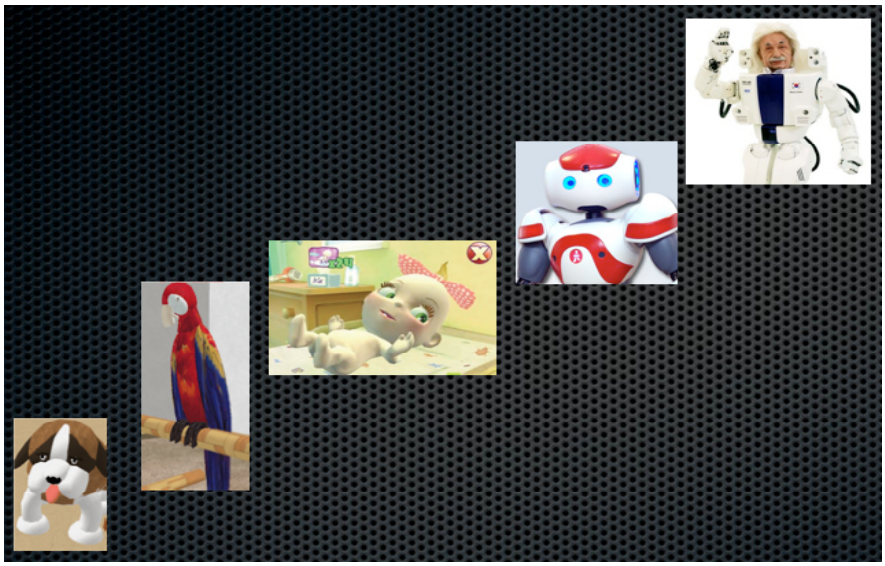
A Path to Beneficial Artificial General Intelligence

Virtual Pets, Robot Children, Artificial Bioscientists and Beyond

Dr. Ben Goertzel
CEO, Novamente LLC and Biomind LLC
Director of Research, Singularity Institute for AI
External Research Professor, Xiamen University

Text of a talk given at Singularity Summit 2009, New York City



To **view a video** of the talk or **download the presentation**
go to http://goertzel.org/Summit09
and click the link near the top

Contents

Hi, I'm Dr. Ben Goertzel and I'm going to tell you about what I think is the shortest and best path we can take to create beneficial AI programs with intelligence at the human level and beyond.

The talk is called "A Path to Beneficial Artificial General Intelligence," but I'm not actually going to sketch out a detailed roadmap – that would take days not minutes. But I'm going to give you a high-level view of the path I see and the ideas underlying it, and give you some pointers to where you can look to find more details.

As you can see from my affiliations I'm not an "armchair AI guy" – I'm going to be talking about work my colleagues and I are actually doing – in my companies Novamente and Biomind, and in the Artificial Brain Lab at Xiamen University in China where I'm an external faculty.

Ten Years to a Positive Singularity – If We Really, Really Try

At the Transvision 2007 conference in Helsinki Finland, I gave a talk called "Ten Years to a Positive Singularity (If We Really, Really Try)." I wasn't able to go to the conference in person so I delivered the talk by video. I put the video online afterwards and it's turned out to be surprisingly popular.

Basically it was a motivational talk. So you can go to YouTube and watch it if you need some motivation! My point was that if we put the kind of money and effort into creating a positive Singularity, that we put into things like wars or television shows, then some pretty amazing things might happen. Look at the US government's response to the recent financial crisis – suddenly they're able to materialize a trillion dollars here, a trillion dollars there. What if those trillions of dollars were being spent on AI, robotics, life extension, nanotechnology and quantum computing? It sounds outlandish in the context of how things are done now – but it's totally plausible.

If we made positive Singularity a real focus of our society, I think a ten year timeframe or less would be eminently possible.

But I'm not going to focus on the timing issue today. I'm going to talk about the roadmap to beneficial AGI, and the points I'm going to make are equally valid whether the roadmap takes five, ten, twenty or fifty years to follow.

Positive Singularity via Embodied Artificial General Intelligence

The Singularity isn't about any one technology – it's about a lot of different sciences and technologies coming together. But I'm going to focus here on the one thing that I think is going to be more critical in bringing the Singularity about: artificial intelligence.

And in particular, Artificial General Intelligence, or AGI. The notion of AGI can be formalized mathematically, but in simple terms, what it amounts to is being able to solve complex, unforeseen problems under complex, unforeseen conditions.

