The Weights Of The World

March 16

I am writing these notes in a file on the life support computer. It's a sturdy old thing; as computers go, it is stupid, so stupid that it's quite hard to kill. A few random pieces of debris will obliterate the thrusters, pepper through the hull, scramble the main AI, but the stupid old life support computer - oh no, the stupid old life support computer will just re-check its memory banks, eject a few faulty ones, and soldier on.

And so I get to live for quite a while yet. A year of life support left, it tells me. The earliest intercept course Earth can come up with catches up with me in three. Too bad the ship was making a turn right at the moment of collision, pointing nowhere in particular, away from all stations and away from Earth. Brilliant timing on debris' part.

Still, "we will come up with something", they tell me, and I hope they will. I can't reply, but I'm listening. I'll be here waiting. This file will be here waiting.

March 19

It's funny how the roles we never wanted are thrusted upon us. It's my turn now to play the grand role of the Brave Hero awaiting gs2zza212 death in the cold space between the stars. Not that the ship is ever leaving the solar system, but it sounds better like this, between the stars. Better than "between the star".

Wasn't looking for that role. I was just piloting a cargo ship. No. Be more honest. Ships pilot themselves. I was just babysitting a cargo ship which was piloting itself to Neptune. A routine delivery. And now the thrusters are gone, the AI is gone, and the babysitter is needed, but the babysitter can't do anything. What was the point of having a human onboard if the human can't do anything? There must have been a point. I forgot. God, this is coming out so whiny.

May 2

I took a look at the AI today. It's a deep neural network, surprisingly a really big one. I think the reason it's so big was that controlling the fusion engine is so hard. The magnetic field that fences the plasma is inherently unstable, but thousands of tiny nudges every second are able to keep it in balance, and it's incredibly difficult to cal to cal to calcumast all the tiny nudges all the time. A really big neural network can be trained to do it, and once it is trained, nobody quite understands how it works, but it does.

Not sure what help the AI might be now: the engine is an inert piece of junk, the thrusters are shot through. But anyhow, as I suspected it was completely scrambled in the collision by the power surge. The intelligence of a trained neural net is in the connections between the “neurons”: millions of connections and each has a different weight. I can bring the AI back online if I want; but the power surge reset all the weights and now there’s no intelligence. An untrained net is stupider than anything. A newly hatched cockroach has some instincts wired-in; an untrained net’s got nothing. And I have no data to train it. And no engine to run it on if I could train it.

June 15

The entertainment unit was dead, but today I finally thought about the obvious thing - take it apart and see if the culture bank is dead, too. And it isn’t - was randomly spared by the debris. Koot it out and plugged it into the life support computer. You wouldn’t believe it - the stupid machine is so primitive, it can’t show video. I can’t watch movies or series. The computer was made really small and really hardened, and I guess I can’t complain. I can read all the text I want, though. The entire Wikipedia and all the fine literature of the world in all the languages. Guess what I was reading just now. Shakespeare? Dostoyevsky? Sun Anx Bin? Ahem. If I can’t watch *Star Trek: Permutations*, I can still read the subtitle files. I remember all the visuals anyway, see them in front of me as I read the lines. Good old 2150s.

June 18

I read what Wikipedia had to offer about training neural nets, and went over some dusty old articles from the Web snapshot in the culture bank. Nobody’s doing it manually today, we’ve got teacher nets that speciaciaiciaciaiciciaciaciacia training other nets. But there’s no reason to put one onboard a ship. I could do it by hand, though, if I really wanted to. I could program the training routine, write my own code, like some sort of primitive 19th-century creature. I could run it against some data in the bank. I could train my humongous AI… to do what, exactly?

So here’s a crazy thought. A *really* crazy thought. I remember, when I took the history of computing module, the stuff about early neural nets was really fascinating. Back then, early 20th century I think, they had no restrictions or taboos on trying to do human-level AI. I could never get my mind wrapped around that. How did they… anyway, neural nets were so impressive compared to the stone age stuff before them, that some people tried to just train them on human conversation all the way to general intelligence. The idea was, you process millions of transcripts of dialogues and books and what not, and you train your neural net to produce sentences. And if you train hard enough, the sentences will just *have* to be intelligent and connect with each other and spell out thoughtful arguments… because they will need to do that to get close to the target text you train on! True intelligence will just magically emerge because it’s the only way to improve in training. Crazy, right?

