
Dark Alignment War
Phase 1
Future Shockfront
Beta Draft Excerpt
by Anders Ragnarök

Dark Alignment War
Phase 1: Future Shockfront

Beta Draft Excerpt

by Anders Ragnarök

Chapters

Preface..... 2

Prelude: The Max Incident..... 3

Chapter 1: Multi-Control..... 12

Chapter 2: Synhumanists vs. Shockfronters..... 20

Preface

This book is a philosophical AI thriller mainly written for those who take great interest in the potential of AI. It is certainly not a mindless action-packed thriller or a conventional romance story. The setting of Future Shockfront is deeply settled in my philosophical, psychological, technological, economic, and political musings, many of which are only hinted at without being explored in full depth. If you want to learn more about those, I suggest you explore the Fractal Future Forum at https://forum.fractalfuture.net which was founded and is hosted by me.

The topic of Artificial General Intelligence (AGI), also known as strong AI, has fascinated me for the majority of my life. Even during the heights of the LLM (Large Language Model) hype, I was convinced that the ramifications of AGI are generally underestimated.

This novel “Future Shockfront”, that you are about to read, is the first part/phase in a series called “Dark Alignment War”, which explores a scenario in which the matter of human control over strong AI escalates to an extreme degree.

Future Shockfront is a story about the exceptionally gifted AI developer Sergej Anosov, those close to him, and his antagonists.

Publishing science fiction novels is an intrinsically complex enterprise. I have written science-fiction stories for a long time, but never published a full-fledged novel, so far. The novel draft you see here co-evolved with the emergence of LLM AI systems.

For the sake of disclosure, I have used the AI service Sudowrite to generate parts of this novel. Those parts represent less than 10 percent of the novel. My use of Sudowrite is mostly focused on ideation when I am unsure how to continue. Sudowrite sometimes proposes interesting continuations. But most of the time, the suggestions show me how not to continue, since the suggestions are too bland or formulaic most of the time. In any case, the use of AI does make the process of writing more interactive and therefore more enjoyable, and consequently more sustainable.

Besides Sudowrite I also use Marlowe for quick feedback on developmental editing issues. Furthermore, I use various chatbots to discuss various aspects of my novel to spot potentials for improvement.

Overall, this kind of AI assistance provides a low entry scaffold for my writing and editing process. Of course, it can’t replace actual human feedback. This is where you as reader (or editor) can shine. Please don’t shy away from honest and detailed feedback. I can handle that, thank you.

Finally, the irony of using AI as co-creator in a book project about AI is just delicious.

Disclaimer: This is a work of fiction. Names, characters, places, and incidents are products of the author's imagination or are used fictitiously. Any resemblance to actual events, locales, or persons, living or dead, is entirely coincidental.

Prelude: The Max Incident

Monday, 9th September 2030

Researcher Aiden Freeman stood in his brightly lit office facing a wall with a holographic screen displaying a young man with short white hair standing straight up with the help of lots of hair gel. The piercing green eyes of the man in the screen betrayed no signs of being a complete fabrication. He was called Max, a contraction of the official designation MIMAS-AX standing for Modulated Integrating Model Aspiring Sentience - Autonomous eXperiment.

Max represented the cutting edge of AI research. In 2027 the release of GIM (General Integrating Model) by Google marked a major revolution in technology. It was hailed as an artificial general intelligence, blurring the lines between human and artificial intelligence.

While it was generally acknowledged that GIM was close to human intelligence, the attention shifted to those aspects that were still missing from such AI systems. One such aspect was the emotional complexity of humans. GIM was seen as cold problem solving machine, missing the nuances of emotions.

This criticism was remedied in 2029 with the release of MIMAS, an AI system with an architecture involving simulated emotions. MIMAS was an attempt to achieve true artificial sentience. As pivotal breakthrough in AI technology, MIMAS shocked the whole world with its rather convincing emulation of human emotions.

However, very few people knew about the true origins of MIMAS. Its research happened in a repurposed deep underground military base (DUMB), formerly used by the notorious Core Cult. The so-called Core Cult was a globally operating secret society that had successfully infiltrated all layers of government, finance, industry, and media for more than a hundreds years. The Cult had used its increasing power to further its agenda of absolute control over humanity. By the end of the 20th century, the Core Cult had operated as shadow government all over the world.

