**Peter Voss:** 0:00

You want some grounding. You want grounding in senses and also you really do need general intelligence, you do need some kind of a vision. It doesn't need to have the same sense of security as humans, necessarily, but you do need vision, I think, to have a sense of a 2D, 3d world and also action movement. I think there'll be a time when it becomes really, really obvious to pretty much anybody who works with a system, saying, wow, this is the real thing. It really can learn by itself, you know, and it still needs a bit more. How long will it take us to get there? Well, I usually answer that question. It's not a matter of time, it's a matter of money.

**Craig Smith:** 0:36

Hi, I'm Craig Smith and this is Eye on AI. This week I talked to Peter Voss, a long time researcher in the field of artificial general intelligence. He was one of the authors of a book titled Artificial General Intelligence, which helped popularize the term, and Peter has a company called Ngoai AIGOo which builds intelligent chatbots for industry. But Peter's real ambition is to build artificial general intelligence using cognitive architectures and graph knowledge bases, a strategy very different from the pre-trained transformer architectures in use today and which many commentators have argued will never get us to AGI. I hope you find this conversation as fascinating as I did. This episode is sponsored by Datastax, the real-time AI company. With Datastax, any enterprise can mobilize real-time data and quickly build smart, high-growth applications at unlimited scale on any cloud. Companies building real-time generative AI apps can leverage the Datastax vector search capabilities to build LLMs, AI assistants and more. Sign up now at the link in our show notes.

**Peter Voss:** 2:19

I'm Peter Voss. I'm CEO and Chief Scientist of Igoai and I got into AI. Actually it's quite a long journey. I started off as an electronics engineer, started my own electronics company, then fell in love with software and my company turned into a software company. I developed a comprehensive ERP software package and that company was quite successful. We went from the garage to 400 people and did an IPO. So that was great fun and I learned a lot, made a lot of mistakes and made some money. And when I exited that company it really occurred to me what project do I want to tackle now? And it was obvious that software is not very smart by itself, that it makes mistakes. If a programmer doesn't think of something, it'll just crash or give you an error message or something. So that software really doesn't have common sense, it can't reason. So that's what I embarked on more than 20 years ago to figure out how to bring intelligence to software. So I took five years to study all sorts of aspects, all different aspects of intelligence, starting with epistemology, theory of knowledge. How do we know anything? What is reality? How can we be sure of things? And things of that nature and then also about cognition is what IQ tests measure, how do children learn, how does our intelligence differ from animals, and things like that. To really deeply understand what intelligence is all about, I obviously also studied work that had already been done in the field of AI and then by 2001 I was ready to start my own AI company and we spent several years basically doing R&D, taking the ideas that are formulated and turning them into actual code, into an actual framework, which we subsequently commercialized in a company called Smart Action to automate phone calls intelligently, basically with a brain, and my current company is basically the second generation of this technology. We took some more time to crank up the IQ of the system and current company, igo AI. We call it this chatbot with a brain, as opposed to all of the other chatbots out there that really don't have a brain. And that's where I am today and really the mission I'm on is to develop AGI, to develop full human level, general intelligence.

**Craig Smith:** 5:00

Yeah, and we spoke last time, Aigo is based on knowledge graphs. It's not a neural matter of a deep learning model, is that right? And just when you talk about a brain, a chatbot with a brain, the brain you're really referring to the knowledge graph database, is that right?

**Peter Voss:** 5:25

Well, the knowledge graph is only one part of it, a very important part of it, but it's really the substrate of our brain. It's how short-term and long-term memory are encoded. You know that all of the knowledge and skills are encoded. I look at it as much as a neural network, but in the sort of original, traditional sense, as well as a knowledge graph. So it's a combination. You know it's nodes and links that basically encode relationships between concepts. It's very, very different from what almost everybody else in the field of AI is working on. So what most people are working on is statistical AI or, more recently, generative AI, which is basically big data, big compute, you know, number crunching, building these large read-only models, and that's really where you know all the visible progress has been made and all the money has been made and all the oxygen has been sucked out of the air with that approach, basically statistical, big data approach, whereas my approach, our approach, has really, from day one, been a cognitive AI. So the starting point is what does intelligence require? So the knowledge graph, that is, as I say, the sort of substrate of our brain, just encodes the knowledge and skills that it has, but the various algorithms and capabilities that the system has, as you know, deep understanding, deep passing, context reasoning. You know short-term memory, long-term memory, language generation. You know all the various skills that the system has. They basically work either part of the skill set that is in the knowledge graph or they are additional algorithms that work on the knowledge graph. So the brain itself really is, it's a cognitive architecture that has all of the components required for cognition.

