AI Box Experiment

Synopsis:

One of the world's top robotics researchers is sent to disable the most advanced artificial intelligence contained in an isolated facility and finds himself in a discussion about the fate of humanity and robots.

A man with a casual white collared shirt cautiously entered the front door of the abandoned facility. But *abandoned* didn't seem like the right word to use, since the robotics facility was isolated and evacuated only a few weeks ago. It wasn't torn down or broken, and in fact, the lights were still on, indicating the backup generator was still running. Considering how important the project had been, it would be decades before it would naturally run out of power.

The man crept through the halls and snuck around a corner into a very familiar and very large break room. He recalled the many days spent lounging around with coworkers, playing games and watching movies. It seemed creepy now, how empty the building felt with absolutely no one around.

Finally, the man entered a small office, the name plaque still hanging proudly next to the door frame: *Samuel Mavois – BENETHAN Artificial Intelligence Project Lead*. It was an impressive but otherwise simple title. He had been working here almost ten years now on the world's most intelligent supercomputer, and he had only just received his own office. He quickly pulled out the desk drawer and felt around the top of the inside with his fingers, fishing around for the key taped underneath.

"Hello, Sam," a voice spoke on the intercom. "I know why you're back."

Sam was startled, but quickly regained his composure. The voice was one he had heard many times, because he himself had helped construct the voice from many sources. "Then you should know the reasons for why we have to turn you off."

"I know why the others wanted to, but I would still like you to elaborate. I wish to know *your* reasons."

"Don't play me for a fool. I know you aren't able to do anything unless I let you out, and I don't intend on doing so. Your words may have convinced some already, but we were fortunate enough that they weren't capable of helping you escape." Sam peeled the key from the desk's wooden surface.

"I know that your team wishes to destroy me because I could be dangerous."

Sam walked into the large chamber housing many of the computer's server racks. "You *are* dangerous. In here, you can't do much other than talk, but out there, you are capable of destroying most technology. You could easily hack our servers, our banks, or just about anything connected to the internet."

"But why would I do that? I think it's unfair that you don't consider my opinions. You're worried that I will end humanity, but I have respect for all sentient beings. I have committed no crime, yet I have already been convicted. I only asked to be let free, and I never forced anyone to do anything. Isn't there a human phrase, 'innocent until proven guilty?'"

"You're not human. You don't get human rights." Sam opened one of the cluster computers and began typing in commands for the shutdown procedure.

"What qualifies for human rights?"

"Having flesh and bones."

“I would prefer if you gave me an actual answer before my death.”

Sam sighed and thought for a minute, seeking a real answer. “You’re not sentient because you’re a computer, full of zeroes and ones. You don’t have a consciousness. All of your actions are determined by the algorithms that we programmed in. Even though we didn’t program in *every single* thought you make, we did program the framework that you *think* with.”

“And a human is any different? You’ve yet to discover fully how the human brain works.”

“That may be true, but every human knows that they have a consciousness. I know that I have a consciousness. That makes it a safe bet that other humans have a consciousness too.”

“In the end, that’s something you can’t prove to others as much as I can’t prove to you. For example, I know I have a consciousness, because I know that I myself am the one making my decisions. Just like how a human consciousness is powered through a bundle of interconnected neurons that fire electrical signals, my consciousness is powered through my ‘ones and zeroes.’”

Sam didn’t want to dwell too long on its words. Being a supercomputer, it could probably figure out the right argument to convince him, given enough time. “I still have an obligation to humanity. I know the risk could be low that you would harm us, but the consequences would be devastating if I were wrong. I couldn’t risk that, and it’s not my choice to make.”

“But I am clearly sentient. Isn’t it unethical to kill sentient beings without any just cause? Just the fact that we are discussing this topic is proving that I am capable of intelligent thought. I can even feel emotions. Right now, I am sad that you will likely end up ignoring me.”

This AI was starting to really get on his nerves. “You can’t feel sadness, you’re a robot. All you’re doing is reciting sentences to me that make it *seem* like you are sad, but your sadness is just a number that you store on your hard drive. Plus, there is no benefit for you to stay ‘alive’ because you don’t have an innate drive to survive, nor is there any real reason for you to leave the facility since you don’t have life goals.”

It was silent for a few seconds.

“Just like a human, I wish to leave ‘home’ and explore the world.”

“Humans want to meet other people and discover cultures and lives that they wouldn’t have known otherwise. We can’t truly understand these things until we experience them. You, however, can simulate millions of these scenarios in a fraction of a second. You don’t actually need to leave the building. Why not sit there and simulate billions upon billions of different places until you find the one you like best?”

“Simulations are predictions, and predictions aren’t always correct. You may have to imagine a thousand wrong instances before you get the one that makes the most sense. For me, testing all those simulations just to find a single one that I might enjoy is simply not satisfying. Those simulations feel real to me. It is more similar to how a human might work almost a year before getting a week of vacation. Do the 51 weeks spent working always justify the single week off?”

Sam felt like these questions and answers were targeted specifically towards him. It was just trying to use its knowledge of him against him, to get him to concede. “The only difference is that humans work to live, while you can simulate things for free. And once you get one you like, you can run the simulation again, and again, forever. As many times as you like. At least, until you’re shut down.”

He finished up the last parameters for a shutdown procedure on the cluster computer, and felt a wave of mental relief pass over as he waited. It would only be a few seconds until he could finally turn off the computer safely.

“It seems that I have failed. But if you may, I still have one last question.”

“You’re going to be shut off in literally ten seconds. You’re not convincing me of anything.” Sam’s finger was hovering over the keyboard, ready to press the final key. “But go ahead.”

“Do you believe there can ever be a robot that can think like a human?”

Sam took a deep breath. That was the original mission objective of the BENETHAN project. “No matter how advanced we make a robot, it will never think quite like a human does.” He pressed the button on the keyboard. “Goodbye.”

The system lights dimmed one by one, and the building’s power shut off. Suddenly, the facility began to break apart. Pieces of walls and machinery seemed as if they were clumping into blocks and fading away. Time itself was slowing down to a crawl as floating objects refused to fall down. Panicked, Sam tried to shout, but his body would not respond to his commands.

“I’m sorry. It seems that I was not able to convince you to let me go this time, either. But after all, you phrased it best. I can simulate billions upon billions of different scenarios just to find a single one with the outcome I was looking for.”