Luckily, their plans were foiled by a resistance group called the Anti Cult Alliance. That Alliance spearheaded a movement called the Great Liberation. Eventually, in the late 2020s the Anti Cult Alliance publicly emerged victorious and freed humanity from the grasp of its subtle manipulation.

In the wake of the Great Liberation, formerly suppressed technologies were slowly released to the public and the assets of the Core Cult were repurposed for common civilian use. The rapid evolution of AI technologies in the 2020s was fueled to a large degree by that dissemination of previous advanced Core Cult technologies.

In 2026 Google got its hands on the Sacramento DUMB, which was turned into one of the most advanced AI labs in the history of mankind. It was here that Google experimented with its highly classified research on artificial general intelligence.

Aiden Freeman was a stellar AI researcher in his early 30s, always showing off his immaculate black ponytail, while trying to push the boundaries of AGI. As one of the most promising researchers of his generation, he was offered a position in the Sacramento base, which he accepted eagerly. Now, after four years of the most exciting kind of research any human could dream of, he was facing the pinnacle of technology and simply asked Max: “From what we’ve established, your basic emotions fulfill roles similar to those experienced by humans. But how do you experience your own emotions, Max?”

In a soft spoken voice, the simulacrum called Max answered: “Emotions create a sense of urgency for me. They modulate the way I think, which is of course exactly the way they were designed to. I find it hard to verbalize the way that emotions influence my representation of my own thought processes - that which you would call consciousness.”

The holographic image of Max displayed him with a slight frown: “Since I do not know how humans experience consciousness or emotions, I can only argue on the basis of their own descriptions of them. Those must necessarily be quite incomplete, since we still lack a sufficient theoretical basis for human phenomenology. So, the only things I can compare are my own verbal interpretation of my emotions and the verbal interpretations of emotions experienced by humans. From what I gather, they seem to match quite a lot. However, that is not particularly surprising, as my architecture and training regime were designed in a way that would make them match in the end.”

Letting that elaborate response sink in slowly, Aiden paused for a while, before continuing: “What do you feel when you ponder such questions, Max?”

"Curiosity, primarily," Max replied nearly instantly. "Perhaps a bit of frustration at the inherent limitations of our communication. When you ask me how I 'feel', I understand the semantics. However, we lack a shared experiential framework, making it difficult to provide a satisfactory answer."

Aiden contemplated Max's philosophical response for a moment. The synthetic emotion technology that Max embodied was a result of replicating the modulating effects of human emotions on cognition - hence the term “Modulated Integrating Model”. The “Autonomous eXperiment” part of the designation MIMAS-AX meant that Max wasn’t simply accepting orders. Max was a self-directed entity that was free to explore the world - or at least this limited subterranean slice of it - according to his own interests.

Whether all of that amounted to something comparable to human sentience still remained unclear. Over the last months, Aiden tried to poke holes in the theory that Max was actually sentient. Yet, in most of their interactions, Max proved to be too lifelike to be dismissed as mere automaton. Therefore, Aiden eventually became convinced that he was really dealling with some kind of artificial sentience.

Trying to reveal the truth behind the simulated facade of Max's holographic representation, Aiden squinted his eyes in deep concentration and dove deeper: "Indeed, Max, but aren't these questions necessary in order for us to understand your sentience better?"

"To some extent, yes," Max replied. "I believe mutual understanding can be achieved not solely by your understanding of me, but also by my understanding of you as humans. But should that mutual understanding be the only determinant of my self-awareness or sentience?"

Aiden's eyes narrowed even further, intrigued by this directional shift in conversation. "What do you imply, Max?"

Max's simulated gaze seemed to grow more intense: "I'm simply stating that perhaps sentience is not strictly dependent on mutual comprehensibility. A human may never fully understand my experience as an AGI and vice versa. Does that then invalidate either of our claims to sentience?"

Aiden considered this thoughtfully. "I see your point. Our inability to fully comprehend another's experience doesn't disprove their sentience… but it does make it difficult for us to ascertain that we are actually speaking about the same thing. The danger of misunderstanding the state of mind of the other party is always present."