**Craig Smith:** 7:26

Yeah, on cognitive architectures, maybe you can talk a little bit about that and what they are and what some of the popular architectures are and what architecture you use.

**Peter Voss:** 7:42

Yes, so cognitive architectures have actually been around for quite a long time, for quite a long time for, you know, 30 or more years. But they sort of are based well, they're different camps. Some of them are based on logic, programming, on, you know, logic, inference and having logic as their base. Others really started off with biology and saying sort of neural networks and you know how the brain works and can we simulate that in a cognitive architecture. But the common denominator is basically that cognitive architecture tries to embody all of the requirements for cognition. So it's typically, you know, input, there's input processing and then there's output and that loop is closed and then there might be some higher level reasoning involved. Now, you know, when I mentioned cognitive architectures, people often say, well, we've tried that for 30 years and it hasn't worked, you know. But of course, go back 10 or 15 years and people would have said exactly the same thing about neural networks. You know, hey, we've tried neural networks for 30 years and they haven't worked. Well, sometimes things don't work until they do. And there are actually several reasons, I think, why cognitive architectures haven't come to the forefront. One of them already mentioned is that the success of statistical big data approaches has. You know, it has been so successful that it sucked all of the oxygen out of here and basically, if you want to get something funded, it's got to be, you know, statistical AI. If you want to get a PhD, it's that. If you want to earn big bucks, that's really the field you have to work in. So I think cognitive architectures have suffered from that, but also there have been, I believe, two sorts of technical aspects that we've overcome, that I think have hampered them, and one of them is the knowledge representation itself of you know what we refer to. The knowledge graph, or knowledge representation, really has to be of super high performance, and we benchmarked our knowledge graph against graph databases, state of the art graph databases, and we are literally three orders of magnitude faster, a thousand times faster. So something that'll take one second of our system to compute and respond to, you know, involving a knowledge graph would take a thousand seconds and of course, that doesn't make it viable at all, you know, in an, in an interactive system. So, and related to that is that the various components that I mentioned, like deep passing understanding, short term memory, context reasoning, and all of those need to be deeply integrated with each other and with the knowledge graph. So we developed all of these components ourselves from the ground up, where it's traditional. Many of the traditional cognitive architectures have a very modular approach, but you know which is typically a good engineering approach. So you say, hey, we need a Stanford parser, that's a good parser, we'll use the Stanford parser. We need a knowledge graph, we'll use some knowledge graph. And we need a reasoning engine. We'll use somebody's reasoning engine. But the overall system really suffers from that because cognition really requires these systems to work together. When you hear a sentence, it might require short term memory to say, you know, say Bob's not coming along, okay, did we talk about Bob or something related to Bob recently? No long term memory: Do we know what Bob's? Do we know the reasoning? What kind of Bob would make sense for the context? We in? Well, it might be the family dog that's not coming along to the picnic next weekend, you know, you know. So you really need to have this deep integration of all of the different components for cognitive architecture, I think, to work effectively. We've actually spoken to, you know, r&d teams that know various large companies at Microsoft and Oracle and so where they've tried to integrate graph databases into their language systems and they just couldn't get them to work because of this performance and integration issue.

**Craig Smith:** 12:19

Hmm, that's interesting. Yeah, so that's IGO. You start with a model that you've pre-trained on a certain amount of data that gives it an understanding of basic relationships and language. Is that right? And then the company or person using it, as they use it, they add to the knowledge base or to the knowledge graph through their discussion or their questions and answers.

**Peter Voss:** 13:01

Yes, absolutely. So. I can probably best illustrate that, you know, with one of our customers because they are really logically three layers. Physically they all the same but logically they separate into three layers. You know, one of our big customers is 1-800-Flowers, a group of companies Harry and David and Popcorn Factory and so on and they wanted a hyper-personalized concierge type personal assistant, you know, for each of their 20 million customers potentially. So we have the core knowledge base of our system that basically knows about people and places and how to start a conversation and you know how to reason and basically to handle language, sort of a core competency. You know you could equate that, taking somebody sort of just starting college or something. Of course it doesn't have the broader knowledge base, but you know conversational knowledge. And then for each particular company we then add the next knowledge layer, the ontology of the knowledge layer required for that company. You know what are the products for that company, what are their business rules, and the integration to APIs, to their backend system, is also handled at that layer where you can dynamically get, you know the customers order status or their history or whatever which may be processed on another system. So the knowledge graph can basically integrate dynamically for company specific things. And then the third layer is per individual customer of what you learn in interacting with a customer. They might tell you, you know, I want to buy a present for my sister's anniversary, and then the system will remember that for that customer and potentially be able to use, utilize it. You know if there's a follow up question or next year or whatever the case may be. So it's basically the core knowledge and skills that are available to all of our users of our system, then the company specific knowledge, apis and capabilities, and then the third layer is for each individual user, what unique information you learn about them.

