AI and You

Transcript

Archive Interview: Ben Goertzel

Episode 106

First Aired: Monday, June 27, 2022

Hello, and welcome to episode 106! We talk a lot about artificial general intelligence on this show, so much that we often say AGI and forget to tell you what it stands for, which is why a couple of weeks ago I devoted a whole episode to explaining ANI, AGI, ASI. Well, my guest today is all over AGI. He founded the AGI Society, he has a book called Artificial General Intelligence, and he was one of the three people popularizing the term AGI to begin with. Ben Goertzel joins me on the show today. You may know him from his book *Ten Years to the Singularity if We Really, Really, Try*, or as the founder of the SingularityNET Foundation, or as the Chief Scientist of Hanson Robotics, the company that created Sophia the Robot, who you may have seen on various talk shows.

We will be talking about the Singularity, and there will be some explanation of that when we get to it. Just so you don't feel left out until then if you've not heard of that, it's a term that came originally from mathematics, for a place where a function wasn't continuous, where it does something weird, then it was absorbed into cosmology, for what's going on in the center of a black hole where the laws of physics go crazy – look up the term 'naked singularity' for more about that – and then the science fiction author Vernor Vinge took it on in 1993 and talked about a Technological Singularity, where the idea was that at some point in the future the rate of technological progress would be so great, changes would be happening so fast, that to human eyes it would appear infinite, that everything was happening at once, and he thought that should have the same name that we give to places in the universe where God is dividing by zero. A few years later Ray Kurzweil took that ball and really developed it, writing a book called *The Singularity is Near,* starting the Singularity University, and writing another book called *The Singularity is Nearer*. He projected that the singularity would arrive in 2045. Vinge actually thought it would be between 2005 and 2030.

But we'll start out talking about AGI. Most of the guests on this show, we're talking about AGI tangentially, but with Ben we get to go at it head on. There are a lot of questions about AGI, and I tried to get through as many as I could. Let's get right into the interview.

Hey, Ben it's great to have you on the show.

Great to be here.

Can you tell us something about how did you get into this field of – well, I'll let you describe it, but artificial general intelligence and your role in that.

I've been interested in AI and in what we would now call AGI since the early 1970s, when I was a kid and saw *Star Trek*, the original *Star Trek* on TV and then got into reading various science fiction novels where there are super human robots and super computers roaming around and at that point, I didn't realize this was a serious research pursuit rather than the subject of fiction only and in the late 70s, or around 1980, I guess it was, I read Douglas Hofstadter's book *Godel*

Escher, Bach, which gave his own views on creating general intelligence and also sort of introduced me to the idea there's a whole community of AI researchers actually working on this stuff, I was maybe 13 or 14 at the time. And from that point on, trying to build thinking machines seemed like a pretty interesting thing to do with one's life and I discovered a book around the same guy called the *Prometheus Project*. I think I discovered that book around 76 77. This was written by a Princeton physicist, named John Feinberg, the book was actually out in the late 60s, I discovered it in the late 70s. He said, he felt within the next few decades, we were going to create machines that think smarter than people, we were going to master nanotechnology and be able to build with everyone out of matter, and we would master aging and death and we'd be able to live forever. And then the question would then be, what do we do with this capability? Do we use it for a consciousness expansion or for just rampant trivial consumerism and materialism? And he proposed, the UN should put to a vote of all humans on Earth, whether the singularity should be developed or advanced consciousness or toward materialism and consumerism? and these were interesting thoughts to me when I was when I was in middle school in the 70s, right that not a lot of people were thinking about these things in but as I proceeded, got my PhD in math and went through various academic roles in mathematics, computer science, cognitive science, I was always inspired by these bigger visions. Like, if you can make a machine that thinks much better than human beings, then as I. J. Good said, the year before I was born in 1965, like the first truly intelligent machine will be the last invention humanity has to make and so, whatever else you want to do, if you make a true AGI first, I mean, then that will probably make that other thing easier, as long as that AGI is configured and educated to want to help you rather than rather than do something else. So, I really came from this sort of science fictional/futures perspective. But then, of course, one quickly realizes it's a deep technical pursuit. So, I went through learning all of math, computer science, neuroscience, cognitive science and linguistics, philosophy, engineering and all the different disciplines that you have to bring together to try to build AGI. Now I introduced the term AGI to the world in August 2004, or 2005, it's heartening to me now. I mean, not so much that that term is in currency, because terms don't matter too much. But it's exciting that the pursuit of building real thinking machines is something that's considered legitimate now and big companies are working on it. You can publish papers on it and there's teams around the world working toward this goal alongside my own team working toward this goal, although, of course, nobody is there yet and while I have my own strong research intuitions about what approaches are likely to be successful, no one can say for sure until the end goal is reached.