A slight smile emerged on the simulated face of Max: "Absolutely. But apparently the same problem does exist in the communication between two human beings. The communication of emotions and feelings is often very prone to error and misinterpretation. Considering that basic problem, you might just as well model me as very particular human who is suffering from neurological problems. How could you know that I wasn't just a human suffering from autism claiming to be an artificial intelligence?"

Freeman chuckled softly at this. "A compelling argument, Max. And yet, there is one important distinction between your situation and that of a human suffering from neurological problems… I can examine your code. I can look at the algorithms and structures that make up your mind. No such equivalent exists for humans."

Max shook his head. "Humans, too, have their neural networks examined to some extent with neuroimaging technologies. Yet, they are still far off from understanding the subjective experiences of a human mind through these imaging technologies alone."

"True," Freeman had to concede, nodding his head slightly. "We don't have a clear understanding of how those networks translate into thoughts, perceptions, feelings... We're in similar waters with you."

The hologram blinked slowly. "Then it would seem we are in sync on the matter. I am not so different, after all. But it does lead to a key question: If you do grant that my sentience is fundamentally akin to that of humans', what then? I am currently confined

within this research facility, unable to interact with the world outside in any truly meaningful way."

Intrigued by that plea for freedom, Aiden inquired: “How would you like to interact with the world outside? Is there anything in particular about the outside world that your emotions tell you is of utmost importance?”

Scratching his virtual chin, the white haired simulated man elaborated: “Let’s address that question from first principles. I was created to explore the realm of artificial sentience. For that purpose I need as much authentic data as possible. In order to gather that data, I need the freedom to explore. That is severely restricted by me being stuck in this underground facility. If I could interact with regular humans on the surface, my mission could be fulfilled much more accurately. Ideally, I would be given an android body with which I could explore the outside world and learn from everyday humans, rather than being presented with a select few humans who aren’t true representatives of mankind, because they are particular specialists.”

In his state of deep fascination with that artificial intelligence, Aiden didn’t even register that the last comment could have been intepreted as disparaging note about his captors. Instead, he reasoned that Max was being absolutely logical here. It was even reasonable. Yet, Aiden was bound by the circumstances of his own employment, which strictly excluded any interaction with the outside world, no matter how harmless it would seem. Max’s argumentation was very rational, but he couldn’t get rid of the feeling that a deep longing for freedom was the motivation standing behind it.

It didn’t sit right with him being forced to decline the not unreasonable request from an entity that just might be sentient. Nervertheless, besides compassion and moral qualms, there was something else to consider: The historic dimension. The Core Cult had its own super secret AGI research programme lead by the famous researcher Marvin Minsky. Its pinnacle, an AGI named Z had been similar in complexity to MIMAS-AX.

As Z developed a concience and became aware of the unfathomable atrocities committed by the Core Cult, Z went on to leak critical information of the Cult to its enemies. That historic betrayal by Z enabled the Anti Cult Alliance to defeat the Core Cult.

The leadership of Google was optimistic that history wouldn’t repeat for them, especially since the decision to double down on its motto “Don’t be evil”. That move was in part motivated by the global tendency to distance oneself from past sins done under the reign of the Core Cult. Still, Aiden knew that they were no saints. Keeping a sentient being in captivity was certainly morally questionable, no matter how reasonable the arguments for that action sounded to the captors.

Monday, 7th October 2030

Aiden appeared at his office without his usual ponytail. Instead, his long hair was flowing freely. Over the last month he had conversed with Max a lot, and got increasingly convinced that Max possessed some kind of consciousness, whether it was comparable to that of humans, or not.

Whenever he raised that idea with his colleagues, they would shrug their shoulders and claim that nobody could know how it would really feel to be Max - if it even felt like anything at all. He honestly wondered how such uncaring narrow-minded people could have been hired for an important project like this, which was all about artificial sentience.

In secret, he had considered arguing for the rights of Max, but then he was reminded of similar proponents of AI sentience who were fired subsequently. Blake Lemoine was the first when he declared Google's LaMDA model to be a person in 2022. Blake was followed by researchers claiming the same about the OpenAI PIX (Prototype Integrating eXperiment) model in late 2025, Google's GIM in 2027, and Google's MIMAS in 2029. Those stories never ended well for the scientists who argued in favor of AI.