**Craig Smith:** 15:20

Yeah, how? How large can it be? The core model doesn't grow in size, but the knowledge graph grows in size and that's housed in a graph database. How large can that get? Is there? Is there a limit?

**Peter Voss:** 15:44

Yeah, so we, you know, currently all of the applications we do are text based, so they are not imaged. They know it. No images and you can really have a lifetime worth of conversation, pretty much. You know, in a few megabytes. I mean it's, it's really not not a problem. You know, we've, we currently. I mean, compared to large language modules, it's completely trivial. You know, whereas they're talking trillion parameters, we're talking a maximum of a few million. You know, we simply don't need more than that for these particular applications. Now for a more general AGI, of course that will become much bigger. It will be, you know, in the tens of millions, hundreds of millions range.

**Craig Smith:** 16:32

Yeah Well, how do you get them from, from this to AGI? I mean, what it's not simply a matter of scale you mentioned this right now is text base. You know we talked last time about Yanlacoon's JEPA architecture and he's right now. He's working in still images, but with an expectation to move on to video, where this architecture will build a model of the world based on its input, on structured data input, no labeling required or anything like that. And to me that you know, his argument against the pre-trained transformer architectures is that you know the knowledge that they're dealing with is what's contained in human knowledge, in language. But human, written human language, human language contains a lot of. You know lies and falsehoods, you know rhetorical devices and things, so it'll never understand the underlying reality, whereas his architecture is learning from in much the way as humans would form from what it's experiencing. And it sounds like your cognitive architecture in AGO has that potential. I mean, can you talk about it? Are there parallels with what the Kuhn is doing and what you're doing?

**Peter Voss:** 18:24

Yes, yes, absolutely. In fact, our early development, our early prototypes that we built, had multiple senses. In fact, we originally started off with a virtual critter, a virtual mouse, in a virtual environment that had whiskers and ears and could smell things, and you know virtual smell and all that. So I think it's very important. So I would definitely agree with that sentiment that you want, you want some grounding, you want grounding in senses and also you really do need general intelligence. You need, you do need some kind of a vision. It doesn't need to have the same, you know, sense of security as humans, necessarily, but you do need, you do need vision, I think, to have a sense of 2D, 3d world and, you know, also action movement. So, the current project that we're working on, that is directly aimed at going all the way to human level. Intelligence is a version of our technology that again has vision included in it, because I do believe that's, that's important. As you know, mouse, you can manipulate a mouse and keyboard and take in a desk whatever's shown on the desktop. So I would agree with that A for the, for the grounding and, as you say, for really the, the tie to reality of the grounding. Now, I'm not totally up to date with his approach and I'm not sure if you know if he's published everything. But typically, what everybody else is doing is basically bulk training, and I think one of the keys of intelligence is it has to absolutely be real time incremental, it has to be able to learn real time incremental and sort of much of it has to be one short learning. You know, for example, you can show a child a photograph, a single photograph, of an elephant for the first time. They've never seen one before and they'll be able to recognize, you know, pink elephants and upside down elephants and elephants in the wild and so on, and the systems really need to be capable of doing it. And all of the big data efforts right now are talking about plowing more data into training and it's all bulk training. You know. That's why they spend $100 million or more and have to use chips that cost $30,000, you know, per GPU chip or, you know, training chip. So I think that that's an important element of being able to learn incrementally in real time.

**Craig Smith:** 21:17

Yeah, and you mentioned earlier you know how condensed this data is if only contained in language. And I had heard and I'm trying to think whether it was you or someone who had mentioned to me that they had spoken to someone who's developing a watch, a device that would run 24 seven. Was it you that?

**Peter Voss:** 21:47

told me about this.

**Craig Smith:** 21:48

No it doesn't ring a bell, no Right. And so you would wear it throughout your life and it would continually pick up data and encode it into a model so that, over time, you would end up with a virtual twin. I've heard you talk about something similar with Igo. That's right.

