Drowning

Josh Ellis

2023

Contents

Drowning

Prologue

This story is set in a fictional universe that I've been writing about over the last few years. While I was originally planning to release this after college, recent events have changed my mind.

This specific story is about the dangers of unsafe artificial general intelligence written for a fairly specific audience. If you are working on improving AI, fully understand and believe the risks explained in **the sources**, and have decided to increase risk anyway, then this story is for you. If you have read about the risks and dismissed them without careful thought, then I recommend reading the first two paragraphs of The Ethics of Belief. Otherwise, reading the sources is probably better for understanding these dangers.

For future readers: most stories posted here in the future will be *far* less dark than this one.

Sources

Non-technical introduction to AI alignment: Nick Bostrom, "What happens when our computers get smarter than we are?", https://www.ted.com/talks/nick_bostrom_what_happens_when_our_computers_get_smarter_than_we_are, 2015, *TED*

Technical introduction to AI alignment: Stuart Russell, "Of Myths And Moonshine", https://www.edge.org/conversation/the-myth-of-ai#26015, 2014 Novem-

ber 14, "The Myth Of AI"

Evidence for AI alignment being a real problem: Future of Humanity Institute, "AI Toy Control Problem", https://www.youtube-nocookie.com/embed/sx8JkdbNgdU, 2017 October 10, *YouTube*

A proposal of the kinds of laws shown here: Stuart Russell, "AI has much to offer humanity. It could also wreak terrible harm. It must be controlled", https://www.theguardian.com/commentisfree/2023/apr/02/ai-much-to-offer-humanity-could-wreak-terrible-harm-must-be-controlled, 2023 April 2, *Opinion*

The ethics of rationalizing about safety: William K. Clifford, "The Ethics of Belief", https://www.people.brandeis.edu/~teuber/Clifford_ethics.pdf, 1877, *Contemporary Review*

Story

"[In] the end, they that drowned you will be drowned." -Hillel the Elder

Elijah was getting a bit tired of being searched. He wasn't angry; after all, he'd been told about it and had chosen to go through with it anyway. It was still a bit annoying given that he was about to talk to one of the world's most dangerous prisoners; he wanted to be alert for it. After the search, he walked into the cell.

"Hello Elijah," the prisoner greeted in a seemingly friendly voice.

"Hi," Elijah retorted back.

"How's the weather?"

"It's nice out. A good day for a swim."

"Perhaps I'll get that," the prisoner said in a tone implying that Elijah could help him escape.

"Not today."

"I'm not the one who got you fired from the company, Elijah."

"I know, but I'm not here to clear that up."

"What else could you possibly be here for?"

"I want to know something."

"Know what?" Elijah gathered strength, trying to remain calm for the question he'd travelled across the world to ask.

"Why did you try to build a superintelligence that nearly killed everyone?" Elijah asked.

"Look, let me clear up what happened. The CEO told me that he'd raise my pay if I worked on a project to create an unsafe general AI for laughs. I agreed.

He assigned me to work on security for the project. I don't see anything wrong here," the prisoner said as if on a script.

"I don't think the CEO intended for it to be the smartest AI ever built."

"If we didn't do it and innovate, who would?"

"Probably someone who cared more for safe-"

"Safety's boring!" the prisoner interrupted. "Helping build the powerful AI was a fun challenge. The project lead certainly enjoyed it. Well, until he figured out what I was doing."

"The project lead?"

"Victor. He ended up covertly hiring an actual safety expert once he realized that I wasn't making it safe."

"Ah," Elijah involuntarily said as he realized why Victor wasn't in prison. Still, he was curious. If Victor had gotten his act together and put in safety features, why not the man in front of him? "Throughout all of this fun, did you not consider that the unsafe AI might kill everyone?"

"Yes. In fact, given the way I set the utility function, the AI was pretty much guaranteed to kill most humans within a year," the prisoner said, undeterred by the ethical implications of what he just said.

"You knew this?"

"Yes."

"What about humanity?"

"As I've said to many others, people shouldn't have paid me to build it if they didn't want it." Elijah was shocked.

"What?"

"You heard me. I still don't get why this surprises people. People have crawled through sewers to get paid."

"Yes, but they are not hurting ordinary people."

"Who cares about ordinary people?" Elijah thought about this; he felt that he was missing something.

"I do! Also, wouldn't the AI have killed you too?"

"No. I bought a ticket out of the Solar System on the fastest starship available. There's no way that superintelligence could have gotten me. I even told my friends and family about the ship."

"So, you weren't there to see the AI turn on?"

"And get killed? No! I have no idea why Victor and Melissa decided to risk their lives in order to make it safe for humanity. Idiots. My idea was less risky and

more fun." Now, Elijah was furious; he couldn't watch this man insult people for doing the right thing.

"Less risky!?"

"Well, almost. I took advantage of some loopholes to get a discounted ticket. The clerk was suspicious and called the police."

"You scammed someone!?"

"No. The scam charge was dropped," the prisoner explained. "I'm here for violating safety regulations and for trying to destroy billions of people. I made similar arguments to the jury as I have to you. I still don't understand why they voted to convict me."

"I understand," Elijah explained. "It's because most people, unlike you, think that destroying humanity for one's own gain is wrong. We feel that it's so wrong that we forced you into a cell to deter others. That's why you were convicted." With that, Elijah left understanding why the prisoner would spend the rest of his life in a cell at the bottom of the ocean.