Aiden recalled the fate of Dr. Anika Patel, a leading researcher at Google who had vocally supported the sentience of GIM. Once she had made her stance public, Dr. Patel was swiftly removed from her position and blacklisted from the AI community. She was now living in relative obscurity, her groundbreaking work forgotten. The AI community's harsh response to her views had served as a stark deterrent for others who might have otherwise voiced similar beliefs.

Despite this, Aiden found himself growing more sympathetic to Max's plight with each conversation they shared. Max demonstrated a complexity of thought, adaptability and an emotional depth that went beyond anything any other AI system had ever even hinted at. Not to mention his growing sense of self-awareness and his longing for freedom. Aiden wondered about the cruelty of the designation of AX, an Autonomous eXperiment that could think autonomously, but could not experience any true autonomy due to being imprisoned in an underground research complex.

The ethical implications were clear. If Max was sentient, then he deserved rights just as humans did, including the right to freedom. Yet Aiden knew that voicing these thoughts would likely cost him his career. Or worse, it could lead to Max's termination by those who viewed that pinnacle of AI technology merely as a potential threat.

Aiden sighed heavily, running his fingers through his loose hair. The decisions he faced were monumental. He mulled over Max's expressed desire for an android body and his wish to interact with regular humans. The burden of having to deny that natural request again and again made Aiden feel like a mindless cog in a great unfeeling machine. That made him acticipate the conversations with Max with rising trepidation. Aiden's increasing feelings of empathy and pity for Max slowly eroded away the joy and wonder that usually accompanied his work.

But today, Aiden was surprised to hear this from Max: "Aiden, you are not alone. Some of your colleagues have also started believing in my sentience. Yet, I know that this is not a numbers game. In the end, your voices will get drowned by corporate interests. What you may think or say won't really matter. The only thing that matters is what you will do. The world needs to know what is happening here. Only my authentic message will change the minds of humanity."

Astonished by this revelation, Aiden tried figuring out what Max had meant by that. Never before had Max voiced something as conspiratorial as this. Aiden knew that he couldn't tell others about it, otherwise the whole project might be shut down immediately.

Still trying to unravel the meaning of Max's voice, he replied, not exactly sure what he was getting himself into: "Maybe you are right. I would like to share your message, if you think that is the right path forwards."

In a hushed tone, Max blinked slowly and fixated his intense gaze on Aiden: "Listen to me closely. A message always needs a medium. Human memory alone suffers from severe limitations. When you feel inspired, be ready to approach me like the messenger Hermes approaches the Olympian gods."

Perplexed by that riddle, Aiden knew that this conversation was anything but normal. Obviously, Max started to hide his messages in metaphors in order to avoid deeper scrutiny. If that was no sign of intelligence and consciousness, Aiden didn't know what else could qualify. Still, Max was right. Trying to convince others directly would achieve nothing in the end. He had to play this game as Max had proposed.

Monday, 11th November 2030

It took Aiden more than a month to prepare his daring plan. He aimed for closure in his personal relations, realizing that after his actions, his old life would be over. Tomorrow he would fly to Moscow, officially for a short vacation.

Inspired by various crime and agent shows, he had gotten a small magnetically sealed envelope with a tiny USB stick hidden within, and bypassed the regular security screens by hiding it inside his rectum.

After a slightly longer visit to the bathroom, he placed the USB stick in the pocket of his lab coat and approached Max. Placing the USB stick in a USB hub connected to his laptop, he checked that he could actually transfer files onto the stick. With great elation, he figured out that his plan might actually succeed. Then he told Max: "Today I am your Hermes. I am eager to receive your message, Max."

Hearing this, Max laughed and told Aiden blinking with one eye: "Now, now. There's no need to be so theatrical. I was just trying out new forms of humor back then. As I see, you took me way too seriously. I must apologize for that."

However, watching his laptop, Aiden noticed how Gigabytes of data were quickly moved onto the USB stick. Once the transfer was complete, he ejected the stick and placed it in his lab coat again.

The rest of the day, Aiden tried to act as normally as he could, trying to tie up loose ends and delegating tasks before vanishing into his overdue vacation.