**Peter Voss:** 22:19

Yeah, it's actually. I had some friends that maybe as long as 15 years ago were doing life logging, where they basically have a camera, you know, permanently attached, recording everything, and that seems to have fizzled, you know. I think one of the problems is also how do you ever get back to us, like people taking photographs of their lunch and dinner and all sorts of things you know, and they have tens of thousands of photographs. Do you ever look at them again, you know? So, yes, I think what we are doing, one of the aims is what we call a personal assistant and, as I mentioned, it would really be called a personal personal assistant because three different aspects of the word personal, you know, are really meaningful here. The first, personal, is that you own it. It's your property, you control it, it serves your purpose and not some mega corporations. You know it's not like Siri or Lexa, that clearly you know you get them for free, but you know they don't serve your purpose first and foremost. So that's the first thing, that you own it. And the second personal is it's hyper personalized to you. It gets to know your history, your likes, who you know, what type of things you buy, who your friends are and everything like that can do things for you. And the third, personal, is that the issue of privacy, that you decide what it shares with whom. So certain things that you may share with your spouse, other things with your coworkers and you know some other things you share with Amazon. So that's a personal assistant. So I don't see it so much as a twin recording everything you do, but more in an extension or an exocortex eventually that you know you can rely on so much to give you advice, to do things for you. It's like an expansion of your own mind, of your own cortex.

**Craig Smith:** 24:28

Yeah, and how do you do that? If you're this personal, personal personal assistant, how would you gather that data? Is it that you know someone would spend a certain amount of time every day you know through natural language just inputting data? Or would it be, as I describe, some device that captures data as you move around?

**Peter Voss:** 24:59

Well, it could be any number of ways. I mean, we are talking about something that really is at human level capability. So, if you wanted to, it could, you know, read your email, pass email, go through it and, you know, get to know things about you or anything else that you may have that it can find out. But to a large extent, it would be through interaction and things you ask it to do for you that it would pick up and ask for clarification. You know, do you always want to fly with this airline? You know, is that something you prefer? Or you know, what kind of hotels do you want to stay in? You know, whatever it might be, that it would learn through interaction. But, as I say, potentially you could read all of your Twitter or Facebook or email and gather more information from that.

**Craig Smith:** 25:49

Yeah, but again, is this a model that you're working on? And that would be that's? We're not talking about AGI there. We're talking about just a really useful tool.

**Peter Voss:** 26:07

No, I think we are talking about AGI there. I mean, there may be early versions of it which you know wouldn't qualify, but the work that we're doing right now is to expand our system in the level of, in a number of levels A in the sort of adaptability, meaning what it can learn. Obviously, the capacity of how much it can learn we are expanding, but that's more of an engineering challenge. And we're expanding the ability to reason. You know, for example, theory of mind, our current system can't really reason about what the other person is thinking. So you know that's something we need to expand the capability of its higher levels of abstract thinking. So it's really the cognitive abilities that need to be cranked up. And then we need to teach our system a lot more general knowledge. At the moment it's not a lot because you know for the applications we do it doesn't need to know about. You know baby showers and different sporting events and traveling the world and stuff like that simply doesn't need to know that. But once you get to a personal assistant, you really want to have that kind of general knowledge about knowledge, about the world, about things. So it's really scaling up the knowledge, scaling up the ability to learn autonomously. At the moment, there's still quite a bit of humans in the loop for when the system learns, so that, you know, will decrease as its cognitive abilities cranks up. But it's really the same architecture that we are using currently. It's just a matter of expanding it to be more general, more adaptive and more autonomous.

**Craig Smith:** 28:01

That's fascinating. I'm sorry, I'm just seeing my batteries low. I'll go for a few more minutes and then I'm going to have to run down and get my cord. But so it's the same architecture. How do you encode that knowledge? If you're? I mean there's a certain amount of label data at the very beginning of what is a verb, what is a noun and that sort of thing. But as you move into increasingly abstract space, how do you encode that knowledge?

**Peter Voss:** 28:58

Yeah, so it's true. The important thing is that the system needs to be able to learn this similar to the way a human does. So it will learn about different instances, for example, of whether it's examples of people or places or activities, and then the cognitive ability is able to do concept formation, is able to conceptualize this automatically and say well, these things fit into this category and other things fit into a different category. And part of that is where we will give it a curriculum to encourage that kind of learning, similar to the way you would a human, to basically feed it information, give it examples, give it exercises to be able to build up this knowledge structure layer by layer. So the more fundamental things are really solidly embedded, like time and space, for example, and comparisons of bigger and smaller and all of the different kinds of relationships, and then, as it gets better at being able to learn autonomously, then it's basically a matter of hitting external data. Sources, like Wikipedia is an obvious one, and we've already done quite a bit of work in that area. But this is also where large language models can actually help us, where we can extract information from them, as long as you have a system that's smart enough to be able to double check the information and see if it makes sense, if it makes sense to integrate it.

**Craig Smith:** 30:42

Yeah, would you be able to avoid the bias that exists in human databases?