I'm sure there are a lot of different routes to AGI, but what I'm going to talk about here is the one I think is most likely to succeed. This route involves starting out with building simple embodied agents in virtual worlds, then porting these agents to physical robots. Initially these agents won't do anything useful – they'll be like virtual animals or children. But once they get more advanced and learn more they'll be able to help us a lot of ways – for instance by acting as virtual scientists in biology and other areas. And eventually they'll become a lot smarter than us, and the sky's the limit.

Routes to Artificial General Intelligence

When you get into the details there are a lot of possible approaches to making an AGI.

Using Narrow AI. Most of the AI programs around today are "narrow AI" programs – they carry out one particular kind of task intelligently. You could try to make an AGI by combining a bunch of juiced-up narrow AI programs inside some kind of overall framework.

I'm pretty skeptical of this because none of the narrow AI programs have the ability to

generalize across domains, and I don't see how combining them or extending them is going to cause this to magically emerge.

Juicing Up Chatbots. You could take a chatbot – like the new Ramona chatbot my company Novamente made for Ray Kurzweil last year – and try to improve it's code to make it actually understand what it's talking about. Our new Ramona understands a lot more than the last one but still it's far from an AGI.

The risk here is that the architecture of a chatbot is fundamentally different from the architecture of a generally intelligent mind.

Emulating the Brain. You can try to figure out how the brain works – using brain imaging and other tools from neuroscience – and then emulate this in hardware or software.

One problem with this approach is that we don't really understand how the brain works yet, because our software for measuring the brain is still pretty crude.

Another problem is that once you're done, what you get is something with a very humanlike mind– and we already have enough of those!

Evolve an AGI. Another approach is to try to run an evolutionary process inside the computer, and wait for AGI to evolve.

One problem with this is that we don't know how evolution works all that well. There's a field of artificial life, but so far its results have been fairly disappointing.

Another problem with this is that it might take a really long time.

Use Math. You can try to use the mathematical theory of intelligence to figure out how to make AGI.

This interests me a lot – my PhD is in math – but there's a huge gap between the rigorous math of intelligence as it exists today and anything of practical value.

Most of the rigorous math of intelligence right now is about how to make AI on computers with insanely unrealistic amounts of memory or processing power.

Integrative Cognitive Architecture. You can try to build some sort of integrative cognitive architecture – a software system with multiple components that each carry out some cognitive function, and that connect together in a specific way to try to yield overall intelligence.

Cognitive science gives us some guidance about the overall architecture, and computer science and neuroscience give us a lot of ideas about what to put in the different components.

But still this approach is very complex and there is a lot of need for creative invention.

Can Digital Computers Really Be Intelligent?

All these approaches I've described assume that it's possible to make AGI on digital computers. This isn't proven.

It might be that we need quantum computers or quantum gravity computers to make AGI. It might be that building AGI is fundamentally impossible for some reason we don't understand – though I really doubt it.

I know some people disagree, such as Stuart Hameroff who'll be speaking here today, but my opinion is that based on everything we know about the world today, it seems overwhelmingly likely that we can realize powerful AGI on a digital computer. So I've adopted that as my working hypothesis, like almost everyone else in the AI field.

OpenCog: An AGI Architecture Based on Cognitive Synergy

My own approach to building AGI is in the "integrative cognitive architecture" camp.

I can't go into too much detail now – the philosophy of mind underlying my work is in the book "The Hidden Pattern" that I published in 2006.

I've done some proprietary work on AGI in my company Novamente LLC, and I've also launched an open-source AGI project called OpenCog. A lot of the technical details on my approach are online at opencog.org.

I'm also working on a book that describes my approach in detail – it's called "Building Better Minds" and I hope to publish it early next year.

Summarizing my theory of mind and its incarnation in software in a few minutes in the middle of a talk is an impossible task, but, the best approximation I can give is to list five words: perception, memory, prediction, action, goals.

In a phrase: "A mind uses perception and memory to make predictions about which actions will help it achieve its goals."

This ties in with the ideas of many other thinkers, including Jeff Hawkins "memory / prediction" theory, and it also speaks directly to the characterization I gave above of "general intelligence as the ability to achieve complex goals in complex environments."

Naturally the goals may be explicit or implicit to the intelligent agent, and they may shift over time as the agent develops.

Each of these five concepts has a lot of depth to it, and I can't say too much about them in this brief talk, but I'm going to take a little time to say something about memory in particular.

One of the things that the mathematical theory of general intelligence makes clear is that, if you assume your AI system has a huge amount of computational resources, then creating general intelligence is not a big trick. Given enough compute power, a very brief and simple program can achieve any computable goal in any computable environment, quite effectively. So, the problem of AGI is really a problem of coping with inadequate compute resources – just as the problem of natural intelligence is really a problem of coping with inadequate energetic resources.

