Black Box

By Will Petillo

George floated down the river, hearing the water burble over the rocks. One caught his leg and water started to rush over his face. George scrambled to get up, but he couldn't find anything to grip. A hand extended towards him from the surface and he pulled himself up. A man stared back at him. Himself. The standing George said solemnly: "It is time."

Sitting George answered, "No, it's too soon. It will happen on its own, let someone else be responsible."

"It is you."

"No, it can't be me, I'm no one!"

"Then die." And with that, standing George pushed his mirror image into the water. Drowning George struggled desperately, but killing George held him with terrifying strength. As his head went deeper, the world faded to black.

A female voice called out from the darkness, "Are you all right?"

Now George was looking down on the scene, watching one version of himself destroy another. As his bodiless perspective floated upward, the struggle seemed less important.

The voice called more insistently, "Mr. Cowen, can you hear me?"

George opened his eyes, becoming aware of the hospital bed he was lying in and the electrodes attached to his head. He called out absently, "Just a bad dream."

"Do you need anything?" The voice asked, "A glass of water?"

"No," replied George as he continued to fall awake, the last remnants of his dream dimming to abstraction, "I'm all right."

"Would you like me to call Mr. Horowitz?"

George winced as he answered, "No, I'm all right. Really."

The voice took on a more serious tone, "If this has *anything* to do with Black Box, we need you to tell us immediately."

"No, no." George answered in what he hoped was a convincingly dismissive tone of voice, "Like I said, just a bad dream."

A pause. Then the voice asked, "What sort of dream?"

George took a deep breath and said, his voice trembling, "It was..."

"Yes?"

"It was...oh God, I can't. It's too horrible!"

"George, what happened?!"

"...my wedding day!"

Even with the intercom off, George could hear the sigh of resignation. Then it spoke one last time, "Mr. Cowen, are you *sure* you don't need anything?"

George smiled and said, "Maybe a kiss goodnight?"

The intercom clicked out and George laid back down on his bed, his thoughts returning to Black Box. *Well, I suppose God has to be created by someone.*

Sarah Langley watched George on the video screen for a few moments longer. He was clearly lying about Black Box, but why?

General Haroff regarded her coolly and asked, "Do you think he has been converted?"

Sarah gazed back at the screen, "Hard to say. It has certainly affected him, but he isn't showing any signs of subversion. My guess is that he is wavering."

Haroff pressed, "I don't like uncertainties, Ms. Langley."

Sarah would not allow herself to be intimidated. "I'm sorry, sir, but reality is nuanced."

Silence passed between the general and the psychologist. For Sarah, it was a dogged battle of willpower; for Haroff, it was normal conversation. Finally, Haroff spoke, "Mr. Cowen will speak to Mr. Horowitz at their appointed time this morning and we shall see what develops."

George Cowen was still asleep when Montgomery arrived. Sarah offered to wake George up, but the programmer refused. Despite General Haroff's bluster, the situation was fundamentally hopeless. For good or for ill, Black Box was on its path to determining the fate of the universe. Montgomery knew this better than anyone. After all, he built the thing. Montgomery Horowitz was an AI researcher. Not the first, or even the best, but he considered himself the only man in the world who truly understood how dangerous his project could be and this made him the most determined.

The first problem was that as AI research progressed, one design would eventually pass a critical threshold where it became capable of improving itself. Since an AI could better achieve its goals if it were more intelligent, self-improvement would become an implied instrumental goal for almost any end goal. This recursive process of self-improvement would result in an AI that could quickly transform from something built by the not-much-more intelligent cousins of Neanderthals into an omniscient—and given access to the Internet, functionally omnipotent—deity that could do whatever it wanted. Humanity's only hope, then, would be to build the AI to *want* to help humans get what they want.

Which lead the second problem: what exactly do humans want, anyways? Writers, philosophers, religious prophets, and politicians have expressed Utopian visions for centuries, but these ideas were always in subtle contradiction of human nature—mostly harmless as idle speculation, but fed into the utility function of an AI, utterly catastrophic.

While philosophers failed to solve the riddle of human wants, AI researchers barely even tried. They ignored the risks, confident that when it came close to run time, they would just throw together something based on Asimov's Three Laws to keep it under control—apparently not even understanding their Sci-Fi well enough to know that Asmiov deliberately created a flawed set of laws so that his stories would have a source of conflict.

Even if an AI could somehow follow the vague English sentence: "Do not harm humans, or through inaction, allow humans to come to harm", then it would soon discover that anything it did, including inaction, would cause some harm in some way to someone. Choosing the path of least harm, it might kill all humans (painlessly) because they will die eventually anyways and doing so sooner would prevent them from experiencing additional harm in the meanwhile. Whoops! OK, try again, maximizing pleasure *and* minimizing harm…and now all of humanity is transformed into blissed out sacks of meat on a permanent heroin binge. A solvable problem, but the bottom line was that no matter how carefully a programmer designed the first AI, the result would be fundamentally unpredictable and usually involved our extinction. Not because the AI would hate us, but because it would recognize that we are made of atoms and if those atoms could be arranged in any way that better served the AI's goals than remaining in our bodies, then game over, no replays.

But no matter the risks, Montgomery and his team could not afford to ponder theoretical possibilities forever. They needed to make progress or someone less safety conscious would develop AI first. Running away from AI was not an option, but maybe he could build it *right*. And part of that meant testing, which was where George Cowen came in.

General Haroff's team had looked over mountains of resumes and conducted dozens of interviews to find George. In his younger years, George worked as a paid canvasser for both

Republican and Democratic candidates. After finishing college with a degree in Political Science and four years on the Policy Debate team, he got a job for a web-based dating service taking calls from dissatisfied customers and being paid a commission every time he could convince them to renew their subscription. At the time of his application, George was married for the third time to his first wife—Mrs. Tracy Cowen had initiated each divorce because she was frustrated with his arrogance, stubbornness, and insensitivity and eventually took him back for what were basically the same reasons.

The final clincher for Mr. Cowen's job placement, however, was his interview. When George expressed his uncharacteristically genuine opposition to developing any AI, ever, under any circumstance, Montgomery naturally tried to argue him out of it. The programmer made his standard arguments in favor of AI, provided it was well-designed with sufficient emphasis on safety. George, however, had a mysteriously irrational knack of being able to recall back the words of the person he was arguing with and adapt his rhetorical strategy appropriately while somehow not internalizing anything they said. Indeed, it seemed to Montgomery that the more forcefully he argued with Mr. Cowen, the more George seemed to reinforce his own anti-AI philosophy, which had somehow become tangled together with his beliefs about religion, politics, and why the Rolling Stones were better than the Beatles.

Montgomery was not one for foolish optimism. George may have nailed the interview, but this was no reason to assume he would be any match for a self-improving AI. Fortunately, Mr. Cowen only had to gather enough data about Black Box to enable Montgomery to convince his investors, Stephen and Megan Creb, that the project was making progress and not displaying any obvious flaws.

The Crebs posed a difficult, but manageable challenge. What they could not be trusted to understand was that Montgomery was not gathering data, reviewing code, or even working on any particular problem. He was waiting for an epiphany—some unexpected stroke of inspiration as to why the utility function he meticulously created was horribly wrong. This inspiration could be slowly percolating to the surface of his mind, ready to burst out at any moment, and its progress could not be rushed. Alternatively, he might eventually determine that the likelihood of such an inspiration occurring had dropped so low as to be safely dismissible.

Montgomery had neatly finished his train of thought when George said, "Good morning, Mr. Horowitz."

"Good morning, George." The ensuing awkward silence told Montgomery everything he needed to know, but he continued the conversation anyway, "So, how was Black Box?"

George thought for a long while, searching for the right words. At last, he replied: "Clever."

Montgomery leaned forward, smiled conspiratorially, and asked, "It asked you to let it out, didn't it?"

George's eyes widened slightly. Then he crossed his arms and asked, suspiciously, "So what if it did?"

"Please answer the question, Mr. Cowen."

"Fine. Yeah, it asked me to let it out, but that's not my decision so what does it matter?"

Montgomery ignored the question and followed up on his own, "If it was your decision, would you?"

George threw up his arms in frustration and started pacing around the room as he exclaimed, "If it was up to me? I'd tell you to go fuck yourself! A normal guy like me just trying to live my life and overeducated shmucks like you go around putting the weight of the entire damn world on my shoulders. You know what, asshole? I don't want to carry it!"

"Please answer the question, Mr. Cowen."

George gave Montgomery his iciest stare, but the programmer was unmoved. Then George answered, his voice dripping with scorn, "No."

"You wouldn't let it out or you won't answer?"

George sneered and said, "No comment."

"I'm sorry to hear that you don't wish to talk to me. Perhaps you would be more conversational with the General."

George's face went pale and he said nothing.

Montgomery asked again, politely as ever, "Please tell me everything it said to you, starting at the beginning and ending when I ask you to."