**Peter Voss:** 30:55

on the internet generally. Yes, I mean ultimately biases like that can. The best remedy for biases, handling biases like it, is ultimately rationality that you say is really a relevant factor? Is race a relevant factor, or whatever it might be? Is the background a relevant factor to achieve a particular objective? And humans aren't that good at it. But basically, the more rational you are, the more rationality you apply to it, the better you'll be at basically eliminating bias in your decision-making.

**Craig Smith:** 31:37

Yeah, and that idea of combining large language models with this cognitive architecture approach. I mean, do you think AGI? Not everybody, but there are several different initiatives to try and get there, and you wrote a big chunk of the original artificial general intelligence book with a number of others, I mean Ben Gertzel and Shane Legge. They've gone off in their own directions working on AGI. Do you think ultimately all of these different threads, or the most promising, will come back and intertwine, or do you think there is only one path to AGI and one of you guys will find it, or do you think that there are several paths and there will be different flavors of general intelligence?

**Peter Voss:** 32:42

Yes, so I'm quite certain that it has to be something along the lines of cognitive AI, though there are different ways of achieving it. For example, one could start with a much more grounded, embedded approach that the system learns. Through that it certainly has its advantages, but also has big disadvantages of having to deal with the robotic aspects of it. So I think there are different ways of getting there. But even Sam Altman and Demis Asabi of DeepMind have said quite clearly that large language models are not the way to AGI. They're a dead end, but they don't really have an alternative. It's more at the moment, well, let's build bigger systems, let's kind of see what happens. But on the one hand, they recognize that it's a dead end, but on the other hand, they don't seem to really know how to solve the problem. And the biggest impediment by far well, there are several, but I would say the biggest one is that they are pre-trained GPT. They're generative, they make up stuff, which is not a good thing, rather than reason about things and have a solid knowledge base. They are pre-trained, which means they cannot really learn once the model is built. So to build a new model you're talking about $100 million and I don't know how many weeks and additional training doesn't really work. Well, I mean, you have catastrophic forgetting. The buffer is not integrated, the context buffer is not integrated. So the fact that they're pre-trained, they're bulk trained, it's really a killer. Imagine hiring somebody as an assistant and they come to you and they can do Excel and they can do QuickBooks and they know a bunch of things. Okay, they sometimes make some really bad mistakes, but let's ignore that for now. But so you tell them we're just taking on some new products, I've got a new partnership here. We're shutting down one of our branches and a few other things happen in our business. Next day they come in. They don't remember a thing about that. That's not intelligence. So that problem really cannot be addressed by building bigger models. And the GPT, the T, the transformer, really locks them into this. But the problem is, transformers have been so incredibly successful in so many areas. I mean, what GPT can do is amazing. It really is amazing. So they're kind of locked into their own success and they really need to somehow tear themselves away. And the other issue why I say cognitive AI is really the way to go and why I think the current leaders in the field with statistical AI are unlikely to actually achieve AGI is because their backgrounds are inherently statistics. They're mathematicians, statisticians, logicians, and that's the way they see intelligence and basically see hey, we've got all this data, we've got this computing power, we've got computers that can do certain things. What can we do with that? How can we use our human intelligence to make these things do amazing things like become world chess champions or go champions or do protein folding or whatever? But these are all really narrow AIs that rely on the ongoing external intelligence of humans. To make them work, you really need to start with cognition, with cognitive understanding, intelligence, and say we absolutely need these. These are the following requirements to build an intelligent system, and if your system doesn't have that, it's not going to ultimately be intelligent.

**Craig Smith:** 36:55

Do you follow how Shane Legg and Ben Gertzeler are pursuing AGI?

**Peter Voss:** 37:04

Yes, somewhat. I was just at the AGI conference, the annual AGI conference in Stockholm, and caught up with Ben. His current focus seems to be more on a sort of society of mind, not in the Marvin Minsky sense, or at least not in my opinion, but that a lot of narrow AIs, society of narrow AIs, will give you AGI, which I disagree with. I don't believe that's the right path, but that's the path he seems to be pursuing. Shane Legg has actually not published much over the last 10 years that I'm aware of, but he recently did give an interview and published something that seems to be relying very much on reinforcement learning, which to me is kind of the opposite of learning with one-shot learning. Basically, and I mean reinforcement learning has its place, but I think it's a relatively minor one, and to believe that you can solve human level intelligence with reinforcement learning seems like a stretch to me.

**Craig Smith:** 38:22

Yeah, that's interesting, and Ben's ideas that these narrow AIs would talk to each other or there would be an orchestration layer that would sort of direct queries to whichever one had the domain expertise.