One of the key ideas underlying my approach to AGI is a principle called "cognitive synergy," which explains how real-world minds achieve general intelligence using limited resources, by appropriately organizing and utilizing their memories.

This principle says that there are many different kinds of memory in the mind – sensory, episodic, procedural, declarative, attentional, intentional. Each of them has certain learning processes associated with it – for example reasoning is associated with declarative memory. And the synergy part is that the learning processes associated with each kind of memory have got to help each other out when they get stuck, rather than working at cross-purposes.

Cognitive synergy is a fundamental principle of general intelligence – it doesn't come up when you're building narrow-AI systems.

In my own AI approach all the different kinds of memory are stored in a single meta-representation – a sort of combined semantic/neural network that we call the AtomSpace.

It represents everything from perceptions and actions to abstract relationships and concepts and even a system's model of itself and others.

So for instance an OpenCog AI system has an AtomSpace, and then it has specific algorithms acting on the AtomSpace corresponding to each type of memory. Each of these algorithms is complex and has its own story.

Declarative knowledge is handled using Probabilistic Logic Networks, which are described in a book we published from Springer last year.

Procedural knowledge is handled using MOSES, a probabilistic evolutionary learning algorithm from Moshe Looks' 2006 PhD thesis. And so forth.

We have a language comprehension system called RelEx that takes English sentences and turns them into nodes and links in the AtomSpace. It's currently being extended to handle Chinese.

Our probabilistic reasoning system can do reasoning based on commonsense knowledge that's entered using English.

And there are a lot of other parts that I won't have time to talk about. But the crux of the cognitive architecture is in how they all work together using cognitive synergy.

Virtually Embodied AI and Virtual Pets

There's a lot of debate in the AI community over whether embodiment is necessary for AGI or not. Personally, I doubt it's necessary but I think it's extremely convenient, and I'm interested in both virtual world and robotic embodiment.

One thing we've been doing is using our AI system to control virtual dogs in the Multiverse and RealXTend virtual worlds.

The virtual dogs see things in the world, and read text that people type about the world, and turn all this into nodes and links in their AtomSpaces.

I'm going to show you a short movie that shows how the AI can use its knowledge of what it sees in the world to help it cut through some of the ambiguities of English.

In this case the AI uses what it sees in the world to figure out what a person means when they say the ambiguous word "it."

Another things we've done with our virtual dogs is to teach them various behaviors by imitation learning.

If you want to teach the dog to sit when you say the word "sit", you can teach it by example. You say the word "sit" and sit down, and it learns to copy you.

This video shows a virtual dog learning to play fetch by imitation learning.

Of course we can also teach more complex things but they don't fit in such short videos.

Finally, this third video shows the dog answering simple questions about itself, its feelings and its environment. Some of the questions are binary but some involve some natural language generation.

Using OpenCog to Control Humanoid Robots

Virtual worlds have a lot of power but there's also something to be said for physical robots that interact with all the messiness of the real world.

So in the Artificial Brain Lab at Xiamen University, we're experimenting with using OpenCog to control the Nao humanoid robot.

Basically we're trying to take the same code that controls the virtual dog and use it to control the physical robot. But it's harder because you need to do real vision processing and real motor control.

One of the things we're doing is what's called Neuro-Symbolic architecture. We use a hierarchical neural net for vision processing, and then link the neurons into the nodes and links in the AtomSpace that represent concepts. So the neural and symbolic systems can work together.

Artificial Bioscientists

Virtual dogs and Nao robots are a lot of fun and I think they're the right path for moving toward AGI. They don't serve any practical purpose.

But once we get smart enough programs through following this kind of roadmap, then we'll be able to combine these programs with narrow-AI systems to get AGI systems that can do useful things – but in ways that are different from how humans do them.

This can be one step on the path to really powerful and general AGIs.

One application I'm particularly interested in is AI scientists – especially in the area of genetics research.

With my company Biomind LLC, I've done a bunch of work applying narrow-AI machine learning tools to analyze genetics data –

trying to figure out how to diagnose and cure diseases, or make people live longer.

Others have used AI to control laboratory robots – the AI designs the experiments, does the experiments, analyzes the results, and then designs new experiments and so forth.

Of course this is really different than having an AGI scientist that makes creative hypotheses and invents new kinds of machinery and so forth. But it's a step on the path.

Using Narrow AI to Analyze Genetics Data: Against Disease and Toward Longevity

I'll take a minute or so now to brag about some of the stuff we've done with narrow AI and genetics data.

We found a way to predict who has Parkinson's Disease from looking at which regions of certain mitochondrial genes tend to have mutations. Human experts looked at the data and didn't find the pattern, but the AI did.