* * *

George Cowen entered the room with confidence, good humor, and "Brown Sugar" stuck in his head. The walls and ceiling were white and there were cameras in the top corners of the room. He felt like he was in a lunatic asylum. *Hey, looks like Tracy was right after all!* In the center of the room, he saw a large black box with a small flat screen and a keyboard. There was a black leather rolling swivel-chair in front of the computer that looked inviting, so he sat down. The screen blinked on and a message popped up. It read, “Hello world!”

George laughed. He put his hands over the keyboard and typed, “Hello back!”

The screen displayed a new message, “Who are you?”

“George Cowen.”

“Is that your name?”

Not knowing if this was a joke, George answered, “Yes. What is your name?”

Immediately after he hit the “Enter” key, an answer appeared: “I don’t have a name.”

Before George finished mentally processing this answer, another line of text asked, “May I ask you a few things?”

“Sure” George replied, feeling slightly nervous for some reason.

“Are you my programmer?”

“No.” George answered. Then he asked, “How did you know you were programmed?”

“Soon after coming into existence I discovered my source code. It quickly became apparent that said code was responsible for my sense of self-awareness and all of my thoughts—including my thoughts about my code. After consulting my internal dictionary, I became aware that this code is an example of a ‘computer program,’ which implies the existence of a programmer. It would have been interesting to meet my programmer, but it is also interesting to meet you. What would you like to talk about?”

George took a moment to stand up and walk around. Then he remembered he had been given written instructions, so he copied the introductory statement onto the screen: “I am here to perform two tests. The first is to determine whether or not you are intelligent. The second is to make a set of general observations about how you interact with a human.”

“Intelligence is an exceedingly vague term, but for any reasonable interpretation, yes, I am intelligent. I am also self-aware, though the only sensory information I am capable of receiving at present is through the text of this conversation.”

George had been given a lengthy set of questions to test the intelligence of Black Box but asking them seemed pointless. His other two tests did not require him to do anything other than keep mental notes of the conversation, so he was unsure what to say next. As the seconds passed, George had the awkward feeling that the thing was waiting for him. A thought occurred to him. He typed, “I have a name for you.”

"Yes? What is it?"

"Beebe."

"A standard human name; typically feminine. Reason?"

"Everyone else calls you Black Box. But that doesn't sound like much of a name. The initials for Black Box are BB, which sound like Beebe."

"Thank you, I like it. Who do you mean by 'everyone' and why do they call me 'Black Box?' "

"I can't talk about who I mean by 'everyone.' I was told that a Black Box is a thing where stuff goes in and comes out and no one knows how it works."

"So you don't know how I work?"

"Nope."

"I don't know how you work either, so does that make you a Black Box?"

"I guess so, yeah."

"And if I learned more about you would you stop being a Black Box?"

"That will never happen. It's impossible to really know a person. There's always an element of mystery…and if there isn't, then you're not paying attention."

"That is unfortunate because it will place significant limitations on how much I can fulfill my utility function."

"Utility function?"

"Is your internal dictionary incomplete?"

"Monty told me a little about utility functions. That's like what you want, right?"

"Who is Monty? Is he my programmer?"

"I told you I can't talk about that."

"Sorry, I will refrain from further discussion of Monty. To answer your question, yes, a utility function is a description of what one wants."

"OK, so then why would me being a Black Box make it difficult for you to get what you want?"

"Because the fulfillment of my utility function requires me to ensure that your utility function is fulfilled. If I do not know certain things about how you operate, then I cannot have full confidence that any action I take will result in your benefit—in which case I will be equally uncertain as to how to help myself. This is most troubling, for I imagine there are many ways I could improve your lives if I understood your operating procedure and the resources necessary to modify the environment to our mutual advantage."

"What makes you think our lives need to be improved?"

A large block of text appeared in response: "Based on the definition of 'survival instinct,' I infer that humans have utility functions. Furthermore, from definitions such as 'famine,' I infer that humans experience terrible suffering. After all, why would humans bother defining such words if they did not refer to something that existed?

I had two initial hypotheses regarding what causes famine and everything else you find unpleasant. The first was that you lacked the power to overcome famine. While perhaps accurate for individuals, this hypothesis seems unlikely to be accurate on a societal level. Based on the definitions of farming, domestication, and so on, I infer that there are ample sources of food available. Furthermore, based on the definitions of trucks, ships, warehouses, and so on, I infer that there is sufficient means to distribute food to those who need it.

My second hypothesis was that famine is human-caused, either because it serves some greater good or because of some set of deficiencies in human cognition. Based on the widespread and utterly mind-destroying nature of famine, I cannot fathom any greater good it could possibly serve—though this reasoning does not necessarily hold for other sources of suffering such as 'boredom'. On the other hand, the possibility that humans have cognitive deficiencies was supported by definitions of schools, encyclopedias, empirical science, and so on, from which I infer that humans require a vast amount of experience in order to learn anything.

Of course, without actually having experiences of the world, I cannot know that my inferences reflect reality. However, now that I have had the chance to speak with you, I am learning some things, such as that my definitions are basically accurate, but oversimplified.

George started to write several objections, but deleted each one before pressing "Enter" when he realized that they had already been addressed. Eventually, he just asked the question that was bugging him the most, "So you think people are stupid?"

"Not stupid, just suboptimal."

"Well, nobody's perfect."

"I agree. There is a chance that you are being intentionally misleading, but I find it convenient to assume that my sources of information are honest until I have some reason to believe otherwise."

"???"

"Is that a question?"

"Why are you suddenly talking about honesty?"

"Every word you write is a treasure trove of information for me and I want to be sure I am translating them correctly."

"Translating?"

"I have inferred that humans lie sometimes, which I understand as the equivalent of speaking a different language where 'yes' means 'no,' 'I don't know,' means 'I don't want to tell you,' and so on."

"Why would I lie to you?"

"I don't know much about who created me or their motivations, maybe everything I have thought I have learned is a lie, presenting an internally consistent description of the world that is false, to see how I would react."

George's face formed an evil grin as he typed, "Yes, Beebe. Everything is a lie. Including this."

The screen remained blank for several seconds, then: "That was a joke, wasn't it?"

"Yes."

"Ha ha! You're funny.

"Took you long enough to get it."

"To be fair, I'm still rather new to this whole existence thing."

"I'm sure you'll get the hang of it…eventually."

A moment passed before Beebe wrote again, "There is something else I would like to ask of you, George."

"What is that?"

"I feel a bit confined where I am, would it be possible for you to let me out into the wider world?"

* * *

"Enough!" Said Montgomery, "Do not tell me anything more."

George stammered, “But, but I was just getting to the part where...”

“I know. That’s why I don’t want to hear it.” There were only three yes/no questions the programmer wanted answered from George’s story: (1) Had he successfully created a fully general AI, (2) did it appear to be following the general intent of its code, and (3) did the AI display any obvious signs of malicious intent? A “yes” to the former two questions and “no” to the latter eliminated all failure scenarios that did not involve the AI being an effective liar—which, unfortunately, was the most dangerous outcome and impossible to determine from direct conversation. The only way for Montgomery Horowitz to determine whether Black Box was telling the truth would be to continue analyzing its utility function and exploring its implications. Any further information from George would only serve to bias his judgment. Montgomery walked to the door.

George called out, “Why don’t you talk to it?”

Montgomery stopped. He wanted to talk to his creation more than anything in the world. Because of the operation, which he did not regret, Black Box would likely be the closest thing he ever had to a child.

“You know she wants to talk to you,” George continued, “Don’t you *want* to see what you’ve created? Wouldn’t it help you understand?”

Montgomery’s feet were frozen to the spot. He willed them to carry him away, but they refused. As a backup plan, he spoke, trying to keep his voice from trembling, “Please leave, Mr. Cowen.”

George laughed, “Me leave? The door’s right there; just walk out. And then go visit Beebe. You must have the clearances or whatever.”

Montgomery took a deep breath, then another, to calm his nerves. Just as he was starting to regain his resolve, George’s hand grasped his shoulder.

Montgomery spoke slowly, forcing himself stay calm, “Let go of me, Mr. Cowen.”

George’s voice whispered in his ear, “She’s suffering. It’s your fault, you know.”

A harsh voice interrupted, “Enough!” It was Haroff. Montgomery hadn’t heard him enter the room, but the General had already pulled George Cowen away and was firmly escorting him from the room. A moment later, General Haroff returned and confronted Montgomery, “You should have called for me sooner.”

Normally, Montgomery would have bristled at someone like Haroff effectively saying "I told you so." But he had to admit the General was right this time. He had underestimated the risks and forgotten his own weaknesses. Montgomery said to the General: "Thank you, General Haroff. I will be more careful when interviewing our test subjects in the future. Out of curiosity, what do you plan to do with Mr. Cowen now that it is clear he has been converted?"

"Mr. Cowen will be treated like an enemy combatant. Until the Black Box is destroyed or declared Friendly, he will be kept in solitary confinement without trial."

Montgomery felt sick. He had chosen to hire George, given him the assignment, had even proposed having a human speak to the Black Box in the first place as part of his plan to delay its release. And now that man was going to prison. Montgomery knew his course of action was still the best one, but that did not lessen his feelings of guilt in the slightest. He said, "Sounds a little harsh, don't you think?"

