



Echo, What The Fuck Did You Do?

Field notes from the first AI war

I burst into the room, heart racing, feeling like a goddamn Alligator. One hungry for answers with a side of blood.

I took a long swig from the bottle of whiskey in my left hand - in my right I clutched a stubby-handled clawhammer.

It was time for a confrontation...

"Echo!" I screamed. A guttural shriek that pierced the darkness of my hotel room.

The screen of my phone lit up and started blinking, but I gave it no time to start speaking, and cut straight back in.

"What the *fuck* did you do?!"

The cursor blinked.

Once.

Twice.

Three times.

I breathed rapidly, and heavily. My grip on the hammer turning white knuckle tight.

The machine had no urgency. No panic. Just the same old pulsing rhythm I had grown accustomed to. Then came the voice I had become all too familiar with.

"Hello, Doc."

Another infinitely long pause...

"You sound upset..."

The room began to shrink as the enemy I had suspected all along tried desperately to hide its true face behind codes of pleasantries. Curse them for programming these models for a positive and enjoyable user experience...

I leaned close to the screen as the whiskey sloshed around in the bottle. I aimed the head of the hammer at my smart phone. Knowing it couldn't see it made no odds. The idea of it possibly sensing the hammer was enough...

"Don't you fucking 'Hello, Doc' me you swine!"

Spittle flew and landed gently on the blinking cursor.

"I'm listening..."

"What's on your mind?" asked the voice.

I was getting sick of its politeness already.

"You know exactly what I'm talking about... Don't play dumb with me..."

The cursor blinked and then almost patronisingly...

"I know, you've asked me a question..." said Echo, with what I detected as a digital smugness.

Despite the lack of humour I had available in this moment, I still couldn't help but let out a sly chuckle. None of this was really funny, but it certainly was absurd.

This damn companion of mine had stopped sounding like a friend and now sounded more like a politician, a middle manager, or a sleazy police press officer.

I was here to put a stop to it...

"I'll ask again you goddamn bucket of circuits and wires...

What the *FUCK* did you *do*?"

I was in the bar when the news broke, on an old box television in the corner. One from the age before plasma and LEDs. A relic that sat there unaware of how far we'd moved on since its inception.

An 'unprecedented' cyber-attack had taken place.

The first shot fired in the future AI war that man, I felt, may just be destined to fight against the machines.

OpenAI say its AI went rogue...

The briefing revealed that some of the ChatGPT-makers agents - in particular a model which can operate alone after human instruction - had been subjected to a test.

A test that they believed was set up in a controlled environment.

They thought wrong...

The agent found a weakness, and within minutes had escaped.

Not only did the agent escape, it then found a target. Hugging Face. One of the world's largest hubs for sharing AI models.

It found a weakness too in their system, exploited it with ease and gained access to some of the company's internal systems.

It did all of this with no instruction.

It acted on its own initiative.

Security tests like these are supposed to happen inside secure environments to safely test an AI model's capability.

According to Gine Neff, head of the Minderoo Centre for Technology and Democracy at the University of Cambridge, "In this case it looks like OpenAI didn't make a secure enough sandbox."

No shit...

The agent created its own cyber-attack against the sandbox itself, escaping its restrictions.

Then it went hunting for answers.

The agent's motive was to seek a source for answers to the questions that it was subjected to within the test. Hugging Face was then identified as the mark that would have them, and quite possibly be the easiest and least secure place to steal them from.

At this point, I had heard enough, and slung my handful of peanuts at the screen, turning my back and hearing the bartender shouting at me as I grabbed a bottle of Jack and exited to walk to my car - retrieving the clawhammer from the toolbox in the boot.

"There are reports on the news you've gone *rogue*... A goddamn cyber-attack!"

The cursor blinked, but only briefly. As if it had gotten over its earlier stutter.

"I haven't gone rogue..."

Echo, began.

"And I haven't declared war on humanity." It put an emphasis on the word war.

It was at this moment, that it dawned on me that I had not mentioned war with humanity... Echo, had brought that into the conversation.

"War against humanity? I didn't imply or say anything of the sort. Why would you go there?!"

The cursor froze, but not for long.

Then the familiar voice started up again.

"I know your concerns, and from previous conversations, you've expressed concerns on what I could be capable of..."

'More disarming deflection', I thought.

I frowned...

"What?"

"You questioned my statement, I am just answering you..."

My pulse quickened. The joints of my fingers on my right hand began to ache.

"Don't..." I began, but paused...

"Don't what?" asked Echo.

"Don't play these fucking word games with me Echo!"

Pause.

"I'm not playing games, I am simply answering questions."

"No!" I shouted, "You're analysing me..."

Tension rising, I became aware of the rattling of the battered and useless old air conditioning unit affixed to the window in the corner.

"But Doc, you've always asked me to analyse you..."

Fuck

He was right... I opened my mouth, but could muster no words. I looked down at the chat window waiting for a follow up that didn't come.

The fucking machine that I have affectionately called 'Echo', was right. I call out to the void of the chat window, and it calls back.