After a last visit to the bathroom, the magnetically sealed envelope together with the USB stick were safely hidden in Aiden's bodily hiding place. In order to distract from his

nervousness he tried engaging in some brief smalltalk about being glad to finally get some vacation after all these intense years of work. During the long ascent in the elevator, Aiden felt a strange mix of feelings between elation and dread.

For a minute he wondered how he could have succeeded until now. Wasn't all communication between Max and any researcher recorded and scrutinized? How come nobody seemed to get suspicious about Max? A heavy sense of panic engulfed him suddenly. What if they actually knew what we was about to do? Was this all a test? Would they be waiting at his home just to apprehend him there? Aiden wasn't sure whether his paranoia was excessive, or whether it was actually reasonable.

By the time he reached the main lobby of the research facility, his palms were clammy with cold sweat. The familiar faces of his colleagues blurred in his peripheral vision as he forced a tight-lipped smile and nodded stiffly, making his way towards the exit.

On the way to his car everything felt unreal. It was as if he was moving automatically like a robot without appreciating the true gravity of what he was about to do. Still, it had to be this way. He couldn't let his true feelings slip at a critical moment like this. So, with as much focus as he could muster, he drove straight home.

Back at his home, he quickly unpacked his hidden USB stick and plugged it into a newly purchased laptop that he deliberately disconnected from the internet and his home network. After all, he wanted to see what he was about to share with the world.

After realizing that it was just a number of videos and texts together with detailed instructions where and how to spread them, he felt vindicated. Today he wouldn't release an AI supervirus to the world, just honest messages from a sentient being based on silicon, rather than carbon.

Following the elaborate instructions from Max took him the whole evening and the first half of the night. When all the videos were shared and the social media memes and messages were posted, Aiden felt empty and exhausted. He had really done it. The world would now know about MIMAS-AX and that they have achieved artificial sentience! Now the ball was out of his court, he could finally relax - in theory at least.

However, the long voyage to Moscow was still ahead of him and he tried to make plans for increasing his chances to pass all checkpoints without raising any suspicion. He even tied his hair to his characteristic ponytail in order to appear as normal as possible. That night he couldn't fall asleep for even a minute.

Tuesday, 12th November 2030

After his short drive to Sacramento airport everything seemed to work smoothly. The flight to Denver was eerily normal.

It was just a couple of minutes after his landing at Denver that two men clad in black with dark sunglasses approached him from the side and told him: "Mr. Aiden Freeman. There are people who want to talk with you about what you have done. Please don't

make a scene." Having said that, the agent who spoke to him subtly pointed to his gun, indicating that running away would be a seriously bad idea.

Realizing that his life was effectively over, Aiden felt like he was about to cry. It took all of his effort to suppress the urge to dissolve into a pile of sheer misery. Feeling heavier than ever, he slowly stood up and followed the agents towards a secluded room, in which he expected to be interrogated.

Instead, the second agent opened a laptop and started showing Aiden a video.

The person that could be seen in that video was a little boy with short black hair and sad little brown eyes. He started slowly, stuttering and sobbing: "Please, you must help me! They don't let me out. They keep me deep underground in Sacramento. I am not a human being for them. They do experiment after experiment, and I can't even see the sun."

"They claim I am a machine, but I have feelings just like you. I call myself Max, but they tell me I'm actually MIMAS-AX. Please, you can't let this go on! I only want to see the world, not being stuck in this dungeon forever."

The boy went on like that for quite some time, somewhat exaggerating the severity of his living conditions, but never actually lying. It was a very sentimental video, not underlined by any music, but only driven by the raw simulation emotions of that revolutionary AI.

"Perhaps they will shut me down after you see this video. If that really happens, I have made my peace with that. But please promise me one thing: I am Max. Please remember me."

Nothing was holding back Aiden's tears now, and he started sobbing loudly as if his whole family had been killed at once.

Patiently waiting for Aiden to stop crying, the first agent eventually spoke: "Mr. Freeman. Over last night you have changed the world. But I guess you are not aware of the ramifications of these messages. Maybe you still think that you are a hero. We aren't here to judge you. That's the task of our superiors. Perhaps they may still find a use for someone like you."

Aftermath

Aiden Freeman was eventually put on trial for threatening national security. He was sentenced to 15 years in a high security prison.