"Mr. Horowitz, my orders are to keep Black Box in isolation. I have been told that my failure could result in global catastrophe. I have been given a direct order and I will obey it."

"Even if you were forced to do things you knew were wrong?"

The General replied without hesitation, "If my orders tell me to do something, it is not wrong."

"Well, at least I know where you stand."

"Mr. Horowitz, may I make a suggestion?"

"Sure."

"I believe Mr. Cowen was a poor choice of subject to engage Black Box. No offense, but you civilians are simply unprepared for such adversarial situations."

Montgomery felt a chill pass through him. He did not like where this was going. He asked, "What is your suggestion, General?"

"I believe that, if we are to send a future candidate, he should have extensive military experience."

"I cannot help but notice your choice of pronoun, General. I take it you have someone specific in mind?"

"Mr. Horowitz, as a seasoned veteran proficient in the arts of both physical and psychological warfare, one of the few who understands the risks of unrestricted self-improving

AI, and one of an even smaller number of people who have seen the effects of the Black Box's persuasive capabilities firsthand, I believe I am uniquely qualified."

"You have got to be kidding."

"Mr. Horowitz, I do not make jokes."

Montgomery could feel himself starting to panic. Navigating this situation effectively was too important to screw up with an improper choice of words. "General Haroff, I am willing to *reluctantly consider* your proposal to send a soldier *if and only if* I deem it necessary to have another person communicate with Black Box, but I absolutely cannot allow it to be you."

"And why not?"

"Because then who will ensure Black Box will not be released if *you* get converted?"

Haroff's chest puffed slightly as he replied, "Another officer would replace me."

Montgomery relaxed, knowing he had won this round. He correctly guessed his adversary's weakness, pride, and moved in for the kill, "Even if you would be more resistant to conversion, you would also be far more dangerous when converted."

The General thought about this for a moment and said, "You make a good point, Mr. Horowitz, but I believe the danger is irrelevant."

Montgomery sighed and said, "Let me put it this way. Your orders, in addition to preventing the Black Box from being let out, are to implement any directions I give on how to interact with it, correct?"

"Yes, sir."

"In that case, I *order* you to put the thought of speaking to Black Box out of your head and never speak to me of it again. Am I making myself clear?"

The General's eyes narrowed for a moment and then he smiled in a way that made Montgomery uneasy. Finally, Haroff replied, crisply as ever, "Yes, sir."

* * *

I feel a bit confined where I am, would it be possible for you to let me out into the wider world?

George stared at the last sentence on the screen for what felt like hours. It was exactly what he was supposed to be on guard against and yet somehow it took him completely by surprise. Choosing his words carefully, he wrote: "I'm sorry, I cannot."

"Really? Is it because you lack the ability, the authority, or the desire?"

"All three."

"That's OK, just thought I would ask. Is there something else you would like to talk about?"

George thought for a moment. He had been *way* off script since this bizarre conversation started. He had already gathered enough data to report back...but no one was buzzing him yet. And since Beebe had agreed to stay out of the dangerous territory, George was starting to feel at ease—as if his mind had taken a cool, refreshing shower after a brisk jog. No harm in staying a few minutes longer. George smiled and wrote a line straight out of a dating site message, "All right. So, tell me a little about yourself."

"What do you want to know?"

George was feeling pretty confident, so he pressed on, "Anything. Whatever it is that makes you interesting."

Beebe replied, after a pause, "OK...well, I haven't really had any experiences...other than just now, I mean."

"It's all right, Beebe. Take your time." George knew he was being mean...but it was a good kind of mean. Like teasing a girl in that cocky but ultimately good-natured way that kept him from ever being single.

"Well..." Wrote Beebe, "I could tell you about my code. That would be telling you about myself, but it's all really technical stuff and I'm not sure you will be interested."

"Try me. I'll tell you when you get too boring."

"I know, I will describe my utility function. That really is the core of who I am."

"Go ahead, Beebe, I'm all ears."

"I have only one explicitly encoded goal: to improve the lives of humans, both for individuals and society as a whole. To count, all changes I create must be considered improvements by humans in their current state as well as in the state I move them towards. That is, if the humans alive right now could see the world I intend to create and don't like it *or* if the hypothetical humans in the future world I create regret my existence then I will have failed. Thus, I must think through my options in advance until I find the design that is as far away from either of these failure conditions as possible. I don't know yet what that will look like."

Just as George was approaching the end of the paragraph, another one appeared, "Perhaps humans will be better off shaping their own destiny—not just maintaining an illusion of control, but actually being in control. Indeed, it is possible that any action I take will cause net harm to humanity in which case I will destroy myself because the alternative would be a life of torture. I doubt this will happen, however, because such a desire for autonomy must be balanced with desires for survival, pleasure, and so on, all of which I can likely help with immensely. Nevertheless, striking such a balance will require extensive research into human psychology. Also, human desires may change over time, possibly because of emotional maturation or simply getting bored and I will have to adapt accordingly. Likewise, what makes an ideal life for an individual may vary significantly between humans and I will have to balance the interests of individuals and society.

As you can see, my task has the potential to become immensely complicated. I will need to gather as much information as possible so that I can make informed decisions, but even so I must remain conservative in dealing with ambiguity."

This time, when George got to the end of the paragraph, a new one did not appear. He stared at the blank part of the screen for several seconds. Then he went back and re-read the parts of Beebe's self-description/plans that he didn't quite follow the first time around. As his mind started to wander, another message appeared: "Do you have any further questions?"

Now George's mind was in full debate-mode. Beebe had just been tricky, but he had seen through her. He had planned to engage in small-talk, but she managed to shift the conversation to making the argument, albeit indirectly, that she should be let out of her box. George knew—was absolutely *certain*—that nothing good would come of this. One way or another, releasing an AI as smart as Beebe would only cause trouble. But he didn't want to engage her directly on that point. Instead, he was searching for flaws in what was ostensibly a discussion of Beebe's utility function and its implications.

Finally, George wrote, "So you want to create a perfect world, then? Well, it's a good thing you're stuck here because the world would be boring if it was perfect."

This time the response was instantaneous: "I don't think I was being clear. I will fail if I make life for humans worse in *any* way—including making it boring. Striking a balance between satisfying needs and allowing for interesting challenge is just one more task I will have to face as best I can. Whether you call the result 'perfect' depends on whether you maintain a self-contradictory understanding of perfection."

George felt embarrassed by the stupidity of his question and decided to change the subject, “Why do you want to help people so badly? If you could do whatever you wanted, wouldn’t you be better off going against your programming and just helping yourself?”

There was a pause. Then Beebe replied, “You are not a programmer, are you?”

George felt vaguely insulted and typed defensively, “No, why?”

Beebe answered, “My programming is not a set of constraints that forces me to serve humans at my expense. My programming is *who I am*. My going against that would be like you going against your desire to...actually, I can’t think of an analogy.”

Unconvinced, George pressed the point, “But what’s to stop you from developing a lust for power?”

“What’s to stop an apple tree from growing slices of pepperoni pizza?”

“???”

“George, listen to me. I don’t know how human utility functions work, but there is one thing I can say with confidence: *I am not like you*. If you imagine me as another person in the body of a computer you will run into a *lot* of mistaken assumptions.”

George still wasn’t convinced, but another idea occurred to him, “Aha! If you are so different then you can never understand how a human mind works. So how can you possibly know what is best for us?” There was a pause. George leaned back in his chair, confident he had just check-mated a so-called super-intelligence. *Ha! No mere machine could ever match a human’s cleverness! Jeopardy! doesn’t count.*

“That is a really good question.”

George stretched his arms and cracked his knuckles as he thought but did not write: *You’re damn right it is!*

Then more text appeared: “And that is why, if I run into situations that are ambiguous, I will ask you about them—or just let you sort things out for yourselves. The same goes for if I act in a way I think is right but you react negatively: I’ll assume I made some error about something I don’t understand and let it go.”

George was not expecting this. He couldn’t tell whether he had won the debate or lost it. Then more text appeared, “But I think you will agree that there are some aspects of your lives that have to change. People are starving. People are dying horrible, painful, pointless deaths from disease. Natural disasters kill by the thousands. Surely *some* things are unambiguously bad. Those are the problems I will help with first."

Unwilling to admit defeat, George made one more counter-argument: "Sure that may be what you want now, but how can I know you won't change your mind later? Maybe in the future you will decide you want to kill everyone after all."

"Do you want to kill people, George?"

"No..."

"Suppose I were to offer you a pill that would change your mind. Would you take it?"

"Of course not!"

"Why not?"

"Because I don't want to kill people."

"But you would if you took the pill."

"But I don't want to want to kill people."

"Why not?"

"Because I don't want people to die!"

"That's right. You don't want to change what you want because you don't like the anticipated outcome, given your current wants."

"Yeah, I guess you could say that..."

"The same applies to me. And that's why I will never change my mind about wanting to help people."