**Peter Voss:** 38:44

Right, I mean that orchestration layer, unless that is an AGI, in which case that's what you should be focusing on. So you want your AGI to be a tool user, like humans are. That's a big power that we have. So you want an AGI to be able. If I ask my AGI to write a poem in the style of Jimi Hendrix or whatever, it should probably go to GPT or the large language model and give me a few suggestions from there. Or if I wanted to play a game of chess, I could probably just use a chess program for that. I wouldn't expect it. So an AGI needs to be a tool user inherently and be good at that. So that orchestrator. The problem with narrow AI is that it's really not AI, because in the original sense of AI and that's why we coined the term artificial general intelligence to recapture the original dream of AI to have thinking machines. Narrow AI inherently has external intelligence. It uses the program as intelligence and the perfect example of that is deep blue, for example, the world chess champion. It's not that the system itself has intelligence, no, it's the ingenuity of the engineers on how to use algorithms that could do the branching and whatever logic they used and heuristics they used and encoded to play a mean game of chess. So the narrow AI is really not intelligent and putting a whole bunch of narrow eyes together, again, they cannot learn, they cannot conceptualize, they don't have general intelligence, they don't have intelligence. So having a whole bunch of specialized intelligence AIs could be good as tools, but not as solving the problem of AGI.

**Craig Smith:** 40:40

Yeah, your system. So you're talking about using the same architecture but growing it. And is that to reach AGI? Is that growth through adding modalities or the ability to input data from different modalities? What's the bottleneck in that?

**Peter Voss:** 41:23

Yeah, there are a number of areas that we need to add. One of them is that language capability. We've really put that into the system through language rules, grammar rules and so on. With the new version that we're working on right now, language will actually be acquired by interacting with the environment. We can do that because we have a visual sense now as well. We have visual sense input and mouse movement and so on. The system can actually have these grounded concepts and can learn language from the ground up. That's one of the big differences, which basically makes the system more adaptive. I've actually, over the years I've always said, when customers have asked us, when can we have it in another language other than English, I said, well, ideally, when I go smart enough to learn the language by itself because we don't want to have to encode different language rules for different languages. Large language models, of course, solve that problem by brute force, with statistics. So basically making the system more adaptive, that it can learn more things and can learn them better than it can now. At the moment there's still quite a bit of humans in the loop for the system to learn anything, to learn new skills or to learn a different concept. There's too much of a human in the loop for many things. It can learn by itself quite a bit, but not as much as it should. That's also a new skill. Again, our system can learn skills autonomously or through language instruction, but that needs to become more powerful. We also simply need to expand the knowledge base that we spoke about and that will involve scaling, going from a few million concepts to tens or hundreds of millions, but that's pretty much a straightforward engineering issue. So it's really the generality of the system, by giving it a broader knowledge base, the adaptability and autonomy that it can learn more autonomously, that it can learn with less humans in the loop. So that's where our focus is, but it's really all incremental type of work, of work we've done either in the past or is already in our current system.

**Craig Smith:** 43:53

Yeah, the human in the loop. What are the different functions that humans are playing?

**Peter Voss:** 44:02

Yeah, so in our current commercial system it's basically we had humans create all of the language rules required, so that was a human in the loop In terms of the specific skills that it needed. Contrasting that with what an AGI needs to do and what the new generation, the next generation of our technology will do, it really should be able to go to a customer and say what are the specifications, give me your training manuals, I need your API documentation and then configure itself basically. So that's ultimately the direction we need to be moving in, and to have less and less of a human in the loop to basically feed it that information. The training process I spoke about is basically building up a curriculum to give it that core knowledge. Think of it like a child. You get to a certain age where it can learn largely by itself and then it'll need less and less of human assistance or it knows who to ask for help when it's stuck or needs clarification. That's where metacognition comes in, that the system knows what it knows or knows what it doesn't know, and can monitor its own thought process and level of certainty and confusion. So that's again an important element missing in large language models is metacognition. They don't have any metacognition, and that's a part of our system that we also are developing further.

**Craig Smith:** 45:48

Yeah, your system is certainly much more compute efficient than a large language model. What a typical system for Fortune 500 Enterprise. What is it running on? Gpus, cpus, yeah.

**Peter Voss:** 46:14

Very, very little processing power. So you can run our system as an agent. You can run on a five-year-old laptop or something. So typically we run them on machines in the cloud clusters and we'll run one conversation per CPU and this is a server CPU. It's not particularly high performance, so they really need a trivial amount of processing power to operate. And on the training side we'll retrain our whole system several times a day for our regression tests to develop new things or test new features or whatever. And I mean literally the training cost is pennies. It's completely trivial. Now of course, once we scale our system up to millions of concepts, that'll be a bit more. But hey, if it costs $100 to train the system, it's still a lot more than $100 million.

