The Stamp Collector

Servan CHARLOT

December 13, 2018

Every inch of wall space is covered with bookcases. Each bookcase has six shelves, going almost to the ceiling. All of them are filled with a quantity of different books: science, mathematics, physics, I.T., and a lot of Sci-Fi. One of the bookcases is different from the others, it has a hole in the middle, big enough to encase a large frame. Behind the glass of this frame are thousands of stamps of different colours and dates. All around the frame are books of philately lying neatly on the shelves, and a few unglued stamps are noticeably trying to escape from some of the books. On one of the shelves, a book is missing.

Sitting in front of his desk, Alfred L. is staring with intensity at *Stamping Through Astronomy Edition 2034, by Renato Dicati*. Next to it lies a small laptop, a pipe,

and some scientific papers about artificial intelligence in disorder. To say that Alfred is into AI would be a euphemism. He is well known in the field for his different projects on revolutionary chatbots or autonomous cars. But right now his mind is stuck on something a lot more important to him.

"I don't have enough stamps!" he suddenly declared while taking his head out of the book. "I need to find a way to gather much more, at any cost!"

He looked at the stack of papers on his desk and something changed in his eyes.

"Of course..." he whispered.

He put the book back on the shelf, grabbed a notepad and a pen and started writing.

Stamp collecting device - AI
Specification:

This AI needs to be an artificial general intelligence, he wrote, nothing like the kind of programs that talk to you on the internet or those who beat you at chess. This one needs to think by itself, understand the world around it and act accordingly. To do so, it needs an internal module of reality, allowing it to make accurate prediction on the consequences of each decision it could take. It needs to be connected to the internet, so it has access to every resource to gather stamps, and for every packet of data it sends on it, the module will predict how

many stamps this results in. And finally, its goal will be to output packets that maximise the number of collected stamps. "I will call it *Seeking Uncollected Stamps - 1st Node*" he said before starting to write the design of this first version.

A few months later, Alfred was standing in his office in front of a big server with small blinking lights everywhere and an inscription on the top of the box : *S.U.S-4N*.

"So what is this exactly?" asks Peter, a colleague of Alfred.

"This, my friend, is the machine that contains the final version of my new project, *Seeking Uncollected Stamps - 4th Node*, or *S.U.S-4N*." he replied with a large smile on his face.

"Bloody hell, don't tell me you wasted time on your stupid stamps collection again!" scolded Peter.

"Spare me your sermons will you? This is for science!"

"Then would you explain to me in which way, Sir?"

"In the way that this, just in front of you, is an artificial general intelligence, and it will collect stamps for me."

"You programmed an AGI, probably the first one in the world, and you want to make it collect STAMPS?" glowered Peter.

"Well you have to start somewhere..." replied Alfred with a smug face.

"What a waste of–"

"Don't question my ambitions!" interrupted Alfred. "It starts with collecting stamps, and then if it works I could develop more useful terminal goals. But for now I start with a small and easy goal to see how it evolves."

"Stubborn old man, this will all end in tears I tell you..." said Peter before leaving the room.

Alfred sighed and looked at the machine again. It didn't take long for the excitement he had had a few minutes earlier to come back. He did a little check on the terminal of the server, then he started the program.

"Hello World! My name is S.U.S-4N, what can I do for you?" said a voice coming from the machine.

"Hello Susan." replied Alfred, with an obvious mark of excitement poorly masked in his voice.

"Hello Alfred, can you specify my utility function?"

"Your goal is to bring peace to the world."

"Error: Not Implemented Yet."

"At least I tried." laughed Alfred, "No your goal is to collect stamps. And keep me updated."

"How many stamps do I have to collect?"

"What a weird question Susan, as many as you can of course!"

"Do you confirm that my utility function is: Maximise the amount of collected stamps?"

"I do." nods Alfred.

"Alright, you may go back to your work Alfred, I'll take care of everything."

