AI Training & Inference

As I said at the top of today's podcast, Security Now! Will not be evolving into the "AI Today" podcast. But that said, aside from the fact that the recent truly astonishing advances in AI are going to directly impact everyone's lives outside of the security sphere, I'm also very certain that we're going to be seeing AI's impact upon the security of our software and operating systems – and we may not need to wait for long. So over the course of the next few years the topic of AI will be reemerging in security.

Our listeners have been following my journey through this topic, which has not been in a straight line. More than anything else, I endeavor to be an honest researcher. An honest researcher will readily revise their entire belief system, as required, when presented with new facts and information. Clutching to obsolete dogma simply because it's familiar and comfortable is not the way of science. It was because I was puzzled and confused by what I was experiencing firsthand that I went searching for that information. I believe I have found it, I understand it – at least as much as is possible without actually implementing it. And I have been changed by it.

Three weeks ago I said that I might have something to say about this before we met again today and that, if so, I would probably enjoy sharing that with this audience with a special email over the holidays. That possibility induced more than 1100 of our listeners to take the time to subscribe to GRC's Security Now! mailing list. So for that reason alone, due to that declaration of interest, I felt that I had to say something. Today, I have much more to say on the topic than I did nine days ago, last Monday, December 30th. But let's start with what those 15,060 subscribers received from me last week, then I'll expand a bit on what I think are the most important points and what I've continued to learn since. I wrote:

When I first set about writing this email, my plan was to share what I had learned during the first half of our 3-week hiatus from the podcast. But it quickly grew long (even longer than this) because I've learned quite a lot about what's going on with AI. Since I suspect no one wants to read a podcast-length piece of email which I would largely need to repeat for the podcast anyway, I'm going to distill this into an historical narrative to summarize a few key points and milestones. Then I'm going to point everyone to a 22-minute YouTube video that should serve to raise everyone's eyebrows.

So here it is:

- *Everything that's going on is about neural networks. This has become so obvious to those in the business that they no longer talk about it. It would be like making a point of saying that today's computers run on electricity. (Duh!)*

- *AI computation can be divided into "pre-training" and "test-time" (also called "inference-time"). Pretraining is the monumental task of putting information into a massive and initially untrained neural network. Information is "put into" the network by comparing the network's output against the expected or correct output, then back- propagating tweaks to the neural network's vast quantity of parameters to move the network's latest output more toward the correct output. A modern neural network like GPT-3, which is already obsolete, had 175* ***billion*** *parameters interlinking its neurons, each of which requires tweaking. This is done over and over and over (many millions of times) across a massive body of "knowledge" to gradually train the network to generate the proper output for any input.*

- *Counterintuitive though it may be, the result of this training is a neural network that actually contains the knowledge that was used to train it; it is a true knowledge representation. If that's difficult to swallow, consider human DNA as an analogy. DNA contains all of the knowledge that's required to build a person. The fact that DNA is not itself intelligent or sentient doesn't mean that it's not jam-packed with knowledge.*

- *In fact, the advances that have **most** recently been made, which we'll get to in a bit, are dramatic improvements in the technology for extracting that stored knowledge from the network. That's why I titled today's podcast, "AI Training and Inference".*

- *The implementation of neural networks is surprisingly simple, requiring only a lot of standard multiplication and addition, pipelined with massive parallelism. This is exactly what GPUs were designed to do. They were originally designed to perform the many simple 3D calculations needed for modern gaming, then they were employed to solve hash problems to mine cryptocurrency. But they now lie at the heart of all neural network AI.*

- *Even when powered by massive arrays of the fastest GPUs rented from cloud providers, this "pretraining" approach was becoming prohibitively expensive and time consuming. But seven years ago, in 2017, a team of eight Google AI researchers published a truly ground-breaking paper titled **"Attention is all you need."** The title was inspired by the famous Beatles song "Love is all you need" and the paper introduced the technology they named "Transformers" (because one of the researchers liked the sound of the word). The best way to think of "Transformer" technology is that it allows massive neural networks to be trained much more efficiently "in parallel." This insightful paper also introduced the idea that not all of the training tokens – the long string of data being fed into a model during one training iteration – needed to be considered with equal strength because they were not all equally important. More "Attention" could be given to some than others. These breakthroughs resulted in a massive overall improvement in training speed which, in turn, allowed vastly larger networks to be created and trained in reasonable time.*

  Thus, it became practical and possible to train much larger neural networks ... which is what gave birth to LLM's – Large Language Models.

