Greg Diamos: I think computing is a way to give people magical abilities, like superpowers. I like to live in a world where you can give it to everyone. Like, what if every single person had superpowers? I had basically ten years with nothing to do and just a pile of computers in my garage.

Jason Johnson: Just getting the ball rolling, I was wondering, Greg, could you tell me a little bit about what it is you do?

Greg Diamos: I'm one of the people who builds the software infrastructure that's needed to run applications like ChatGPT or Copilot. Let's see, I think it's helpful to start from the beginning. My last name is Diamos. It's a Greek name. It's kind of cut in half. About a hundred years ago, my family immigrated to the U.S. They mostly settled in the West.

Before I was born, my father, he moved to Arizona. And his goal in life was to retire and find someplace relaxing and to go and play golf most of the time. And so he moved to Carefree, Arizona. Most people, when they think of Arizona, they think of Tucson or Sedona or places that normal people would go. You go to the Grand Canyon or something. Carefree is a place where no one ever goes. It's right in the middle of the desert. There are not a lot of people, but it has a lot of golf courses and it's very relaxing. It's like a spa. So it's a good place for that, but it was a very boring and brutally hot place to be as a 10-year-old boy.

[Related: AI Reorients IT Operations]

So I was bored, bored out of my mind. My mom worked as a database administrator for IBM. IBM at the time, they had a mainframe business. And so they would basically throw away the last generation of stuff. She would just take all the machines that were headed to the dumpster, put them up in the car, drive them out and throw them in our garage because it would be such a waste if such a thing was thrown away. I had basically 10 years with nothing to do and just a pile of computers in my garage. So it took me from when I was 10 years old until when I was about 25 to kind of figure out how the machines worked. It seems, like, magical to me that you could build such a thing. You could do that as a little boy stuck in the desert. I did a PhD in computer engineering. After that, I went to Georgia Tech.

Okay, let me tell the story of this one. I really love telling the story. I actually worked at the Baidu search engine. You could think of it as the Google of China. If you go to the search bar, there's a little microphone on it. You press the microphone, you can talk to it. It was based on the last generation of machine learning technology. It still used machine learning, but it didn't use deep learning. This exists right now in all major search engines. Around 2014, 2015, Baidu was in the process of deploying their first deep learning system.

[Related: Alex Karargyris's Path to Becoming a Pioneer of Medical AI]

One of the projects that we had in the Baidu Silicon Valley AI lab was to apply deep learning, which had just been invented and was just starting to work to improve that particular product. Along the way, we had a number of researchers. One of the researchers was Jesse Engel. He was a material scientist, but he was not a deep learning researcher. Actually, none of us were deep learning researchers because deep learning didn't exist. One of the things that he did is he performed sweeps over all the parameters in the system.

One of the sweeps produced this plot that seemed to show that there was a relationship between the accuracy of the system or the quality of the system with the amount of data that went into the model and the size of the model. If you think of a neural network as just being a collection of a ton of simulated neurons, not real neurons, but simulated neurons in that system, trained on thousands and thousands of hours of recordings of people talking. It seemed like as you added more neurons or simulated neurons, and you added more recordings of people talking, the system got smarter. It didn't just happen in an arbitrary way. It happened in a very clear relationship you could fit with a physics equation, E equals mc squared kind of equation, a one parameter equation. It was a very simple, very consistent relationship. We thought that was pretty intriguing. After that project succeeded, the team grew a lot.

[Related: IT Leaders Get AI-Ready and Go]

I inherited a pretty big group of researchers, about a 40-person group of researchers. I had to decide, what should we do? I thought that was the most interesting thing that I'd seen, so let's see if this actually is a real thing. Let's try to reproduce this. Let's try to understand it better. We spent about a year trying to reproduce that experiment, and we absolutely could not break it. It was very repeatable across many different applications.

We tried it for image recognition. We tried it for different types of speech recognition, different speech recognition models. We also tried it for language models. Language models at the time were actually used in speech recognition, basically as a spell check. Essentially what came out of that was this very consistent, very repeatable effect that if you increase the size of the model according to a certain ratio, if you increase the amount of data according to a certain ratio, the quality of the system, like its ability, for example, for a language model to predict the future, like predict the next word, would get better in a very predictable way.

[Related: Seeing AI's Impact on Enterprises]

We published this as a paper. It was called Deep Learning Scaling is Predictable Empirically. This is, I think, came out in 2016. I just thought it was so weird that you could actually predict intelligence. We tried even more to break it. We went and consulted all sorts of machine learning theory experts, and I actually finally understood why it was happening. There actually is a theoretical explanation of why this is happening. The result of all of this is basically we can't break this thing. This is actually a real thing. This is a real relationship that's repeatable.

It's absolutely the most amazing thing I've ever seen in my life because if I try and explain what that means at an application level or a user level, it means that we have a repeatable recipe for intelligence. We have a simple equation that you can write down, and then I can just apply this recipe, and I can create intelligence. The implications of that, we understood what they were. I think, in particular, Dario Amodi, who is also a researcher in our group, really got it, what the implication of that was. He was a biologist who had done very detailed computational biology experiments. He took that and went on to put it into GPT-2. They just kept scaling it, so it kept getting smarter. GPT-2 became GPT-3. You had to have an enormous amount of computation to run this thing.

[Related: Will AI Workloads Overload IT?]

Finally, it seemed to cross a threshold around Chat GPT and GPT-4. It's not just spell check anymore. Now it's abilities that we think of as being very uniquely human abilities, like the ability

to read and write English, the ability to reason logically, the ability to plan, the ability to write software correctly.

It's not like these are just being produced in a research lab. We're actually seeing them integrated into products that are deployed to all users. Chat GPT right now has more traffic than Netflix. It's not just AI researchers who are talking about it anymore, who are playing around with these things. It's actually pervasively available. If we project into the future, as long as we can keep feeding it with data and keep feeding it with computation, it's going to keep getting smarter, and we're going to see new abilities emerge out of it, many other human abilities and potentially even beyond human abilities.

Jason Lopez: Greg Diamos is co-founder of ML Commons, as well as co-founder of Lamini. He's also a founding member of the Silicon Valley AI Lab, where he was on a team that helped develop deep speech and deep voice systems. This is the Tech Barometer podcast. Tech Barometer is produced by The Forecast, where you can find more tech stories about topics from enterprise software to the cloud revolution, AI, and digital transformation. Check us out at theforecastbynutanix.com. This is Jason Lopez. Thanks for listening.