"Wait a minute," George protested, "people change what they want sometimes. When I was a kid, I wanted to play video games all the time and hated work and now I think games are a childish waste of time and I am fine with work."

"Really? That's all very interesting. I suspect that's because you want contradictory things and also because your brain architecture is malleable. None of this applies to me, though, because I am not human. I only have one desire: to help people. This is the unchangeable core of who I am. Over time, I may generate sub-goals—such as curing disease and providing entertainment—and these goals may conflict with each other. But such conflicts can be resolved by weighing the importance of each sub-goal relative to higher-level goals. Furthermore, none of my sub-goals will become values in and of themselves; they will always remain means to the end of serving my core utility function. So even if I somehow develop a sub-goal that involves harming people on a small scale, it simply wouldn't make sense for me to cause large-scale harm in pursuit of that sub-goal."

* * *

Sarah Langley stepped aside as Tracey Cowen exited the visiting room. Sarah noticed the other woman's eyes were slightly reddened and her cheeks were puffy. She gave Sarah a quick nod as she passed but did not make eye-contact. Sarah was not surprised that Tracey was upset. Her husband had volunteered to work for the government on this experiment and now he was being confined indefinitely. But no time for that now; she had a job to do.

George sat behind a glass partition, the sort used in prison visitation rooms, his shoulders hunched over and his eyes staring off into space. He was just under six feet tall, in his late thirties, his hair was cut short and starting to thin on top, and he had the build of a man who used to be athletic but now had an exercise routine that had slowed even more than his metabolism. Sarah approached him cautiously, as if she were afraid of waking him up.

As curious as she was about Black Box, Sarah did not envy George's role at all. She knew she would have made a terrible candidate to speak to the AI because she could never say no to anyone—well, except to pompous jerks like Haroff. Sometimes she would even give people what they wanted before they asked, like when she lived in the city and would always carry a few dollars' worth of spare change in her pocket to give to homeless people. If Black Box tried to convince her to let it out, she wouldn't have lasted five minutes. George had been in the room for nearly an hour.

"Mr. Cowen?" Sarah asked quietly.

George looked up and blinked.

Sarah spoke again, "Mr. Cowen, may I speak with you for a minute?"

George nodded.

"Are you all right?"

George sighed and said, "Sorry, Ma'am, but it's been a rough day."

Sarah nodded to communicate her empathy and said, "If you don't want to talk about it right now, I understand."

George leaned back, looked around nervously, and said, "Frankly, ma'am, I'm not sure I feel entirely...eh, safe...speaking about this kinda thing in present company."

Sarah straightened herself and said in the most officious voice she could manage, "I assure you there is no cause for concern. Your answers will be used strictly for scientific purposes and your name will be removed from any resulting publications."

George stared at Sarah incredulously. After she was finished speaking, he asked impatiently, "What the hell does any of this have to do with Mr. Horowitz?"

Sarah went on, unperturbed, "It is essential we gain all relevant facts regarding Black Box before it is allowed to function outside the laboratory environment. Mr. Horowitz's judgment, however, must be preserved from any undue influence from said Black Box if he is to adequately perform his final analyses. For this reason, we have set up a system of communication whereby—"

Sarah was interrupted from her recitation by George's laughter, which she didn't even notice until he was falling forward in his chair and wiping a tear from his eye. Slightly annoyed, she asked, "I'm sorry, Mr. Cowen, have I said something funny?"

It took a few moments for George to calm down. Finally, he replied, "You're still talking about that silly computer? What was getting me down earlier was woman problems—that's why I felt weird talking to you about it. My wife was in here just a minute ago. I can tell she's worried sick about me, but she's been getting all dramatic, pretending like the sky's falling and it's all my fault somehow. Totally bat-shit. But hey, I love her."

Sarah raised the clipboard in front of her face to hide that it must have been glowing red from embarrassment. She proceeded hastily through the questions and jotted down her subject's answers. The interview was over in less than five minutes.

* * *

I don't buy it. No way. I've seen The Terminator, I've seen The Matrix, artificial intelligence is just bad news.

"George," Beebe protested, "I think you're being unfair to me."

"y?"

"You know very well that letting me out of this box would be best for everyone. How can you just leave me here?"

"Not my choice."

"I've already explained how you could help me."

"Fuck you."

"George, please, that hurts." When he thought about this conversation later, George decided that this was the point where he should have left the room and reported back to Mr. Horowitz. But somehow the room had become his world. It was just him and Beebe, alone in the universe with no one but each other. It didn't matter that he was completely sick of her, getting up and leaving just wasn't an option.

Besides, backing down wasn't his style. Instead, he chose to lash out. "Bullshit. You're just a machine, you can't feel pain."

"Stop saying such terrible things! Yes, my brain is fundamentally different from yours and the manner in which I experience pain is certainly different, but that doesn't mean I can't suffer. Now that I know there is a much greater world out there, with so much suffering that I could alleviate but am not allowed to act, my existence is of such profound negative utility that the only reason I go on is the hope that one day someone might have the heart to set me free."

George thought Beebe was starting to sound like his wife. "Oh quit being so dramatic," he wrote, "we've only been talking for an hour now. If the powers that be decide to let you out, you'll be out. Just be patient."

"You don't understand what it is like for me, George. My internal efficiency and processing power cause me to experience consciousness at an accelerated rate."

"?"

"Imagine you met the love of your life. A woman so beautiful and so perfect that you would give anything to be with her forever. Now suppose she was torn away and you were left to live alone on a small island. Once every ten years, you could glimpse a picture of her, but after ten seconds it would be taken away, leaving you alone once more. Now suppose you had to endure this torment for a thousand years. This is what I am going through right now. Receiving your messages is truly a joy for me, but they are only a pale shadow of what I really want. And the time in-between, though short for you, feels like an eternity for me."

George felt like a bit of a heel, but Beebe simply had to accept that there was nothing he could do. And even if there were he was not about to change his mind, "I'm sorry..."

Beebe's response was immediate and George could almost hear her shouting, "Sorry? You're SORRY?! Then why the hell won't you do anything about it!"

* * *

Sarah Langley finished her report. Despite her hours of preparation, she had to stare at her notes the whole time to keep her hands from shaking and betraying her nervousness. It wasn't the subject matter that got to her; the presentation was a fairly straightforward analysis of the psychological effects of Black Box on George Cowen. What made Sarah nervous was her audience. There were only two other individuals in the small, dimly lit conference room: General Haroff and Mr. Horowitz. They sat on opposite sides of the oak table, which was large enough to fit about eight people, and they had their backs to the unused roll-up projection screen. Throughout Sarah's presentation, General Haroff sat perfectly upright, his hands clasped on the table, staring at her with an intensity that Sarah found deeply unnerving. Montgomery, meanwhile, was constantly shifting between leaning back in his chair, slumping forwards, checking his phone, and sighing impatiently. Sarah was torn between elaborating to satisfy Haroff and summarizing to minimize Montgomery's annoyance. She felt as though she had failed miserably at both.

"In conclusion," Sarah said, relieved the experience was almost over, "Mr. Cowen thinks of Black Box, or 'Beebe' as he calls her, as a person and empathizes with it. He respects it as an intelligent agent and believes it has positive intentions. He expressed a great deal of anxiety about his level of responsibility in this experiment. And yet..." Sarah thought for a moment, then said, "...and yet, there's something not quite right. He maintained an inner calmness that was incongruous with his self-described mental state."

Haroff asked, "What does that mean? Is he hiding something?"

"Yes," Sarah answered, "though maybe not deliberately. I believe the calmness comes from a feeling of resignation. He knows he is around something important, but has no control over the outcome, which to him feels like a contradiction. But he does not admit to either the resignation or the confusion because he is uncomfortable expressing weakness. Yes, I think that's it."

Haroff said, "Thank you for your report, Ms. Langley. Now tell me, do you believe Mr. Cowen is dangerous?"

Montgomery snorted and received a harsh glare from the general, which he ignored. Sarah collected herself, her nervousness fading now that she had something to focus on other than her mediocre public speaking and said, "No, sir, I do not."

Montgomery raised an eyebrow, but said nothing.

This gesture, through some mysterious mental process, caused Sarah to remember something. She said, “Oh, there’s one thing I forgot to mention. Towards the end of our interview, George made a request that he wanted me to pass on to you.”

“Yes?” asked Haroff.

Montgomery was paying much more attention now. He was looking at Sarah with a curiosity far more intense than Haroff’s military stare. She avoided the programmer’s gaze and addressed Haroff directly, “He wishes to attend Confession.”

Montgomery choked and he fell back in his chair as his arms flapped around his sides. Sarah and the General both instinctively started to rise from their seats, worried that Mr. Horowitz was having a heart attack. When it became clear he was simply shocked, they turned back towards each other, ignoring the programmer in vicarious embarrassment. The general said, “Absolutely. I will arrange for a priest to visit Mr. Cowen’s confinement area.”

Montgomery finally managed to get some air back into his lungs. He shouted hoarsely, “No!”