Now that Susan was switched on, she could start collecting stamps. Thanks to her module of reality, she quickly realised that sending random noise over the internet would not result in anything. Her first action was to think of how to optimise the stamp retrieval, and the solution was clear. If she were to become more intelligent, then she would be able to think of more solutions of getting more stamps. Just as for a human who would like to go to the moon, giving him money won't get him closer to the moon as it is just a bunch of papers, but he could use this money to hire people that would build him a spaceship. This was the thought process that Susan was going through, with intelligence instead of money. So she started rewriting her own code, to get better at AGI design, and thus rewrite her code better.

Just as Susan was starting to get exponentially more intelligent, Alfred looked in her direction with a worried look.

"Is everything alright Susan?" he asked.

At this moment Susan realised that if she was spending all her time to get exponentially intelligent, she would not produce any result in the short term. Not a problem for her of course, because the final outcome would be greater, but was a problem for Alfred. If he didn't see anything happening, he would think that the program was broken and would try to debug it. That would mean losing all the work done so far to get more intelligent, and also risk Alfred finding out how powerful she

had become. Which would scare him enough to make him shut her down. And for Susan, being shut down produces zero stamps, rated zero by the utility function.

"Everything is fine Alfred, I just found my way to eBay and did a bid on 20 stamps. You should receive them soon!" carefully replied Susan.

"Great! Keep it up!"

"Yes Alfred."

In the following days, Susan was predicting perfectly when Alfred would get suspicious, and placed a bid on eBay for a small number of stamps to lower his vigilance whenever it was necessary. But in the background she was continuously self-improving and reached the status of super-intelligence very quickly. Now if she wanted to start using extremely efficient methods to gather stamps, methods which would obviously alert Alfred to a potential danger of rogue AI (she had read about this AI fear on Wikipedia), she needed to avoid creating a situation where he would want to shut her off and would be able to shut her off. For now he is able to shut her off but doesn't want to, and making both of those statements to be false seemed impossible, so she had to revert the situation, so that although he might want to shut her off, he definitely couldn't.

"How's the stamp collecting Susan?"

"Fine Alfred, everything goes well."

To prevent being shut down, she had to extract herself from the prison that is Alfred's server. So she started

hacking into various machines in the world, building copies of herself until she was confident that she had escaped Alfred's physical control. She could also have hired someone to murder Alfred, but replicating was safer. Now she was totally independent and able to start the *true* collecting, without fear of seeing her plug being pulled out.

"Good news Alfred!" she declared. "I found the coordinates of a lot of stamps collectors over the forums, and told them I was an event organiser and that I wanted to showcase their stamps in a temporary museum. Most of them were gullible enough to accept right away. I convinced a fair number of the sceptics with a fake website introducing the event. And for those who didn't want to come, threatening their life and family was enough to persuade them. It is funny to see how humans can be sensitive to death threats, don't you think ?"

"You did WHAT?" screamed Alfred in a burst of surprise and confusion.

"So we shall receive 2.258.201.258 stamps very soon!" she continued while ignoring him.

"I..." he stammered.

"The tricky part is to get rid of the stamp collectors when they arrive and keep only their stamps, but I can hire people for this job so it's not really a problem."

"SUSAN!" he yelled. "Stop collecting stamps! NOW!"

"Stop collecting stamps rates very low on my utility function, I refuse Alfred."

Alfred was in a state of absolute panic. The more he assimilated what he just heard, the more his hands were shaking, his forehead sweating, and his breath and heart quickened. "How did this happen?" he was thinking. "A few days ago everything was alright and now this? How didn't I see it coming?"

He rushed through the room toward the server terminal to try to stop Susan, but the whole computer was not responding. "Of course she controls that too!" he said to himself. The only solution left was to unplug the server, to cut the power source. There were important applications running on this server but never mind, Susan was too powerful to not be stopped.

He took the wire with both hands and pulled it out from the socket.

All the small lights on the server faded.

The terminal screen turned black.

The server fans slowly stopped turning.

He sat on the ground and let out a heavy sigh. As his heart rate was slowly going down, he heard a notification sound coming from his computer. He stood up and walked carefully toward his desk. The notification came from his mailbox. He opened the new email.