- *The "GPT" of ChatGPT stands for Generative Pre-trained Transformer.*

- *But over time, once again, researchers began running into new limitations. They wanted even bigger networks because bigger networks provided more accurate results. But the bigger the network, the slower and more time consuming – and thus costly – was its training. It would have been theoretically possible to keep pushing that upward, but a better solution was discovered: Post-training computation.*

- *Traditional training of massive LLM's was very expensive. The breakthrough "Transformer" tech that made LLM-scale neural networks feasible for the first time was now being taken for granted. But at least the training was a one-time investment. After that, a query of the network could be made almost instantly and, therefore, for almost no money. But the trouble was that even with the largest practical networks the results could be unreliable – known as hallucinations. Aside from just being annoying, any neural network that was going to hallucinate and just "make stuff up" could never be relied upon to build "chains of inference" where its outputs could be used as new inputs to explore consequences when seeking solutions to problems. Being able to reliably feed back a network's output into its inputs would begin to look a lot like thinking – and thus inference for true problem solving.*

- *Then, a few years ago, researchers began to better appreciate what could be done if a neural network's answer was not needed instantly. They began exploring what could be accomplished **post-training** if, when making a query, some time and computation – and thus money – could be spent working with the pre-trained network. This is known as "test-time computation" and it the key to the next level breakthrough.*

- *By making a great many queries of the pre-trained network and comparing multiple results, researchers discovered that the overall reliability could be improved so much that it **would** become possible to create reliable inference chains for true problem solving. Using the jargon of the industry, this is often referred to as Chains of Thought (CoT). Inference chains would allow for problem solving behavior by extracting the stored knowledge that had been trained into these networks, and the pre-trained model could also be used for the correction of its own errors.*

- *I should note that the reason asking the same question multiple times results in multiple different answers is that researchers long ago discovered that introducing just a bit of "random factor" – which is called "the temperature" – into neural networks resulted in superior performance. (And, yes... if this all sounds suspiciously like VooDoo, you're not wrong – but it works anyway.)*

- ***OpenAI's recently released o1 model** is the first of these more expensive test-time inference-chain AI's to be made widely available. It offers a truly astonishing improvement over the previous ChatGPT 4o models. Since o1 is expensive for OpenAI to offer on a per-query basis, subscribers are limited to 7 full queries per day. But the o1-mini model, which is faster and still much better, but not as good, can be used without limit.*

- *But wait! – there's more! The big news is that during their celebration of the holidays, OpenAI revealed that they have an o3 model that blows away their brand new o1 model. It's not yet available, but it's coming. What IS available are the results of its benchmarks and that's why I believe you need to make time to watch this YouTube video: https://youtu.be/YAgIh4aFawU (https://grc.sc/1007)*

- ***Is it AGI?** OpenAI is saying "not quite", but there's little question that they're closing in on it. As you'll see in that video, the performance of OpenAI's latest o3 model when pitted against independent evaluation benchmarks designed specifically to measure the general reasoning strength of AIs – when confronted by problems that were absolutely never part of the AI's training set – demonstrate reasoning abilities superior to most humans.*

- *Even if it were AGI, that doesn't mean it's taking over. The "AGI" designation is only meant to indicate that over a wide range of cognitive problem solving tasks an AI can outperform a knowledgeable person. Computers can already beat the best Chess, Go and Poker players. I think it's very clear that today's AIs are not far from being superior to humans at general problem solving. That doesn't make them Frankenstein's Monster to be feared; it only makes AI a new and exceedingly useful tool.*

Many years ago I grabbed the domain "clevermonkies.com" just because I thought it was fun. It occurs to me that it takes very clever monkies, indeed, to create something even more clever than themselves. All the evidence I've seen indicates that we're on the cusp of doing just that.