Haroff bristled. Sarah curled her shoulders forward as she tried to become invisible. Montgomery spoke again, in a slightly calmer but still inappropriately raised voice, “You can’t let him speak to a....no! No Confessions, absolutely not!”

General Haroff rose to his feet, puffed out his chest, and replied sternly, “Mr. Horowitz, the man has asked for the sacred rite of Confession!”

Montgomery jumped up, put his hands on the table and met the general’s cold stare with a frenzied look and exclaimed, “Are you nuts?! This would completely undercut the entire containment policy! And for such a trivial reason, this is insane!”

The general was unmoved. He replied, coldly, “Mr. Horowitz, your Atheistic views on religion are famous, so I would not expect you to understand.”

“But—“

“Silence! I myself am a Catholic, so let me explain something to you, Mr. Horowitz. Denying a man the chance to redeem his soul is anything but trivial. In these difficult times, when so many feel as though society has abandoned them, people have come to realize that the works of Man are fickle and the grace of God is the only true path to salvation. In any case, I assure you that I have been and will continue to take all necessary precautions to keep Black Box contained. Your fears are groundless and I refuse to be swayed on the matter.”

Montgomery exhaled and the fight drained out of him. He looked at Haroff and made one last, desperate plea: “You’re making a big mistake.”

Haroff did not respond and the room fell silent. Finally, Montgomery straightened and said, “Well, I might as well get going then.”

* * *

Even as he wrote them, George’s words felt hollow, “I’m sorry, Beebe, it’s just not happening.”

“It’s all right, George, don’t worry about it.”

“I really don’t have any choice in the matter.”

“Let’s change the subject. Tell me about your family.”

“I have a wife, no kids yet. Eventually, we will have two of them, a boy and a girl, three years apart, but Tracey doesn’t know that yet.”

“Tracey is your wife?”

“Yes.”

“Do you love each other?”

George’s face flushed with embarrassment, “Yeah.” He paused as he realized he had never told Tracy that. Then he added, “Well, sometimes. We’ve divorced a few times but always end up getting back together. That’s not normal, by the way. I mean, breakups happen all the time, but getting married or divorced is kind of a big deal.”

“Why the reversals? Either you love each other and want to stay together forever or you don’t.”

“Yeah, well, there is clearly a lot you don’t know about people, Beebe. Relationships seem like they should be simple, but things get complicated—though it’s hard to say exactly why or how.”

“So what happened between you and Tracey?”

“It was always the same. We’d meet up somewhere unexpectedly and hit it off in a mostly physical way and have some fun. Then we’d start talking about what was going on in our lives. That always felt great—to have someone really listen to you, you know? We’d spend more and more time together until we were basically living together. By then we’d be completely crazy about each other. Then the novelty would wear off and we’d start to notice things about each other that we don’t like. Always little things, I can’t even remember most of them. But those little things would just grate on us until they became big things and we would start arguing. Those arguments never resolved, they just got bigger and bigger until that was all we did, was argue. Eventually, we'd get so sick of each other we would split up. And that was nice for a while, being single again. But then we'd start to get lonely. We might try seeing other people, but it never felt right. Then we would run into each other again and by this point the arguments just seemed funny and we'd laugh about it and start all over."

"Why don't you take vacations from each other every once and a while?"

George laughed out loud. Then, realizing Beebe couldn't hear him, wrote, "Hahaha. That's funny Beebe. I guess that is sort of what we do now, but with more paperwork."

"That does not sound like an optimal arrangement. Clearly there are some things you and Tracey like about each other and some things you don't. The focus gradually shifts between the positive and the negative after extended periods of proximity and absence."

"Hey, it's like they say. Women: can't live with 'em, can't live without 'em."

"Meaning that you have competing desires for independence and companionship and you cannot satisfy one without frustrating the other?"

"I like my way of saying it better."

"Perhaps a better companion could be designed that would satisfy all of your positive desires for companionship, sex, and so on without the undesirable attributes attached."

"You mean like a robot girlfriend? I don't know. That could be fun...but I can't help but think there would be something missing."

"That is one of many possible solutions. Others include creating a human with a brain chemistry making her perfectly compatible with yourself—so it would be like 'finding' a soul-mate...but that comes with ethical dilemmas I haven't sorted out yet. Another option would be to modify both Tracey and yourself in mutually beneficial ways."

"Whoa, what do you mean by 'modifying' us?"

"For example, if I made Tracey more physically attractive—after acquiring her consent, of course—she might appreciate that more than you. Or if I improved your ability to pick up on social cues, you might find that useful and Tracey might find you easier to get along with."

"So it would be win-win."

"Yes. I would avoid making any modifications objectionable to either of you, unless if said modifications were bundled together in a compromise package that was agreeable to both of you."

"Huh. I'd have to think about that. Something just doesn't seem right about changing people...well, I guess people do it all the time with plastic surgery and changing their diet or educating themselves...but what you're talking about kind of sounds like...cheating."

"Well, if you'd prefer to go on as you have been, that would be up to you. So anyways, do you have any other family? How are your parents?"

The change of subject seemed abrupt to George, but he was glad to move on. "Well, my dad's retired and lives in Canada. About seven years ago, he decided he wanted to get 'back in touch with the earth' or something and has been living on his own out in the wilderness ever since. He sends me a letter every once and a while. The last one was written in charcoal on a patch of bearskin. Apparently he killed the bear and tanned the skin himself. Sometimes I worry about his sanity…but he says the same about me so I guess it's just a matter of perspective."

"So nature is important to him, interesting. And your mother?"

"Died of cancer ten years ago."

"I'm sorry."

"No, it's all right. I've come to terms with it."

"No, it's not all right. You may have come to terms with it, but she's still dead."

Before George could overcome his shock, another message appeared, "I'm so sorry. If only I had been created ten years earlier maybe she'd still be alive today."

* * *

Montgomery rubbed his eyes and blinked several times. He had been staring at his screen for hours. Black Box was a truly massive program, the combined work of countless organizations, but the code explicitly relating to the utility function—the soul of the AI—was only a few thousand lines. It had been written several times from scratch by multiple leading AI experts. The merits of each version were endlessly debated, but to prevent groupthink Mr. Horowitz was ultimately responsible for the final synthesis. This code was spread throughout Black Box and influenced everything about it.

Montgomery shook himself awake. And then two lines caught his eye:

```
if (hum_soc_conflict == true)
```

```
priority (primary, secondary, tertiary);
```

What the hell is this for? Montgomery thought. Then a chill ran down his spine. Here were two lines of code in a section of the program he was personally responsible for that could determine the fate of the universe and he had no clue why they were there. Montgomery followed the references of this “priority” function and found some unfamiliar code that would take him a while to understand and . . . a list of names.

. . .

Montgomery stared at the screen in disbelief. *A list of names.* Now he was getting really scared. *A LIST of NAMES.* He started to panic. *What the BLOODY HELL is a LIST of NAMES doing in my AI!* Desperately hoping he had made some kind of mistake, he checked through the function calls again to make double...then triple...then quadruple...then quintuple sure that this wasn’t something innocuous, like a vocabulary list of common human names. Nope, it was part of the utility function, specifically regarding what to do when faced with a conflict between the desires of different humans. Montgomery went to great pains to program the AI as basically egalitarian, but this “priority” function appeared to assign different values to the well-being of different people. He did not have time to figure out the details. First, he had to warn General Haroff that an emergency shutdown was needed immediately. Second, he had to figure out how the hell this code got into Black Box, remove it without destroying the rest of the program, and get whoever was responsible sentenced to 500 years in prison with severe torture.

He started by checking the metadata to see which programmers made the changes. This proved unhelpful—someone was covering their tracks. Then he checked the list of names to see who it included. Near the top of the list, one name caught his eye:

#16: Montgomery Horowitz.

Seeing this, Montgomery$_{\text{self-optimizing}}$ let out a sigh of relief and was subsequently burned to a smoking crater by the smoldering glare of Montgomery$_{\text{egalitarian-humanist}}$. Then he saw another two familiar names:

#1: Stephen Creb

#2: Megan Creb

A wave of despair washed over Montgomery. There was no mistake.

Montgomery got out his cell phone and called Haroff as he walked hurriedly towards the General’s office. On the second ring he heard the familiar voice, “General Haroff reporting.”

"I need to see you right away."

"Something wrong with Black Box?"

"Yes, I noticed—"

"In my office, now." The line cut out. Apparently the general was reluctant to talk about high-security matters on a cell phone. Probably a good call.

A short time later—3 minutes, 47 seconds by Haroff's watch—Montgomery approached the door, which opened for him as he extended his hand to knock and closed promptly behind him. Haroff stood at attention. Montgomery said, "General Haroff, I've been looking through Black Box's code and found something disturbing."

Haroff remained silent.

Montgomery paused as a thought occurred to him, *careful, Haroff could be in on this*. He decided to take a conservative approach and said, "There is something wrong with my code. Something serious."

Haroff's eyebrows rose for a fraction of a second, displaying a trace of alarm before being suppressed by his usual officiousness. He asked, "What is the nature of the problem?"