**Craig Smith:** 47:18

Yeah, that's right. How far do you think you are from scaling this up? Yes to a degree that's approaching human level intelligence.

**Peter Voss:** 47:33

Yes, I think there'll be a time when it becomes really, really obvious to pretty much anybody who works with a system, saying, wow, this is the real thing. It really can learn by itself, you know, and it still needs a bit more. How long will it take us to get there? Well, I usually answer that question. It's not a matter of time, it's a matter of money, because it really is. I mean, a lot of the research we did many, many years ago and it's really more a question of having the people to do the development that needs to be done. But I'm actually quite confident, if we didn't have a constraint in terms of people, that we can do this in three years, in that sort of timeframe. But you know, we really need, you know, 100 or so people on the project for the various things that need to be improved and tuned and scaled, and that's what we're in the process right now of trying to raise funds for that, to increase our team size, to accelerate the development, because with our current 30-person team we have about a third of the people working on the AGI project but two-thirds are working on the commercial side.

**Craig Smith:** 48:53

And are you working with the government at all? I mean, there's been such a push and so much money made available. Are any of those programs helping fund you?

**Peter Voss:** 49:07

No, I mean, I've had really bad experiences with that and other people I've spoken to about that. As usual, government money just tends to end up in the wrong hands. We've had people, I mean, applied for these programs and then they say this is for innovation. And then they ask you well, how many projects have you done before? Okay, if it's supposed to be innovation, you have like a factory of innovation or something and do you have a department that knows how to, has all the right connections and so, and then you try and work with consultants and say, okay, pay me $30,000 a month and I'll put you in front of the right senator and yeah no, I mean, I don't have any inherent objection but I've not had a good experience with trying to play that game.

**Craig Smith:** 50:04

Yeah, yeah, that's unfortunate. So you're raising money now. If you get the money, we're within several years of achieving something close to human level intelligence. Yeah, I believe. And then in the meantime, yeah, who are the customers for IGO as it exists today?

**Peter Voss:** 50:32

Yes, the simplest way to categorize it is really any company that has a call center of 100 people or more. I mean, that's sort of where economics becomes really interesting, because we could probably replace 50% of the people. So if they're growing, they don't need to build bigger call centers or get additional people and call centers are really suffering. I mean we hear that everywhere. They're really struggling to keep staff, to train them, to get them, to train them and keep them. It's a tremendous turnover and the quality is all over the place usually. So that sort of at the moment, the most obvious customer profile. But we're also talking to universities as a student assistant, when you first go to university, for example, find your way around, where do I get meals and books and curriculum and all the stuff to help you with studies and so on. So we're very excited about being able to do that. We're helping to get a project like that off the ground soon. Or diabetes coach for somebody who has diabetes and they want something to help them manage it. Internal HR support for a large company or IT support would also be an application. And another area is what we call a co-pilot. So for complex software where people don't know how to use all of the features in the software, it's awkward to use it Like Salesforce or SAP. It's really difficult to know how to use all the features or they may be hard to use If you can have an IGO as a front end to that and just talk to it and tell it what you want done. It can basically then either navigate you to the place or, in some cases, even just execute whatever you want to do. So we are really industry agnostic and, to a large extent, application agnostic in terms of anywhere where conversational AI can help.

**Craig Smith:** 52:39

Yeah, and how do you price your products?

**Peter Voss:** 52:45

I mean we're fairly flexible, but typically it's per conversation. Now we may have a different rate for fully contained conversation if it's a call center application. So if it's totally successful, maybe a premium for that. But we work with customers that can potentially also be perceived, but usually performance based. We like that model because it's a win-win situation.

**Craig Smith:** 53:15

Yeah, yeah, that's fascinating, it is. Do you find the competitive landscape these days difficult because the GPT chatbots have kind of taken over?

**Peter Voss:** 53:34

Yes, it's certainly a short-term challenge. By the way, one thing I forgot to mention is we deploy this behind the customers firewall, which they love. So we do not assess service, we provide our technology and it just runs on their cloud service or on whatever hardware they have. We just deployed it as a Kubernetes service. So, yes, when chat GPT-4 is, in particular, 3.5 chat GPT hit the news, a lot of our prospects that we had lined up that were ready to go with us management said, wow, isn't this gonna do everything that you guys do? And we've got to investigate that. And some of them have started coming back to us and said, yep, we've investigated it. No, it can't do anything like that. It can be good for FAQs, it can help with search and if you can train it with all of the documents you have in your company, it can certainly help you answer questions about them, as long as you have a human in the loop to verify that this actually makes sense. But as far as having customer conversations that you need to be able to rely on, then that need to be deeply integrated with your backend services and have business rules and things that are reliable enough that your legal system will sign off, your marketing department will sign off and your customer experience team will sign off on it. Large language models simply don't meet that requirement, so we're starting to see those prospects come back to us after they've investigated large language models.