All right, how do you think we're doing on the timeline of that compared to perhaps what you thought when you were starting to get into this and how long you thought it might take then; is it taking longer than you thought? Is it going faster? What's that velocity like?

Well, when I was a young kid, I figured it would take tens of thousands of years so I want to build a spaceship go away from the earth and near lightspeed and come back a million years later when I was already embedded superAGI/nanotech. Then when I read Feinberg's book, the notion of the singularity in different language, but it was the same notion as the singularity came to my attention and I realized, there's exponential progress and our intuitions aren't well attuned for it and you can do twice as much in one year as you did in the previous year and it takes a

while for the human brain to acclimate itself to a regime of exponential progress because our minds are tuned by evolution for periods of linear progress of steady state society, as was the case, before the advent of civilization. Australian Aboriginal society has been around the same for 60,000 years and that's what we got accustomed to in an evolutionary sense. And the reason I bring this up is, I think we're actually well on track for Ray Kurzweil's, target of human level AGI by 2029. And I think Ray did a fairly careful job of plotting out Moore's Law and progress of brain scanning, progress in complexity of software and he came up with the time around 2029 is when he thought human level AGI would come about, and I think, plus or minus a few years one way or the other. That still seems about how things are going now and I think whether it seems that way to an individual depends in part on how well they've sort of internalized the rhythm of exponential progress, because it's certainly true that we don't have an AGI now, we don't have a baby AGI now. I mean, we have quite impressive AI systems that are quite cool and do some amazingly impressive and very useful things. We don't have a system that can generalize robustly beyond its experience in programming. We don't have a system that can make creative, imaginative leaps beyond its experience, like a person can, so we don't have it, whatever this or that marketing department made to me telling me right? On the other hand, the nature of exponential progress is that you're doing more each year than was happening the previous year and I think looking at things in that light, I think 2029 plus or minus a few years is quite viable. I mean, I'm hoping to get there a few years ahead of that and I think that getting a human level AGI by 2026 or 2027 is not beyond the pale. I think if our OpenCog Hyperon project goes maximally nothing screws up and funding lines up and our AGI chip hardware project loads up and isn't stalled. Like if everything comes together well, we can beat Ray's deadline by a couple of years. Now, if there's another mega pandemic and more, World War III risks and recessions and things slow down a bit, certainly reality can intervene and you miss Ray's target, by a few years. I mean, it's not an exact science, even Ray understands that. But I think we're on track for that magnitude and Ray believes that you have human level AGI in 2029 and still believes, and then the singularity in 2045. This is where I disagree with him, I don't think there's going to be a 16-year gap. I think once you get to human level AGI, it can be just a few years to get to superhuman AGI because you got a machine that can rewrite its own code and rebuild its own hardware. Now I wrote a book, which I think you you've taken a look at, which was titled *Ten Years to the Singularity if we Really, Really Try* and I mean, the we meant humanity, not just me and my friends and family and I think humanity has not really tried, right, and what Ray's estimates based on curve fitting, baked in the level of effort that humanity really has to put into frontier science and technology, which is a tiny fraction of its effort and there's no doubt that with massive concentration of resources on building AGI we can do much faster than we could with the current concentration of resources. I don't think we need humanity to focus its attention in a single-minded way on AGI to get there within like a decade or so from now, which is amazing if you take it seriously, and not many people do.

Well, I want to break down the term *AGI* because it's getting so much use these days. It's become a word that needs some differentiation. I asked Stuart Russell on the podcast whether we were hardware-bound or software-bound on AGI and he said software-bound; that we could do it on today's hardware if we had the right software. That's a guess, of course. But it

means that we need some kind of stages in here, we need some kind of markers and milestones because otherwise, people seem to treat AGI as an all or nothing kind of thing. Like I'll know it when I see it. But how do we know when we're 50% of the way there? This is one of the big problems for me in interpreting the progress.