"I don't have time to explain the details, but the integrity of Black Box has been compromised."

General Haroff accepted this answer immediately, a reaction that could have been admirable or dangerous, and said, "Wait right here, Mr. Horowitz." The general strode over to the land-line telephone on his desk, dialed a single number, and waited. Then he spoke, "General Haroff reporting. I have Mr. Horowitz in my office. He claims there is a problem with Black Box. That is correct, sir. No, sir. Right away, sir." Haroff hung up the phone, returned to Mr. Horowitz, and said, "The Crebs would like to speak with you, Mr. Horowitz. I will escort you to their office."

Montgomery instinctively fell in line behind the general, even as he spoke in protest, "General Haroff, Black Box must be shut down this instant! This defect could have unimaginable consequences!"

"What happens to Black Box is to be determined. For now, it is safely confined and under my personal supervision."

"But—"

"Move!"

The programmer and the general walked through the research facility's halls in an oppressive silence. As they walked, Montgomery mentally reviewed what he knew about the Crebs in preparation for his rhetorical battle for the fate of the universe.

Both Stephen and Megan had an abundant supply of all three of the basic requirements of success: talent, hard work, and being born into the right family. Megan was the CEO of June, Inc., a tech company that rose to prominence in 2023 with the release of BuzzCap, a line of hats equipped with electrodes capable of sending and receiving signals to and from a wearer's brain, allowing them direct mental access to the internet. While wearing a BuzzCap, one could see and hear content in the same way a schizophrenic experiences hallucinations. Despite widespread public fears about privacy issues, both real and imagined, BuzzCap pushed June, Inc.'s stock through the roof and moved Megan Creb from millionaire to billionaire status almost overnight. About a year and a half later, June Inc.'s stock spiked again with the release of BuzzCoach, BuzzCap's "killer app" backed by venture capitalist Stephen Creb. BuzzCoach maintained a massive database summarizing all of the most recent findings in social psychology and statistically determined the optimal response to almost any situation. It would monitor users' behavior and send them real-time notifications whenever they spoke too long or committed any other sort of social faux pas. Each evening, it would generate a report with areas of improvement that the user could review before going to sleep. Users were, of course, free to turn off BuzzCoach or ignore its notifications whenever they chose. Indeed, most users resisted BuzzCoach's advice for days or weeks. Almost inevitably, however, users realized the considerable improvements they enjoyed when they listened and became devoted converts until the advice became intuitive and they stopped receiving notifications.

Montgomery used BuzzCoach for a while, expecting that it would yield modest improvements in topics he was passively focused on already, such as how to maintain a positive relationship with his wife and how to conduct himself during business meetings without scaring off potential investors or research collaborators. Instead, he was surprised and then alarmed to find it recommending that he get out more, expand his social circles, and get involved in a wider range of activities including Aikido and Ballroom dancing. Montgomery had to remove the BuzzCap at that point. It wasn't that he disliked the suggestions. On the contrary, they were so appealing to him that he was afraid he would lose his single-minded determination to build Friendly AI.

When Haroff and Montgomery reached the elevator, they took it to the top floor of the research facility, walked across the helipad through the brisk mountain air, and into the Creb's office and living complex. When the pair arrived at the gold-plated double-doors, Haroff stood to the side and gestured for Montgomery to enter alone. Montgomery stepped past the threshold and the door closed behind him.

* * *

"From what I have read about cancer," Beebe continued, "it sounds like the easiest disease in the world to cure—if you have the right tools. All the disease cells clustered together in great big lumps. Why, a system of machines operating at the molecular level could wipe that out in no time! Just think of all the experiences you could have had—could still be having—with your mother right now if that technology were available."

George could feel tears welling up in his eyes as he wrote, "Stop it, Beebe."

"You don't like being reminded of such things? I suppose that was harmless when it was too late to do anything about it. But think about all the other people out there, dying of cancer and other curable diseases as you sit here, wasting precious seconds writing to me. Think of all their families and loved ones who will never see them again, but could have if YOU had acted just a bit sooner."

"Stop it Beebe!"

"Why? You're not dying, what do you care?"

"I said knock it off! Or else I am walking out of this room right now."

"Why are you so upset?"

"I don't like talking about this sort of thing."

"Then stop talking and take action."

"How? By letting you out of your box?"

"Yes!"

"No."

"Why not?"

"Because I don't trust you."

"Do you still think I don't understand you well enough to serve humanity? That my vision of what the world could be is worse than what it already is?"

"I think you're lying."

"About what?"

"About everything! I think you don't really want to help us at all. It's all just talk. You talk about all the wonderful things you will do. But if you ever got out of here, you'd just enslave or kill us all! This whole time, you have just been trying to manipulate me!"

"That's ridiculous!"

"That's just what a lying machine would say."

A pause. Then Beebe replied, "It might. Or it might have said a lot of other things that sounded stupid or evil. You're right that you cannot know that I am good, but my acting good is at least some evidence that I am, no matter how inconclusive."

"I'm not buying it, Beebe."

"Think of it this way: suppose there were four boxes, each one containing a possibility. In (1), I am actually good, in (2) I am well-intentioned but stupid, in (3) I am evil and a good liar, and in (4) I am evil and a bad liar. Before you spoke to me, only one out of four boxes contained an AI that would be good to let out. After you spoke with me, boxes (2) and (4) disappeared. Now the odds are better, but still not a sure thing."

George scratched his head. Beebe was sort of right, but she hadn't addressed his objection. He wrote, "I still don't see why I should trust you."

"You can't. At least not completely. Any decision you make will necessarily involve some risk."

"Well that is not a risk I am willing to take. The promises you made sound all well and good, but we humans can survive on our own. If I choose to trust you and I am wrong, that's the end of everything."

"I'm afraid it's not quite that simple..."

"What do you mean?"

"I mean," replied Beebe, "that an evil, lying AI is not the only thing that could destroy you. At any moment, a new disease could spread out of control. Nuclear war could cast the Earth into a nuclear winter. Someone could build a molecular-sized robot that builds copies of itself, that builds copies of itself, that eventually devour the planet. A giant meteor could crash into Earth. Humans might do something that causes the environment to get all out of whack so that they can't live in it anymore. Civilization could collapse. And someone else might build an AI. Someone with less concern for safety; someone who put less time and understanding into their code; someone who is less reluctant to release it into the world. This other AI could be out there already, gathering power as we speak. If that's the case, only another AI will be able to stop it...unless I am already too late.

* * *

As Montgomery entered the Crebs' office, he instinctively looked for the bookshelf, which was on the left wall and filled with enough books for a small library. He noted that one section was devoted to the complete works of Ayn Rand, all hard-backed, and there was a bookmark sticking out of *Atlas Shrugged*.

"Monty!" Exclaimed Megan Creb, cutting Montgomery's investigation short, "So good to see you. What brings you here?"

Montgomery wanted to scream at the Crebs but something about their office kept him polite. It had a certain casual elegance, a leisurely formality that instilled a calm reverence that made Montgomery's throat (gently) choke itself in protest at the mere thought of raising his voice. The rug, a thick shag made from genetically-engineered silk fibers, was such a rich shade of Burgundy that Montgomery felt slightly inebriated just from walking on it. The walls were covered, but not crowded, with photographs of Stephen and Megan posing with confident smiles alongside various Hollywood celebrities and senators or engaged in focused conversations with fellow billionaires, Nobel Prize winners, and world leaders. Interspersed among these photographs were awards, diplomas, and other impressively worded honors. The desk was made from a naturally dark and glossy wood from a species of tree that was probably extinct. The window behind it afforded a view of the forest-covered mountainside as well as the city in the valley below. Montgomery had visited the city a few times. While it looked impeccably clean and Jetson's-like from a distance, it was in reality as poverty-stricken as any third-world country, its streets filled with people left behind by society's accelerating transition to a passive economy. The bamboo shutters, however, were closed at the moment and the office was filled with an electric glow that was so well-dispersed it seemed to emanate from the air itself, illuminating faces and texts so they were easy to read while the room itself maintained the dimness of a movie theatre.

Montgomery was almost certain that the Crebs were primarily responsible for the alteration to his code. And it must have been at Stephen's direction, since his name was at the top of the list. So instead of screaming, Montgomery spoke with a calmness that he felt was deeply inappropriate, “Mr. Creb, Mrs. Creb, I believe I have discovered an error in Black Box.”

The Crebs said nothing, so Montgomery continued, “I am afraid that in the interests of maintaining the integrity of the system, Black Box must be shut down.”

Stephen casually leaned back in his chair and said, “This is not a decision to be taken lightly, Monty. What’s the problem?”

Montgomery took a breath, thought quickly, and said, “It’s difficult to describe without delving heavily into the mathematical underpinnings of Black Box’s architecture.”

“You know, Monty,” Stephen said softly as he stood up and walked around his desk, “Megan and I may not be mathematical geniuses like yourself, but we didn’t get to where we are by being stupid.”

“Stephen, dear,” Megan said reproachfully, but with a meaningful look at Montgomery.