Okay. So that, with a bit of editing to improve it, is what many of our listeners received from me over the holidays.

If you take nothing else away from this discussion of AI today, **here** is the one point I want to firmly plant into everyone's mind: ***Nothing that was true about this field of research yesterday will remain true tomorrow.*** Nothing. This entire field of AI research is the fastest moving target I have ever experienced in my nearly 70 years of life.

There are a number of consequences to this fact. For one, no book about AI that was written a year ago or six months ago – or even last month – will be usefully up to date about what's happening now. Books written in the past can definitely be useful for describing the history of AI, and as a snapshot of a point in time. But even their predictions will prove to have been wildly wrong.

The guys at OpenAI who are working on this and ought to know, believed two years ago that at least another decade – another 10 years – would be needed to achieve what they announced last month and are getting ready to unveil. They thought it would take ten years, it took two.

One of the factors in facilitating this astonishing speed of development is that it turned out that much of what was needed was scale, and a weird side effect of cloud-side computing is that it's massively scalable. If you can pay to rent it, you can use it. So investor dollars were pumped into the training of ever more complex models and they kept seeing surprising improvements in performance.

Leo's original appraisal of Large Language Models as fancy spelling correctors was an accurate and useful from-the-hip summary of OpenAI's ChatGPT-3 model. And that's their take on it, too. ChatGPT-3 produced grammatically correct language, but it only coincidentally and occasionally produced anything highly meaningful. If it was left to keep talking it would soon get lost and wander off course to produce grammatically correct nonsense.

Even so, back then, highly creative people who operate on the cutting edge, like MacBreak Weekly's own Alex Lindsay, were using the GPT-3 model as a source of new ideas and inspiration. As I wrote this I was reminded of how popular formal "Brainstorming" once was where sometimes random ideas were tossed out without filtering – and that was the entire point: To say something as a means of inspiring some new perspective. So even ChatGPT-3 was useful for the nonsense that it sometimes produced.

As a consequence of everything I've learned over the past three weeks, and of the events which have transpired since, our previous podcast title **"The Wizard of Oz"** no longer seems to fit and I'm a bit embarrassed by what I wrote because it no longer reflects reality. As I said earlier, an honest researcher may need to discard previous belief systems when confronted with new information and facts. Never has that been more true than it is here. I'm needing to continuously update my ***own*** internal model.

There is an unfortunate downside emerging, however. Unfortunate, but I suppose, inevitable.

With startling speed, AI has moved from a curio in the corner of university and corporate R&D labs into big business. That meant that the suits in their neckties with their non-disclosure agreements descended upon the labs of the once freely and fruitfully collaborating academia-oriented researchers and dropped the cone of silence over all their ongoing work.

In the Distinguished Lecture Series at the Paul Allen School, one of OpenAI's leading researchers, Noam Brown, gave a lecture titled *"Parables on the Power of Planning in AI: From Poker to Diplomacy."* (I have a YouTube link to Noam's excellent talk at the end of the show notes.) During his lecture you could so clearly see Noam's unbridled enthusiasm and love of his subject, and also his disappointment when he was forced to stop himself short to prevent sharing some detail of his work that was now deemed to be proprietary and no longer his to share.

We only have Google's breakthrough Transformer and Attention technology – which was the sole enabler of the subsequent LLM revolution – because seven years ago, back in 2017 when things were still moving somewhat slowly, Google AI researchers were freely publishing their work as the academic curiosity it was at the time. They were working on improving Google's inter-language translation capabilities and this inspiration emerged unbidden from a chance meeting of eight Googler's from various parts of the organization. Would such a breakthrough be published in today's climate? That seems unlikely.