I don't think that's a big problem. I don't think it matters. I mean, I understand why it would be convenient, but I'm not sure that's important.

It is if you're investing in it, you want some idea of when it's going to happen.

Yeah, it would be convenient to have for government grants and for various sorts of investors, but then that not everything that will be convenient can be delivered by reality. So, I agree with Stuart Russell, first of all, that getting to human level AGI now is essentially a software problem, I think, however, a good analogy is deep neural networks, which everything done today with deep neural networks could have been done without GPUs. On the other hand, having GPUs and multi-GPU servers not only makes it faster to run your computer vision and natural language model, it allows research to happen faster because researchers can try various successful and unsuccessful ideas more quickly, to then find things that work more quickly. So, certainly we have the hardware to deploy the needed software. But to create human level AGI I'm almost sure of that. On the other hand, having better hardware could let us do the research faster and fine tune our software faster and I mean, I'm basically a software guy and the math and cognitive science guy. I'm working with one of my companies True AGI in partnership with another company Simuli run by an AGI researcher called Rachel St. Clair, you should interview some time actually. But we're working on an AGI chip, which we hope can do for AGI what GPUs did for deep neural deep neural networks, which is not to make possible wouldn't have been possible otherwise, but just to make cheaper and faster what would've been more expensive and slower otherwise and thus accelerate research, and I mean, I'm sure, Stuart Russell understands that. But again, getting to your other question about defining AGI and quantifying AGI and quantifying the path toward AGI; I want to go back to a workshop I did I - think this was 2012, I'm not sure the year I think it was 2012, which is a great Rush album too, by the way, but I think that was 2012 at University of Tennessee and Knoxville and I think there was a workshop on Roadmap to AGI with about a dozen AGI researchers from academia, mostly academia, a couple people from industry, and we set out to figure out exactly what you just alluded to. The purpose of that workshop was to figure out how would we define incremental progress toward AGI in a way that you could measure your progress toward AGI. And our goal at that time was to help work with DARPA and IARPA, other US grant funding agencies to define grant funding programs for AGI research. Because for US government grants or any other government grants, having metrics to evaluate incremental progress is very important. I thought about metrics a lot more between 2003 and 2011, I lived in Washington, DC, and I was primarily working on US government agency focused AI projects to make a living while working on AGI research in my spare time, because government was the main source of AI funding in that in that period, right? So, in that domain, you think a lot about metrics and milestones. You know what came out of that workshop, on the other hand, was a paper in AI Magazine called *Mapping the Landscape of AGI*. So, we started out with a roadmap to AGI and we ended up with mapping the landscape of AGI and the reason

for that shift of metaphor was as follows. What we found is that every AGI researcher at that meeting, had a different idea of what's the path to AGI, and the metrics and milestones that they would accept, dependent on their path. So, there was one guy there, Josh Hall, J. Storrs Hall, who was robotics-oriented and he basically figured, once you had a robot that could go into your kitchen and figure out how to make coffee in a random American kitchen then you're 90% of the way there to AGI and if you don't have that, you're just fooling yourself, and you're not doing anything like what the human's doing. So, he just figured, like, humans evolved out of creatures that perceive in a 3-D world that move and manipulate stuff and then our abstract thinking and language are all built on top of that, right? So, his idea was, your incremental progress milestone should be basically how robustly and flexibly can you move around, see stuff and manipulate stuff and navigate in unpredictable environments, and he wanted an incremental milestone of that nature, right, and that - I don't say that's a bad path, I just say it's not the only path. So, there was another guy there, whose focus was on making AI systems read like little kids, and he was doing logic based reasoning for natural language processing and he was making AI systems do reading comprehension exams, from elementary school, middle school, and so forth and again, there's a fairly good argument that you can see an incremental path, right, you start by reading early readers, you go through, you can read, third, fourth, fifth, sixth grade, you can read high school level, eventually you're getting to human adult level general intelligence. But his idea was if your system can't even read it, the kindergarten level is basically an animal, he's not going to consider that too much progress. So, his incremental milestones were more language comprehension based, right. Someone else was looking at embodied vision processing and trying to extract the semantics from movies, and then looking at video games and using the semantics you extracted from video games and he's just: sensor and actuator problems are just hardware problems, right? That's like making the wheels on the car. If you can perceive in a virtual world and act in a game world, that's just as good as going into physical world. So, what we need is progress on integrated perception, action, language and social interaction, whether it's in a game world, or in physical world doesn't matter. So, his incremental milestones were about the complexity of the situation in a virtual world or video game world, the complexity of the situations that you can achieve goals in and he could articulate milestones based on that. And so then getting these researchers to accept each other's milestones was not possible, because what you saw was the path to AGI that someone thought was most likely is totally wrapped up with their own research and what kind of system they're building, which is totally wrapped up with what incremental milestones you would accept. So, then we just put forth the World One metaphor of, we're trying to get to the top of Everest, we all sort of agree on what's at the top of Everest, like we all sort of agreed, if you could make a robot that could go go to MIT and graduate, going through the same exact procedures as the human does, it's going into class, it's sitting there in the desk, it's listening to the teacher, it's understanding what the assignment is, it's doing the assignment, it's doing teamwork with other students when needed, right. It's putting the pencil on the paper when needed, or booting up the software when needed and typing. I mean, it's, it's also solving the math and writing the code. We all sort of agreed if you could make a robot college student, then that's human level AGI and we also agreed that if you got to human level AGI by some other method, the human level AGI could probably learn to be a robot college student before too long. So, we had some agreement on what was at the peak of the