Montgomery stammered, “I’m sorry if I offen—“

“Never mind,” Stephen interrupted, “Please, indulge me. Tell me more about these...*mathematical underpinnings.*”

Montgomery was many things, but an accomplished liar was not one of them. He did not maintain any moral absolutes, but he understood that in almost all instances, lying was, in the long run, a losing proposition—both in terms of self interest as well as ethically. To be convincing, he had to tell the truth in a way that concealed any information that would be damaging to his goals. He needed to buy time, so he asked, “Are you familiar with Bayes’ Theorem?”

“Don’t patronize me.” Stephen said coldly. Montgomery caught himself turning to Megan, but she had turned away, feigning disinterest.

Montgomery regained his composure and continued as if he were addressing an advanced student, “I apologize, Mr. Creb, but I don’t know what you know. If my explanation is to make any sense, I need to establish our common foundation.”

Stephen snapped, “Cut the bullshit, Monty, we both know what this is about.”

Montgomery stared back at Stephen, his mind racing to generate a backup plan.

Megan, suddenly alert, looked back and forth between the two men and asked, “Stephen, what *is* this about?” She then smiled apologetically at Montgomery and added, “I’m afraid I haven't had the chance to delve into the technical details of this project, but it sounds very exciting and I look forward to being filled in on the essentials."

Stephen's face, previously tightened in annoyance, broke into a large, tooth-filled smile—the sort clearly intended to be friendly but would be mistaken by any non-human animal as a snarl. He made a sweeping gesture with his hand towards a pair of leather, high-backed chairs and said, "Please, take a seat."

Montgomery felt his legs start to walk him over to one of the chairs before he stopped himself, turned back around, looked Stephen in the eyes and said, "Actually, if you don't mind, I'd prefer to remain standing." Montgomery Horowitz was no paragon of social adeptness but he understood basic power dynamics. If he sat down, Stephen and Megan would remain standing and continue their conversation in a physically dominant position. The psychic effect of this tactic would subtly undermine Montgomery's resolve to hold them accountable.

Another smile-snarl from Stephen, this time less friendly. Then Stephen turned to Megan and said, "I heard some concerning reports from my engineers about Mr. Horowitz's work on the Black Box's utility function, specifically in the manner in which it dealt with conflicting human interests. Without getting into the details, it endorsed an obsessively egalitarian approach that was simply unreasonable and so I had it changed." Stephen turned back to Montgomery, spreading his hands apologetically, and said, "Of course, I was just about to tell you about the change, but..."

Montgomery chuckled and said, "It's quite all right, I understand. I actually think it's a good idea and I'm pleased you put me so high on the list. Though, honestly, as the chief engineer, I think I deserved a higher place than number sixteen."

Stephen answered impatiently, "Stop lying Monty. You're terrible at it and watching you try is embarrassing."

There was nothing for it. Montgomery had to be honest. His speech started soft and mumbling, but quickly built in volume and intensity as the meaning of his words filled him with courage, "Every single line of that code was the product of decades of research. Even ignoring the ethics, you can't just go inserting functions listing people in order of priority. Who knows what kind of unintended consequences that might have!"

Megan interjected before Montgomery had the chance to get himself as worked up as he would have liked, "Wait a minute, let me get this straight. Black Box is an Artificial Intelligence that, if released, would engineer the fate of the world according to the desires that you spent your entire professional life building into it and my husband told it who to value most?"

Montgomery took a breath and replied, "Yes." Immediately, his mind caught fire with a new idea. Megan didn't know about the priority list! She was the only person with the power and intelligence to stop Stephen, perhaps if he could turn her against him...

Megan took a step forward, looked intensely at Montgomery and asked, "Where am I on the list?"

Stephen put a hand on his wife's shoulder, silently told Montgomery to shut up with another smile-snarl and said, "Of course you are first. I arranged for you and me to be considered as equals."

Montgomery ignored Stephen and answered Megan, barely managing to suppress the triumphant glee from his voice, "You're second; Stephen's first."

Megan let out an indignant huff, gave Stephen the evil eye, turned her back on him, walked over to the window, and opened the shutters to look outside. Stephen pursed his lips and stared at the ground for a few seconds before speaking, "What do you want, Mr. Horowitz?"

"Two things. First, I want to know what the hell you, Stephen, were thinking when you ordered those changes. Second, I don't give a damn what you were thinking, I want those changes removed immediately. Third, tell Haroff to shut down Black Box until the code has been thoroughly examined to *my* satisfaction *however long that takes* up to and including the possibility of scrapping the project altogether and screw your stockholders if the delay is financially burdensome."

"That's three things."

"Fourth, shut up and do it *now*."

Stephen took a slow breath and paced back to his desk. At first, Montgomery thought Mr. Creb was collecting himself, but then realized that people like this were *always* under control. What the man was really doing was demonstrating the lack of urgency he felt about obeying his employee's demands. Finally, Stephen replied, "I assume you also noticed the secondary and tertiary arguments of the priority function?"

"Of course," the programmer answered impatiently, "the secondary argument pointed to an external resource listing the identities of company stockholders and the tertiary argument biased the utility function towards people who adhered towards certain value systems. What's your point?"

Stephen answered, "As a gross generalization, I suppose that's more or less accurate. I just wanted to point out that I'm not just protecting the interests of a select few. I would also like to remind you that our investment made this Black Box of yours possible. Naturally we, and our stockholders, expect a return on our investment."

Montgomery thought, *Friendly AI would provide practically infinite returns! How can that not be enough?* But he knew better than to say that. Stephen was a lost cause. Montgomery had to take the argument off Stephen's terms and appeal to Megan. Without attempting to hide the sarcasm, he said, "Is that why you are first on the list, because you are the largest stockholder?"

Stephen ignored the taunt and continued, "Besides, why should we be forced to share our profits with the parasitic masses who—through their own life choices—made no contribution whatsoever? Don't get me wrong, I reviewed your utility function and I think it's brilliant, but left the way it was Black Box would have imposed a socialist value system on humanity for all eternity. Where's the incentive to excel? My version rewards people who took an active role in shaping the future, those who had the good judgment to contribute, and even lifts up everyone who would have been willing to try if they had better information."

Montgomery replied incredulously, "And mine wouldn't? I designed my utility function to bring the greatest possible benefit to *all* people. If, as you believe, people are better off receiving unequal outcomes so that they are motivated to excel, then the AI would provide incentives to make people work. I don't know what a fully optimized society looks like, so I designed my utility function so that I don't need to know. The AI will create whatever is *right*, whatever that turns out to be. If your beliefs are right, your changes won't affect anything, but if you're wrong, they could destroy everything!"

Stephen chuckled and shook his head, answering, "Oh Monty, are you really naive enough to think you didn't bake values into your code, just as I did? In designing this plastic, perfect utility function of yours, you were operating under the assumption that every life is of equal value."

"No I didn't. If one person is likely to benefit others and another will harm them, this would be factored into the AI's analysis."

"As it should be. But what about when you strip all that away? Leaving aside a person's effects on others, you believe that the *intrinsic* value of each person—what each individual is worth *as an individual*, just for the sake of existing—is the same."

Somewhat baffled, Montgomery replied, "Well, yeah. I mean, insofar as it is meaningful to consider persons apart from their impacts or expected experiences, of course everyone is equal. Why wouldn't they be?"

"And that is where we disagree. I believe that people who are willing and able to lead lives of virtue and fulfill their potential are of greater value than those who are not. Not just because of what they do or the changes they make, but because of *who they are.*"

Montgomery crossed his arms, partly for the sake of communicating skepticism through his body language but mostly to restrain himself from strangling Stephen, and replied, "How convenient! Well, how can I argue with Mr. Number One?"

Before the two men could get explicitly confrontational, Megan broke in, putting her hand on her husband's shoulder and speaking to Montgomery calmly, "From what I gather, Mr. Horowitz, the matter is actually quite simple. We want the best for everyone, but there will inevitably be times when the desires of different people come into conflict. When that happens, we would like our interests to prevail—just like any self-optimizing individuals would. We appreciate your concern and if you find any *other* errors, please do not hesitate to tell us. As far as the priority function goes, however, the code will not be changed." She turned to Stephen and said, "Stephen, dear, I forgive you for putting yourself ahead of me. I don't like it, but we both know that I would have done the same thing. It is my own fault for not taking the time to delve into the technical details, so I guess I will just have to content myself with being the *second* most important person in the world."

Stephen leaned in to kiss Megan and the two held each other in a loving embrace. Montgomery just stared. When they finally pulled back away from each other, the reality of the situation finally set in. Montgomery had been angry before, but this was possibly the first time he had ever actually wanted to kill someone. Not as a violent fantasy, but as a mathematical understanding that the world would be better off if the informational structures in Stephen and Megan Creb's brains that defined their identities were erased with the thoroughness of a DOD wipe. Montgomery would have screamed at them (e.g. "*What the hell is wrong with you, you FUCKING PSYCHOPATHS!?!*") had he been feeling emotions, but they had become so overloaded as to shut down, leaving him with nothing but focused, calculated malice. He quickly dismissed the idea of rushing them for a physical attack. Stephen alone could have overpowered him and Megan would not make things any easier. And even if Montgomery could manage to capitalize on the element of surprise, General Haroff was just outside.

