work. But yeah, he was very important. He helped me a lot in Sussex and in Edinburgh.

Vassilis Galanos 59:42
Yeah, I guess a bit of a follow up on the future vision, given the contemporary historical context that this interview is taking place in. Would you advise contemporary researchers in terms of their association with military uses of AI technologies? Currently we've we're facing the Israel-Palestine and Russia-Ukraine conflict. There's lots of debate around the misuse of AI technologies in warfare. And I wonder what's your message to new generations around that?

Geoffrey Hinton 1:00:19
I think it's fairly clear that all the major defence departments are working on killer robots. The US has explicitly said, it's always going to be under human control. But that doesn't mean sort of from moment to moment. In other words, it's going to be able to make autonomous kill decisions. That's what that means. And they're hoping to replace a lot of their soldiers by robots in the not too distant future. If you look at all the regulations around AI, they always have a little clause in them that says none of these regulations apply to military uses of AI. The European regulations explicitly have a clause that says that.

It's clear, we're going to get very nasty killer robots, or killer drones. And we're not going to get any effective regulation of those until after very nasty things have happened. It's going to be like chemical weapons in the First World War. After the First World War, they were so nasty that people could agree to have Geneva conventions about them. That pretty much have held. I mean, Putin isn't using chemical weapons in Ukraine right now. Maybe he would if he got really desperate. And they weren't used much. They were used by Saddam Hussein, when the Americans were on his side. And the Americans didn't censure him for it. He used them against the Kurds during the Iran-Iraq war. The British didn't censure him for it either, but they would have been a bit embarrassed, because Winston Churchill had actually authorized the use of, well, some people say he'd authorized the use of chemical weapons against the Kurds. The historical evidence for that isn't that good.

Geoffrey Hinton 1:01:59
It's clear the British, they're not in a position to complain to other people about waging warfare in nasty ways.

Vassilis Galanos 1:02:11

Very interesting. Well, we are now exceeding the hour of this conversation. I'm always tempted to ask a million other questions, but I think we have covered the vast terrain. I want to give you the opportunity to add anything else, if you want to maybe something you think is worth covering, but we didn't have the chance came,

Geoffrey Hinton 1:02:37
Here's my primary advice for young researchers, Find something where you've got a sense that everybody else is doing it wrong. There's something wrong with the way people are thinking about it, and see if you can figure out what you think is wrong with it, and fix it.

Vassilis Galanos 1:02:56
It's a great way to end this conversation, I think. And not just for AI researchers. But I think for any researcher today, navigating the funding landscape, and convincing people that sometimes crazy ideas are worth exploring. Well, thank you very much for sharing invaluable insights, Geoffrey, and your experiences. Your contributions have shaped the landscape of AI and we are really honoured and pleased to have spoken with you today. Thank you.

Geoffrey Hinton 1:03:39
Thank you, bye for now.