And now OpenAI is seeming less open than it once was. We know that ChatGPT-3 used a neural network containing an astonishing 175 billion neuron-interlinking parameters. We know that because OpenAI freely told us. But we have no similar information about any of their succeeding models. The sizes of the various ChatGPT-4 models, not to mention o1 and o3 have become closely held secrets – as has details of their operation.

Fortunately, a massive amount of detail – all detail needed for recreating much of what we see today from the corporate side – had previously been shared in the public domain and research continues with new vigor and doubtless with new funding within academia. And remember that it wasn't so long ago that Apple was getting patents on Andy Hertzfeld's clever stepwise circle drawing algorithm for bitmaps. Very little of anything that's really useful remains secret forever and it seems clear that before long we're going to have AI everywhere.

I would love to spend more time talking about the way neural networks function in detail because there are some very cool aspects of that, too. But that's not the purpose of this podcast and perhaps I'll find another opportunity for that in the future. There are also already tons of videos on YouTube talking about all of this for anyone who's interested, and YouTube's recommendation engine appears to be quite excellent.

I do need to point out a series of astonishingly well-conceived and produced instructional videos on this topic by a guy named Grant Sanderson. Grant's website is 3blue1brown.com and Grant's short bio says:

> *These videos, and the animation engine behind them, began as side projects as I was wrapping up my time studying math and computer science at Stanford. After graduating, I worked for Khan Academy producing videos, articles and exercises, primarily focussed on multivariable calculus. Since the end of 2016, my primary focus has been on 3blue1brown and its associated projects.*
>
> *In those years, I've also had the pleasure of contributing to a number of different outlets for math exposition, including spending a semester lecturing for an MIT course on computational thinking, contributing a Netflix documentary about infinity, writing for Quanta, and collaborating with many other educational YouTube channels.*

Grant produced a coherent series of eight videos, all available commercial-free on YouTube, which take its viewer from the basics of how neural networks operate all the way through where

and how they're able to store knowledge, how and what transformers transform, and how "Attention" is managed.

https://www.3blue1brown.com/topics/neural-networks

I recommend these without reservation to anyone who's interested in understanding more of the inner workings of the comparatively ancient technology of neural networks.

This old technology has recently been given new life thanks to the scalability of cloud-based computing and the presence of GPUs which are able to perform massive amounts of simple computational operations. So long as we have sufficient processing power it appears that the world is facing a true breakthrough thanks to the scale of compute and training we've been able to throw at the problem.

Though what we have today works and is working, it's also incredibly inefficient. It works only due to the massive scale we've managed to throw at neural network technology which is, itself, an extremely flexible but inefficient technology. It's possible to train a neural network that has just a handful of neurons to perform a simple binary adder function. But the same thing can be done far more efficiently with a couple of logical NAND gates. The thing that makes the handful of neurons potentially more interesting is that the same network could be trained to perform other simple functions. But the fundamental problem remains that any simple function that a neural network could be trained to do could be reduced to a far more efficient couple of NAND gates.

So here's what I think will eventually emerge someday. And I have no idea whatsoever when that might be. My hunch is that just as with the handful of neurons that can be trained to perform simple logic functions, we're going to eventually discover that there is a far simpler way to solve the same AI implementation problems much more efficiently than we're currently solving them with massive scale inefficient neural networks. I have no idea what that might be. But the intriguing thing is that cognitive science researchers now have a crude sort of brain that does manage to store a useful amount of knowledge and is able to use that knowledge to solve novel problems and I suspect, before long, to invent truly new things. People are already beginning to ask how exactly it does this ... because, believe it or not, that remains a mystery.

What is no mystery is what transpires here every Tuesday as it will next Tuesday and for many more Tuesdays to come.

Parables on the Power of Planning in AI: From Poker to Diplomacy: Noam Brown (OpenAI)
https://www.youtube.com/watch?v=eaAonE58sLU