mountain and I mean, we also I would agree, the peak of that mountain of human level AGI is the foothill of the mountains of superhuman super AGI right. But what we didn't agree on is do you take the south face up, the north face up the east face up, the west face up to the top of the mountain of human level AGI right. And we sort of even sort of agreed that all the paths could get us there. But some people just thought, getting there in a way that doesn't involve a robot is dumb, but there's going to be a hundred times slower. Whereas other people thought involving a robot is done and will make you a hundred times slower. So, totally didn't agree on which route. But when you get into metrics, I mean metrics do depend a lot on your idea about what is the path for getting there. And while this was 2012, we see the same thing right now. So, I mean, you have Google DeepMind, you have some researchers now saying, well, we're basically there, right? We just need to make our models bigger and we need to make our models faster, and train them and more and more data, and then we're there. And that's really because according to their approach, they're buying into a set of metrics that happens to rate their own approach as very far along those set of metrics. Like they made a system that one neural model, they trained on 600 different tasks, and they can figure out which task is being given a new model, you can see how someone would think that's fairly long along the path. I could also see how someone would say, there's no creativity, there's no imagination, there's no agency that right, so then you've made zero progress.

So, I think that this is we're talking about a definitional problem here that is exacerbated by the college student example. Because that's an anthropomorphic example, and in humans, general intelligence also comes along with a body, emotions, creativity, free will, and consciousness; and how many of those things are included in your definition of AGI, would be there as well as the intelligence?

Yeah, so if you want to define general intelligence, in a really rigorous formal way, there's a literature on that. Marcus Hutter had a book, universal AI and Shane Legg who was my former employee, who later went on to co-found Google DeepMind. He wrote his PhD thesis on this with Marcus at IDSIA in Lugano, Switzerland, and they gave a formal mathematical definition of what it means to be a general intelligence now, I later wrote some papers trying to improve on their definition and one can debate it but the nature of Shane and Marcus's definition is, A system is generally intelligent to the extent it can achieve an arbitrary computable reward function in an arbitrary environment. Now, I don't think that's a good definition of general intelligence, in the end. But I think it has something to teach us. One thing it has to teach us is that humans are really, really, really stupid by that definition, right? Like we were very bad at optimizing computable arbitrary computable reward functions in arbitrary computable environments. We can't even run a maze in 750 dimensions, right? I mean, so if you give us seven-dimensional Go on the 489 x 489 x 489 x 489 … board we're terrible at playing it. So, these are pretty simple reward functions to define. So, I think that definition itself is too stiff and rigid and I'm a big fan of Weaver, aka David Weinberg's theory of open-ended intelligence, which models general intelligence is more being about just a system individuating and preserving itself as an agent and then transforming itself and growing in a radical way. So, I think this whole reinforcement learning paradigm with reward optimization is a bit limiting, but it does teach you how dumb humans are in the grand scheme of things. So, when you're talking