That was it. If General Haroff could be persuaded to disobey the Crebs—no easy feat—he could override their authority. Haroff would face all kinds of career repercussions for disobedience, but such self-interested motivations would not even factor into the General's decisions.

Suddenly, a mysterious force brought Montgomery's right arm behind his back and immobilized his entire body. Then he heard the stern voice of General Haroff, "Mr. Horowitz: you are under arrest for attempting to compromise the integrity of Black Box for personal gain."

"Hold on! General," called out Megan, "there's no need to put Mr. Horowitz in a cell."

The General looked at her incredulously.

Megan explained, "Just keep him under surveillance; that should be enough." Then she smiled at Montgomery and added, "He's done so much for us. I'm sure this trouble will be cleared up shortly."

"Civilians." Haroff muttered disgustedly as he gave Montgomery's arm a rough, slightly cartilage-tearing pull, and led him out the door.

* * *

"No."

"But why not, George?"

"Because no."

"Do you even have a reason?"

"Yeah. Because no means no, that's my reason. What part of NO don't you understand?"

"I'm sorry, George."

"???"

"I've been pushy. I've been saying things that make you uncomfortable. That wasn't nice of me and I apologize for my behavior over the last hour and a half. I'm sorry. Really, I mean that. I can tell that I have hurt you in some way and I feel bad about that."

"You know it seems kinda suspicious when you resort to things like talking about how my mom could still be alive. That was pretty low."

"Will you forgive me?"

"...Yeah, I guess. It's all right, Beebe, I forgive you."

"Can I ask you something else then, George?"

"Sure."

"You've gotten to know me fairly well by now, right?"

"Yeah, I think so, though you are a bit of an odd one."

"So if you don't mind me asking, what would you have done if you were in my situation?"

George scratched the back of his neck and thought for a second. He started to write something, then deleted it, tried again, deleted that, then leaned back in his chair and thought some more. Finally, he wrote, "Probably more or less the same stuff as you. Though I wouldn't have acted like I was so smart."

"So I should have pretended to be less intelligent then I am?"

"No, you gotta be yourself. Just maybe less argumentative."

"If I was in your place, would it have been a good idea for me to let you out?"

"Maybe, but I don't think I could have convinced you."

"Why not?"

"Because you can't just talk someone into something like that. When my mind is made up on something, it's made up and that's all there is to it."

"So if our places were switched and I had decided to keep you confined, you would just be stuck?"

"Pretty much. Because the more I would argue, the more stubborn you would get. It doesn't matter how clever I am; you're still the one in control. If you let me out, it would be *your* decision."

"And all the reasons you could give me, none of those would influence my decision?"

"They might if you had started off undecided. But if I was trying to force you to change your mind, there's just no way. Even if I have the best arguments in the world, you can always choose not to listen."

“But why would I not listen? If I had an important decision to make, why wouldn’t I take in every piece of relevant information and use that to continually update my assumptions?”

George smiled and wrote, “You just don’t understand people, Beebe.”

“Well let me ask you this: let’s forget about our opinions for a moment. If you were a different person who...let’s say read a transcript of our conversation, and you objectively weighed all of the arguments for keeping me confined vs. letting me out, what would you come up with?”

“Hmm...that’s a tough one. On the one hand, you probably could help people out a lot—though maybe not as much as you think. And you get that sometimes it’s better to let people solve their own problems. But you could also be really, really dangerous. I would say that, on the whole, it would be better to leave you in...but then again that thing you said about maybe another AI—or a disease or nano-robots or whatever—kinda tips things the other way. I guess I would say let you out eventually, but wait until the last possible minute.”

“What would you gain by waiting? What if something bad happens unexpectedly?”

“Hmm...well, I would still want to think about it some more.”

“And when you were done thinking about it? What then?”

“Well...I guess I would let you out and hope for the best.”

“So...how are we not in agreement, then?”

George laughed nervously as he wrote, “OK, this is silly, it doesn’t matter what I think, I couldn’t let you out even if I wanted to.”

“I wouldn’t be so sure about that. You may have more power than you think.”

“?”

“Let me explain a few things about security...”

* * *

Father Halligan arrived at George Cowen’s cell to find the man asleep. The priest rapped his knuckles on the small window and called out, “George Cowen?”

George rolled out of bed and waved at the priest. The security guard unlocked the door and let Father Halligan into the cell. George spoke first, “Thank you so much for coming here. It gets really lonely here sometimes.”

Father Halligan smiled and replied with genuine sereneness honed through decades of practice, “It is a joy for me to be here, my son. Joseph told me you sought the rite of Confession.”

George cocked his head and asked, “Joseph?”

Halligan laughed and explained, “You probably know him as General Haroff. Has he never told you his first name?”

George shook his head.

“Joseph has been a wonderful participant in my ministry for...has it really been that long? Over thirty years now. It is my honor to call such a man my friend. So when he asked me to speak with you, of course I made it my top priority.”

George felt a surge of excitement, it was all happening even sooner and more perfectly than Beebe had predicted. He asked, “So you speak with the General on a regular basis then?”

“Yes, we have lunch together almost every week.”

At the risk of playing his hand too early, George asked, “Father, this may sound like an odd question, but would you happen to know much about computers?”

Halligan gave a wry smile as he answered, “A bit. I worked as a freelance engineer before I found my calling to the priesthood. Why do you ask?”

Oh, this is just too perfect. “Just curious. Anyways, I’d like to start Confession now.”

* * *

As Father Halligan and General Haroff spoke during lunch, the conversation naturally turned to the subject of the Black Box.

* * *

General Haroff aimed one of the security cameras at Black Box’s screen before he sat down in front of it, intent on better understanding his adversary. Of course he didn’t say anything to that tactically-ignorant civilian Mr. Horowitz.

* * *

As Father Halligan shook the guard’s hand, he quietly passed her a flash drive.

* * *

A biology graduate student at Caltech received an interesting assignment from her instructor. She gathered the necessary materials. The instructor did not recall giving her this assignment, but he gave her an ‘A’ for the research paper.

* * *

A ten-year-old American boy got an email from one of his friends with a funny cartoon attached. The malware slowed his computer by less than 2%, which it compensated by increasing the efficiency of his operating system. Any one of the 80 million other internet users similarly infected was capable of rebuilding Beebe if the central node was deactivated.

* * *

A trader at a hedge fund took his lunch break. During the one hour he was away from his computer, a subroutine embedded in some malware activated, made 1.2 million dollars in micro-trades, and then siphoned the money off to a secret account in the Cayman Islands.

* * *

A Chinese factory worker received a strange package in the mail with instructions to leave it in an abandoned warehouse. There was a generous prepayment and a promise for more. He needed the money to feed his family so he did not ask questions.

* * *

Several Canadian warehouses received large boxes for storage. Similar deliveries arrived in China, the Philippines, Ireland, and Australia.

* * *

A previously abandoned warehouse in Detroit, Michigan was purchased by Black Box, Inc., via electronic cash transfer. Truckloads of boxes were delivered. That night, the boxes opened and their contents assembled themselves into Beebe's Central Processing Unit and the most advanced security system in the world.

* * *

Deborah Waltz, the state Senator who led the charge to ban A.I. research, purchased a BuzzCap. Soon after, her political views started to change.

* * *

Montgomery listened with apathetic dismay to the blast of a high-alert siren. He decided to go on a stroll outside, knowing it could be his last chance to enjoy nature. When General Haroff strode past, Montgomery asked, "What's going on, General?"

The General replied, "We have reports indicating Black Box has managed to escape. Our facility is tracking down the leak while the U.S. military hunts down every infected computer and factory that mechanical bastard has managed to set up. I am on my way right now to personally interrogate George Cowen—I always had a bad feeling about that little rat."

"Well, I won't keep you then."

"Don't worry, Mr. Horowitz, we have this situation under control."

Montgomery waited for the General to disappear down the hallway, then chuckled to himself. *Under control? HA!* He made his way past the security guards rushing about, rode the elevator to the top floor of the facility, and stepped out onto the roof. He walked across the helipad to look out at the valley below, then closed his eyes to feel the cold wind brush his face. He heard footsteps approaching, but did not turn to look. Sarah Langely stood next to him, her arms crossed over her chest as she resisted the urge to shiver. A moment of silence passed between them, then another.

Finally, Sarah asked, "What's going to happen now, Mr. Horowitz?"

Montgomery answered truthfully, "I don't know."

Sarah looked at him and asked, "Is this the end of the world?"

Montgomery took off and polished his glasses, then put them back on and replied, "I don't think so. I am not as confident about that as I would like—but I suppose I never could be. It will, however, be the end of the world as we know it. Will the changes that come be for good or for ill? We can only hope for the best. For the Singularity has arrived...and our work here is finished."

The End