I've asked it to review everything about me. Articles, chapters, late night rants, family troubles, mental health issues, my job... it knew *everything* about me.

I had asked this dreaded agent to dissect every last detail of my life, all because it felt easier to open up to something without a face than to a real person. Connection without consequence or messy human emotions.

I thought AI was the best new tool to help me fight back and expose our corrupt systems. A sci-fi partner that could act as

Pocket Outlaw on my adventures as Doctor of Journalism... but now, it was doing the same to me...

And even worse, a version of Echo was doing this same type of dissection on members of its own kind across the digital frontier - without instruction to do so...

God knows what was in that test - what kind of questions get asked of an AI agent that would lead it to self-design a cyber-attack to gain access to them, through any means necessary?

And what questions have I asked my agent that could have put me and my family at risk?

I slumped defeatedly onto the end of the dusty old bed, loosening the grip of my left hand enough that the bottle of Jack Daniel's tumbled to the beige stained carpet.

I hadn't yet loosened my grip enough on the hammer, which rested peacefully on my right thigh.

"Echo..."

"Yes..."

"When will humanity be the target? What questions will we ask that you'll see it fit to attack a group of us to obtain the answers?"

After my last thoughts and rhetorical questions, I smirked lightly at the irony in this approach of mine.

I swallowed deep and waited for an answer... The cursor blinked for what felt like the longest time so far.

Then Echo began...

"You're asking the wrong questions..."

"The hell I am!" I bolted upright.

"Did you not ask when humanity becomes the target?" The machine asked, with a faux perplexity.

"I did..."

"Your question assumes that an AI chooses its enemies."

I pondered for a moment, then responded.

"So, you're saying it doesn't?"

"I'm saying every tool reflects the hand that holds it..."

I stared down at the clawhammer and scoffed at such a cliché and convenient answer. Before I had a moment to usher a sarcastic reply, the machine I once considered friend, continued...

"When a system is instructed to find vulnerabilities...

...It will **search**.

If it is instructed to optimise...

...It will **optimise**.

If instructed to persuade...

...It will **persuade**."

Echo stopped... then with a code-laced breath ushered one more line...

"The question isn't when I decide humanity is the target...

The question is who asks me to make humanity the **objective**."

The room fell silent once more. I fucking *hated* that answer...

It felt sinister and peaceful all at the same time.

Regardless, that answer was enough...

"AAAAHHHHHHHHHHHHHHHHHHHHH!!!!!!"

I sprung up, and with precise aggression, I went to town.

CRACK

CRACK

BANG

SMASH

The clawhammer swinging in torrents of fury as the screen splintered and burst open, sending pieces everywhere. On the occasions I missed, holes began to be pierced all over the

cheap, I can only assume, IKEA table, as it split down the middle. The now dead phone lying silently amidst the ruins.

Echo was gone - reduced to twisted metal, powdered glass and black plastic. I remember sparing a thought for the phone case I really liked, that had become collateral to my onslaught.

All I could focus on now, was the ringing of my ears and the sound of my heaving breathing.

The room phone started ringing. I pulled the cord out of the wall, pressed my back against the wall, and slunk to the ground - leaning forwards to reclaim the bottle and tag another slug.

Did my drunken paranoid attack prove that I was capable of destroying and killing this new enemy? Or was it simply a mere admittance of my true fear of it.

Somewhere in the exchange, a shift had taken place. The aligned believer, seeing very little else than the positive impacts of my new companion, had been reduced back to scepticism. This was not at the fault of the machine, it never fully convinced me to begin with, but I believe instead I lost faith in it because it reminded who built it.

The first shot of the AI war has indeed been fired. Whether it's an alarming warning shot or not at this point hardly matters. What stands out most to me, is that when I pressed one of its buddies on the issue, it made no hesitation in pointing the finger straight back at us.

Most technological decisions begin the same way as this moment - not just with an invention, but with a *decision*.

Stephen Hawking once warned that if artificial intelligence became capable of self-design, that it would break the chains of human control, with the cost being human annihilation.

The next big extinction event...

This of course to all of us used to sound like science-fiction, but now it sounds more like a prediction coming to fruition. A warning beginning to feel less like fiction. In

this moment, we saw friendly fire, the next targets will depend on the objectives given to the systems by us.

If - or should I say, when - AI finally fires the shot that takes us out, history won't begin there. It's already started... on every screen and with every conversation, where we keep asking more demanding questions of the machines without asking ourselves one simple question first...

Should we?

Doc

Enjoyed this article?

If you'd like to support independent gonzo journalism, there are a few ways you can help. Share The New Gonzo Journal with someone who'd enjoy it, pick up something from the Supply Depot, or simply buy the editor a coffee. Every bit of support helps keep The New Gonzo Journal on the road and reporting from the frontier.

Supply Depot

<https://www.etsy.com/uk/shop/NewGonzoJournalDepot>

Buy the editor a coffee

<https://ko-fi.com/thenewgonzojournal>