about human level AGI, you're talking about a level of general intelligence which is very very low compared to mathematically-definable AGI and certainly very very low compared to the maximum level of general intelligence you could have in our physical universe, even setting aside abstract mathematics. Once you get to femto computers and atto computers and futuristic compute fabric. So, thinking about AGI as human level AGI, it is a very messy sort of fuzzy, nasty thing to define, because we're talking about something that's generally intelligent in the particular ways that humans have evolved to be and there's a lot of a lot of arbitrariness to that, right. Like we're really good smart. Some things were really dumb and other things were insanely dumb and doing long division in our heads, except for a few autistic people. I mean, so much that some aliens who are good at that might decide we're not intelligent at all, because we can't even do; we take hours to divide 700-digit numbers by each other, like what why are we so dumb, right? We're not good at three dimensional rotating objects in our head to figure out what tunnels they could fit through. There's stuff in the way seems very basic that we're very bad at but we're not so terrible at figuring out when someone's trying to bullshit us, right? We're pretty good at walking across the street in New York; we're really good at making up funny things that other people will like and composing music. So the mix of things that were intelligent or stupid that I don't think is fundamental or easy to rigorously pin down. I think it's just what we have what we evolved to be good at, which is how you end up with something like the robot college student test, because that's just like, this is being like a smart human and MIT was designed to measure in a way it's designed to measure Are you a smart human or not? Right, and it's a very non bullshitty way to do it. Because over many centuries, people have tried to figure out ways to bullshit their way through college. So, college has been designed to make to make that a little more challenging.

Do you think it'll take embodiment to get there; they will have to be able to walk into the classroom in addition to learn the course?

No, I don't think he would have to. But I think that once you get to human level AGI, that human level AGI would be able to figure out how to walk into the classroom without that much extra work. So, I think if you have...

Chicken and egg, would it need to know how to get into the classroom to become AGI?

I don't think that's a need, I think that's a convenience. So, I think that having a robot body moving around in everyday human world can be extremely helpful to an AGI system that's trying to gain human level intelligence because there's a lot about what humans do that can be learned best in that way. But I don't think it's the only route to get there and it's an interesting question how far you can get to running around in like 3-D video games and virtual worlds. Like there's a certain fraction of what you get from embodiment in the robot and then there's a certain fraction that you miss. So, there's, I think there are many different pathways that could work to ascend to the top of the foothill that is human level AGI and I mean, there's not only going to be one golden path. Robotics is a real pain. I mean, I spent a fair bit of time maintaining a humanoid robot, the Grace medical robot, which I have here on Vashon with me right. So, I mean, maintaining a robot is annoying, you have to get out the soldering iron and fix connections and so forth and then, Grace rolls around, you have to carry her up and down the stairs. So, there's a

certain friction to dealing with the physical robot. On the other hand, certainly, there's a lot that's learned from understanding how everyday objects are used in the human everyday life scenario and the relationships between people and objects; and that is probably helpful for evolving human understanding in many ways, including ways we haven't explicitly even figured out yet.

That's the end of the first half of the interview; it's split into two halves for attention span and download size. Second half will be next week.

I personally think that AGI and the Singularity are completely interdependent. I think you don't have a singularity without AGI and you don't have AGI without a singularity. It might matter that it is artificial superintelligence rather than AGI, although people like Ben and myself think that ASI is going to follow immediately after AGI anyway.

In today's news ripped from the headlines about AI, Intel has software called 'Bleep' that will remove bad language, for whatever your definition of bad language is, from speech in real time. You could, in principle, put it on a kid's computer when they're gaming and it would remove what you'd decided to be language you don't want them to hear as it comes in. They many categories including "misogyny," "swearing," "ableism and body shaming," "white nationalism," and the "N-Word." It's in beta at the moment and runs on newer Intel chips.

Next week we'll conclude the interview with Ben Goertzel, when we'll talk about Blake Lemoine, the Google engineer who was suspended for claiming that their LAMDA AI was sentient, Ben's SingularityNET Metaverse, and his prediction for when we will have AGI. That's next week on *AI and You*.

Until then, remember: no matter how much computers learn how to do, it's how we come together as *humans* that matters.

http://aiandyou.net