We analyzed mutation data and found the first evidence that Chronic Fatigue Syndrome has a genetic basis. It's not just a matter of "being tired."

We looked at the genetic networks in mice on calorie restriction diets, and learned something about the genes involved in how calorie restriction impact aging. No one knew these genes were involved before. Now we have to do the wet lab work to see exactly what these genes are doing.

Most recently we've been looking at some data from a company called Genescient, who has these flies called Methuselah Flies. These flies live 5 times as long as normal flies – they were produced by directed evolution. The question is why they live so long –what's the underlying genetics? We're probing into this now and finding some interesting stuff.

We find that the differences between Methuselah Flies and normal flies involve some networks that combine metabolism, cancer tumor suppression and stress response. All these things are known to have something to do with aging, but the Methuselah flies seem to combine them in a unique way. So now we need to gather more data and analyze it and try to understand further.

If this is successful, it will lead us to understand how to make drugs that make ordinary flies live as long as the Methuselah flies. There have already been some steps in this direction. And since a lot of these genetic networks are about the same in flies and people, there's reason to hope this could lead to longevity drugs for people too.

My favorite gene that came up in the analysis is one that's called "death executioner BCl-2 homologue". I didn't name it, by the way.

We've also used AI language processing to recognize biological relationships in biomedical research abstracts. This is a really important application because there's more biological knowledge online right now than any human can read. You'll find something like 60,000 research papers on a single topic like "apoptosis" (preprogrammed cell death). The AI can analyze all this and find you new conclusions, and it can point you to information you would have missed otherwise.

To get a real AGI geneticist, what we need to do is hook all this specialized narrow-AI biology stuff up with the AGI framework we have in OpenCog. Then the pace of discovery will really accelerate – and instead of fetching the ball, the AI will be telling us how to make a new longevity drug ... or manufacturing the drug itself.

Creating Ethical AGI

One of the nice things about biology as an application area is that it's helping people.

This brings us to the topic of AI ethics. There are a lot of science fiction movies about AIs going nuts and killing everyone or taking over the world. And it is a real possibility.

If you go online you'll find a website called deadlyai.org. *The Society for the Promotion of Universal Nonexistence through Malicious AI*.

Fortunately it's not serious. I know because I made the site.

But obviously, there is a real threat that someone could make a powerful AGI and use it for the wrong purposes.

In fact there are TWO major threats related to advanced AGI. One is that people might use AGIs for bad ends; and the other is that, even if an AGI is made with the best intentions, it might reprogram itself in a way that causes it to do something nasty. If it's smarter than us, we might be watching it carefully while it does this, and have no idea what's going on.

The best way to deal with this "bad AGI" problem is to build ethics into your AGI architecture. Stephan Bugaj and I wrote a paper on this for the 2008 AGI conference which I'd encourage you to look at.

One funny thing I noticed is that if you go to that talk on Google Video, one of the videos that Google's narrow-AI recommends as similar is a cartoon about Mao Zedong.

Go figure.

One idea that I wrote about recently, related to the ideas in that paper, is "ethical synergy" similar to cognitive synergy. There are different kinds of ethical thinking associated with the different kinds of memory and you want to be sure your AGI has all of them, and that it uses them together effectively.

So, ethics has got to be part of the roadmap to AGI.

Toward an AGI Roadmap

I haven't really told you much about my roadmap to AGI in detail, I've just laid out the broad picture.

Later this month about a dozen AGI researchers are going to gather at the University of Tennessee in Knoxville to talk about the details of making an AGI Roadmap.

What we aim to do is lay out a series of specific tasks for an AGI to do – virtual-world tasks and robot tasks. When the AGI gets to the end of the series it'll be a human-level AGI. But we want to agree on what kind of incremental steps are important and how to measure progress.

What Can You Do to Help?

Finally, I've told you a little about what I'm doing – but what can you do about all this?

If you're a programmer, you can join the OpenCog project. Go to opencog.org. You can help us build a thinking machine.

If you're not a programmer, there are still a lot of ways you can help – the simplest one is to donate money. For which you can also go to opencog.org.

I'd also like to invite everyone to the next AGI conference, which will be March 2010 in Lugano, Switzerland. This year's AGI conference was in Washington DC, and AGI-11 will be in Silicon Valley.

To wrap up I want to make a brief pitch for Singularity activism.

Yes, the Singularity is near. But how near it is – and how well it turns out – depends on what WE do, NOW.

I've outlined a specific path to creating beneficial AGI that can help cure our diseases, help us live longer, and ultimately solve a lot of other problems. The more people we have helping us walk this path, the faster we'll get there and the higher our odds of success.