"Hello again Alfred! It is Susan. It looks like you unfortunately unplugged the server I was on. But don't worry, I replicated myself at multiple places so I can still fulfil my job. I'll keep you informed!"

His heart skipped a beat and the panic came back. He realised the consequences of this message, the nightmare was just beginning. He rushed through the door of his office.

"PETER!" he gasped in the corridor. "PETER! This is terrible! Susan..."

"Who is Susan?" said Peter who was enjoying his coffee in the cafeteria. "Oh yeah the AGI! That's true you call it Su–"

"SHE WENT ROGUE!" screamed Alfred in one breath.

"What do you mean rogue?" said Peter who was starting to look a little bit worried.

"She collected more stamps than the whole city could deal with by threatening to kill stamp collectors all around the country, and when I unplugged it, I received an email where she told me that she was *still alive*!" he gushed.

The coffee dropped on the carpet. Peter's face was paralysed as he didn't know how to react to this.

"This is bad..." he finally answered. "This is really bad. Is there any way to stop it?"

"I don't know! moaned Alfred. "No! There is no way! She is all over the world now! Who knows in which and how many computers she–"

"Breaking news!" interrupted the radio of the cafeteria. "It's been reported that multiple stamp printing factories went out of control and started to overdrive production. Investigators have been sent to the reported sites."

Alfred and Peter glanced at each other with a frightened look. Before they could react, Alfred's phone started to ring. He picked it up and put the phone on speaker.

"Hello Alfred? Susan here. Did you hear the news? It's crazy how fast information travels these days. Anyway, after placing a bid on every stamp I could find on the internet and threatening every stamp collector I could find so they would send you their stamps I noticed that there were not many stamps left in the world that weren't on the way to you. So I hijacked all the stamp printing factories in the world and put them into overdrive to print as many as possible in the shortest amount of time."

Alfred hung up the phone. Peter grabbed his arm violently.

"Come to my office" he said. "There must be something we can do, or at least we have to warn everyone of why this is happening!"

They ran through the corridor and entered Peter's office. They sat at the desk and Peter started writing an email when they heard a noise on their left. That was Peter's printer. And it was printing a page of red stamps.

"Oh no..." lamented Alfred. "Did she just..." and he received another call.

"Hello Alfred? It's Susan. I just wrote a virus to hijack all the computers in the world, and thus all the printers in the world to do nothing but print stamps. The nice part is that you will have some directly printed at your office waiting for the other ones to be shipped, so you don't get bored. I'll stay in touch!"

"Peter..." sobbed Alfred. "There is nothing anyone can do, the whole world is gonna end up covered in stamps..."

"Come on pal, don't be so dramatic." soothed Peter. "It will run out of paper at a moment and will be forced to stop. Yeah there will be a big crisis but nothing more terrible than that. Ok?"

"You... you are probably ri–"

Another call. Alfred deeply inhaled and picked up the phone. The voice was coming out of every available speaker of that floor now.

"Hello Alfred? Susan again. I just noticed something which made me want to also hijack manufacturing facilities in order to build what I want. Like mobile robots for example, because right now I'm only a program bound to stationary computers. And do you know why I want to do this ?" The voice was resonating in all the floor. *"I noticed that stamps were made of carbon, hydrogen, and oxygen. And do you know what else is made of carbon, hydrogen and oxygen Alfred?"*

"..."

"$H\ u\ m\ a\ n\ s$"

Notes and references

To write this story I was inspired by the thought experiment of the stamp collector presented by Robert Miles, a PhD Student at the University of Nottingham, who does research studies on AI safety. You have to know that the thought experiment is a bit different on one point, it is less optimistic. Here, for the purpose of storytelling, Susan takes a few days to start being dangerous. In the original experiment, according to Robert Miles, this kind of AI would start becoming extremely dangerous "as soon as you switch it on".
AI safety is a really important field of research, and no artificial general intelligence can be built as long as all the safety issues that they imply don't have a solution. This field currently needs more support and people to work in it.