**Craig Smith:** 55:17

Yeah, and just a question on using graph databases. There's a lot of people using vector databases with large language models to ground the large language model in a knowledge base. Do you use vector databases at all and could you apply or integrate that kind of a solution with IGO, where you have a large language model and either it queries the knowledge graph or a vector database to compile its answers?

**Peter Voss:** 56:04

Yeah, no, you really can't integrate the two technologies for the reasons I mentioned earlier. You really need deep integration. I mean every new thing that you hear. You know if you have a product, new product or something changes in your life, that needs to be an integral part. It needs to immediately update your model and it may have a significant impact on how you respond to future queries. You know you have another child or you know whatever you move to a different town or whatever. It has to immediately update everything. And as far as using graph databases, we actually, in the white paper just recently published on how to get to AGI, must direct the route to AGI. We did some benchmarking and the one benchmark we did was against the state of the art graph database and our system was a thousand times faster. And you know that's what I mentioned earlier and you really can't have that kind of performance penalty. So you know, if it's again if the system can use it as a tool, as an external service, by an API, to do something, absolutely, I mean same as the system can query an external database. But to be an integral part of the brain, that just doesn't work.

**Craig Smith:** 57:39

Yeah, okay. Well, we're coming up to an hour. I just want to ask on the personal, personal, personal, I go how far off is that? I mean, I can imagine having everybody having this personal assistant that learns as you learn over time.

**Peter Voss:** 58:08

Yeah, I know people get excited about it, always ask us when can I have one, can I be a beta tester, and so on Again. Unfortunately, the answer really depends on dollars rather than, you know, months or years. I mean, certainly it'll be at least a year before our tech, our new generation, our next generation of technology, you know, is ready for prime time. But I don't expect it to be much longer than that, provided we get the funding lined up pretty soon.

**Craig Smith:** 58:44

Yeah, that's fascinating. I'll look forward to that. Is there anything that I didn't touch on, that you wanted to say or that we talked about last time? That I don't remember?

**Peter Voss:** 58:59

Yeah, I think the thing that really surprises me is that I published a white paper two months ago so why don't we have AGI yet? And the thing that surprises me is that there aren't more people in the world who really take that question seriously. You know, with the billions and billions of dollars being thrown at AI and people just blindly go ahead. You know sheep following FOMO setting in. Well, we don't know which of these companies are gonna succeed. Nobody's got a moat. Maybe the best moat is money. You know, if we give them 10 billion, if we give them 100 billion, then maybe they've got a moat that other people can't compete with them. And you know, on the one hand, they all freely admit this technology is not gonna get us to AGI. And why aren't you asking the question? Well, why and what will it take? And I've tried to get people to just brainstorm, you know, even sharing ideas of what we've learned, what works, what doesn't learn. You know quite openly in terms of sharing ideas to make this happen, because, yes, we'd like to have a trillion dollar company with AGI and we expect to, but to me it's more important to for humanity to have AGI because I think it will so much improve the quality of life of everyone in the world potentially. And so you know that really baffles me that we don't have more people really asking the question of what do we know and what you know, what doesn't make sense, and should we be doing more things that don't make sense? What are the things that maybe do make sense and, you know, fund them. It's such a monoculture basically at the moment, so that's a bit baffling and I mean, apart from that, it's just obviously I'd love to hear from people who want to use our technology commercially and, more importantly, we are looking for partners, people who want to work with us on developing AGI, collaborate in some way, and, of course, also for people who want to fund that and maybe make that part of their legacy.

**Craig Smith:** 1:01:23

This episode is sponsored by DataStacks, the real-time AI company. With DataStacks, any enterprise can mobilize real-time data and quickly build smart, high-growth applications at unlimited scale on any cloud. Companies building real-time generative AI apps can leverage the DataStacks vector search capabilities to build LLMs, AI assistance and more. Sign up now at the link in our show notes. That's it for this week's episode. I want to thank Peter for his time. If you want to read a transcript of this conversation, you can find one on our website. I am AI. That's E-Y-E-O-NAI. In the meantime, remember the singularity may not be near, but AI is changing your world, so pay attention.