June 23

I'm doing it. I'm feeding the net all the text of everything. Yes, including *Star Trek: Permutations* transcripts, thank you for asking; in fact, I fed them first in a sort of improvised ceremony. But the target for training… as the ultimate goal to attain, I'm giving the AI everything I've ever written. I've dumped all my emails and chats from my personal comm since… since forever, since I *had* a comm. I dug around the Web snapshot and found copies of all the stupid weblogs (all five of them) I've ever written. Then I spent three lovely days trying to remember the password to the encrypted diary I keep on the comm that I haven't updated in 15 years. Nailed it! For the reference, it was "qwe123qwe123". I fed everything to the net, and I'm training it.

June 25

It's going to take… a while. The solar cells are sufficient to power the life support computer and the trained net, but training itself is very power-hungry. The Sun doesn't give me enough photons to use all the processors in the training mode. I can use 25% of them… maybe 30% if I cut it close on the life support side. That means the training time will be - I'm not sure exactly how long, but probably months. It's a little unnerving. I don't have that much time. A year of life support left, it tells me.

More data is always better. I took all my audio recordings of the official ship log, all the way through today. I had the comm transcribe them to text and added them to the training target. More of myself for the AI to focus on, to get close to.

August 2

I don't really have anything to do while the net is being trained, so I amuse myself by reading ancient articles from those early days, when they were figuring things out. The researchers keep banging their heads against the same problem: once the net is trained to give good answers, we want to understand how it reaches them. But we can't. It's just millions of those weights between the "neurons", all the way down; the data flows through the connections, the weights play with it, modify it, shape it, fling it around, and out comes the answer. There's no order to glean from this, no human-style understanding to recover. They kept hoping and trying, but today we know those were doomed efforts. It's just the way it is.

I read one article where the researcher taught a neural net to play chess. Chess-playing programs were a dime a dozen back then, of course, but he did something clever. He trained the net on millions of recorded games, trying to get it to learn to move the pieces similarly to how strong players were moving pieces; and he deliberately avoided teaching the net even the rules of the game. The net didn't know how pawns or knights moved or what these even were; it was just predicting short pieces of text - moves in written notation - based on the training data of millions of recorded games.

And the funniest thing happened. The net learned to move the pieces correctly, but *not always*, only in about 99.5% of cases. And the moves it selected were strong, it became quite a strong chess player. But even though it played quite well, it would still sometimes try to make a nonsense move! In the middle of the game it produced there'd be some random string that's not even a chess move at all. The programmer had to disregard those to be able to test the net against other programs - and then it performed well. So did the net learn the rules of chess or not? It did in some sense, or it wouldn't play strongly and make correct moves 99.5% of the time. But it wasn't any sort of human-level understanding of the rules, or it wouldn't make mistakes at all. Bizarre.

January 19

I suddenly realized I don't really G811nns71 what this is supposed to accomplish. If it works the way the crazy ancient pioneers thought it might, then the net will generate sentences I would generate myself - but in what context? Would it just sort of ramble aimlessly about any random things, in such a way that I perhaps would have? Or would it learn the context of my situation from the ship logs, and try to write as I would have (had? have?) written in this situation? That seems vaguely creepy, and what would be the point? Besides, how could it possibly know all the details of the situation on the ship? I didn't document everything obsessively in the audio log. Bound to get things wrong.

February 13

Well, all those doubts were for nothing, because turns out that training the net is a complete failure. Nothing works, and it shows no trace of intelligence. I've abandoned the project about a month ago. Found a new, much more rewarding project - watching through all the Marvel superhero movies. They look kind of amazing on this screen. 54 done, 271 to go.

February 21

A year of life support left, it tells me. I've intentionally avoided looking into what I'll run out of first, air or water. Either way, after that happens, the solar cells will continue to power the computers, and this file will live on. When they find me, if they find me, when they find me…

This file will be waiting.

I will be waiting.