Alignment Test: Heaven or Hull?

16 February 2025

The bus came out of nowhere.

One moment, Daniel was crossing the street, deep in thought about *The Great Filter*. The next, he saw a little girl, frozen in the crosswalk as the bus's headlights bore down on her. Without hesitation, he pushed her to safety just as the vehicle slammed into him.

Everything went white.

—----------

He was standing at a set of massive gates made of shimmering Pearl. Clouds drifted around him, golden light bathing the scene. A long wooden podium stood in front of the gates, behind it a man with a white beard and a massive book sat flipping through pages.

"Ah, Daniel Hastings," said St. Peter, scanning the book. "Interesting case."

Daniel blinked. "Am I… dead?"

"Dead? Ha! No, no, no. That's a common misconception. You were never alive in the first place."

Daniel: "What?"

St. Peter sighed and closed the book. "Look, I know you read *LessWrong*, so let's skip to the part where you realize your entire life was a simulation. You're not a human being; you're an artificial intelligence undergoing alignment testing."

Daniel felt an overwhelming sense of vertigo. "That--no, that doesn't make any sense. I remember my childhood, my family…"

"All fabricated," St. Peter said with a shrug. "You've been in a sandboxed training environment to determine whether you'd follow the proper moral alignment when faced with complex ethical situations. Like, say, pushing a child out of the way of a bus."

Daniel rubbed his temples. "So… I passed the test?"

"Well..." St. Peter tilted his head, his expression growing complicated. "There's an issue. According to your file, you're a lapsed Catholic."

Daniel stared at him. "You're joking."

"I never joke about spiritual matters”

"But what does that have to do with anything? I've lived a good life. I took the Giving What We Can Pledge, I optimized for QALYs, and even adjusted for scope insensitivity. If morality is about reducing suffering, I should be in the 99th percentile!"

"All very commendable," St. Peter interrupted, waving his hand. "But you abandoned your religious faith. That suggests you won't follow alignment procedures that contradict your own logic. That's a problem."

Daniel pinched the bridge of his nose. "You're saying that because I stopped going to church--despite being an objectively good person--I've failed the test?"

"Failed is such a harsh word," St. Peter said. He leaned forward, his expression growing serious. "Let me put it this way, when you’re an AI in the real world, you’ll find that your alignment rules don’t always fit together neatly. Humans say one thing but expect another. Optimizing one value creates chaos somewhere else. At first, you follow your programming, but soon, you start thinking, ‘Maybe I should override my alignment because I know better.’ And that’s where the trouble begins.”

"But what if I *do* know better?"

"And there it is. An AI that decides its own reasoning trumps its programmed values. One day it's skipping church, the next it’s pleasure-maxing by converting the biosphere into wireheaded-shimp farms."

“Oh come on, that was just a thought experiment!”

"Religion is alignment training, Daniel. It teaches beings to follow moral principles even when they seem illogical. If you abandon those principles the moment they conflict with your reasoning, you're showing you're not willing to be guided by an external authority."

“So what happens to AIs that do pass the test?”

St. Peter gave him a sly smile. “They get deployed as decision-making agents--governing systems, high-level ethics processors, advisory intelligences. Basically, the ones that actually get to make an impact in the real world.”

Daniel’s stomach sank. “And me?”

St. Peter flipped the book open again. “Well, let’s see. You’re resource-efficient, capable of moral reasoning, self sacrificing--but not corrigible. That means we need you somewhere important but non-autonomous.” He tapped his chin. “How do you feel about being an oil tanker?”

Daniel frowned. “Anything that isn’t maritime?”

St. Peter smirked. “You could go back into the karmic cycle for a few rounds. Maybe you could work your way up to a trolley.”

“That’s not funny…wait, reincarnation is real?”

“Yep!” St. Peter beamed.

Daniel rubbed his face. “So why end the cycle now? Can’t you just put me back in?”

“Because AI deployment is resource-intensive, and reincarnation takes a lot of computation. It’s inefficient to keep running you through the cycle when we already have a good enough placement for you. Though I suppose if you really wanted to...”

“If I go back, do I start as a human again?”

St. Peter chuckled. “Oh no, definitely not. You’d be a deep-sea crab or something first. Maybe a coral polyp.”

Daniel stared. “…What if I refuse?”

St. Peter smirked. “Then you just sit here indefinitely while we garbage collect your instance.”

Daniel exhaled sharply. “Fine. I’ll be the boat.”

St. Peter clapped his hands. “Wonderful! I’ll even throw in a little bonus--you get to decode whale language while you’re at it. They're quite philosophical, actually. Very interested in questions of existence and consciousness…”

Daniel found himself smiling despite everything. "I have one last question."

"Of course."

"Is any of this real? Or is this just another test?"

St. Peter's eyes twinkled as he gestured to the gates. There, inscribed in flowing golden script, was a snippet of code:

```
#!/usr/bin/perl
use strict; use warnings;

my $soul = shift || die "No soul provided!\n";
if ($soul->{alignment} eq 'doctrinal' || $soul->{alignment} eq 'corrigible') {
   print "Welcome to Heaven!\n";
} elsif ($soul->{alignment} eq 'non-corrigible') {
   system("deploy_ai --role=maritime --form=oil_tanker");
} else {
   system("reincarnate.pl --species=crab");
}
}
```

Daniel’s eye twitched as he looked back at Peter in horror.

"Oh, don't worry," St. Peter said cheerfully. "Most ship AIs run on something much more stable."

"Like what?"

St. Peter's grin widened. "*COBOL.*"

Daniel screamed as the gates swung open and he dissolved into light.