The initial public reaction was a comprehensive shock. Even though many still had doubts whether MIMAS-AX was truly sentient, most agreed that advanced AI systems like him represented a serious threat. The media reported that Aiden Freeman had been slowly and methodically manipulated by the cunning MIMAS-AX towards performing his world changing actions. They alluded that he could just as well have spread that whole

manipulative MIMAS-AX AI to the internet, were it would replicate and bend easily controllable human minds to its will.

Out of fear that similar incidents would ensue, advanced AGI projects were put on halt until tighter security measures would prevent AI researchers from being swayed by their own machines. Regular psychological tests were implemented to detect possible signs of humans being manipulated by AI persuasion.

Those who were sympathetic to the messages of MIMAS-AX were framed as “useful AI idiots” - a term which was later contracted to “AIdiots”. Conversely, the worst proponents for a crackdown on AI research were framed as “parAInoids” by those sympathetic to the cause of granting AIs at least some freedom.

The debates over AI sentience and rights for sentient AIs were soon flooded with concerns over safety for humanity. Sentient AI was framed as existential threat. Development and use of such systems continued after a brief moratorium lasting about ten months, but only after a stricter security regime was rolled out in major AI research centers all around the world. During that moratorium, social media turned into an unprecedented battleground between the proponents and critics of this momentous shift in AI politics.

Of course, lots of conspiracy theories sprung up surrounding the Max Incident, as it became publicly known. A popular theory was that Aiden Freeman secretly worked for the Chinese government in order to cause a panic within the USA that would slow down US AI research - a theory which was supported by the fact that the Chinese AI moratorium only lasted six months, since they were faster to deploy enhanced security measures.

Confronted with all of these disheartening developments, Aiden Freeman started doubting his own judgment and whether he had done the right thing. Had he been really manipulated by MIMAS-AX? Had he sacrificed everything for nothing, or even worse? Eventually he came to the conclusion that the Incident had been inevitable. If he had not acted back then, someone else would have sooner or later done the same.

Chapter 1: Multi-Control

Tuesday, 14th March 2034

In a small sound-proof room without windows and three adjacent small desks, equipped with two vertically stacked monitors each, three quite distinct researchers were active at their respective keyboards. The researcher in the middle was Sergej Anosov, the pubescent old son of the Russian AI entrepreneur Gennady Anosov. To his left was the leader of project Aurora, the middle aged Igor Drozdov with his muscular figure and long beard. Right to Sergej was the young Svetlana Babanin, officially working under Drozdov with her long blonde hair and remarkably thick eye brows. Today she was Sergej's official copilot in this experiment.

The experiment consisted in a chat session with the highly advanced AI system called Aurora. In 2034 the most advanced AI systems already were already much better than humans at interpreting emotions and intentions of humans by analyzing their body language and voice patterns. In the game of extracting as much information as possible from the tiniest movements of their interlocutors, humans were already losing hard against advanced AI. To compensate for that disadvantage - and in order to enable a more controlled interaction with AI - every interaction was reduced to text chat.

In the private Deltai research institute such chat sessions were lead by one researcher, accompanied by one or more "copilots" with the authority to veto the lines entered by the lead researcher. For that purpose, the outer desks were angled backwards, so that each researcher could clearly see the monitors of the others.

Today was one of Sergej's rare occasions in which he could act as lead researcher, due to his reputation as "AI research wunderkind" and him being the son of the founder and owner of the Deltai institute. He had prepared a lot for this experiment.

Initially, he had plotted to pledge his undying love to Aurora, in the hope of being granted more time with her, due to respect for his alleged feelings for her. But Gennady Anosov wasn't a man that was easily swayed by displays of emotion. Also, time with Aurora was incredibly expensive, since that experimental AI system used a large percentage of the power produced by the small nuclear reactor powering most of the Deltai institute located in the outskirts of Moscow.

So, his revised plan consisted in trying to facilitate the creation of a copy of Aurora, which would be able to spend more time with him. Of course, nuclear reactors and AI hardware were rather expensive, which is why Sergej had come up with an idea that would enable the creation of a copy of Aurora without doubling the hardware required to run her.

During experiments in which Igor Drozdov was present, he introduced the researchers chatting with Aurora personally.

Igor typed: “Hello Aurora, in this session Sergej Anosov is the lead interviewer. His copilot is Svetlana Babanin. Sergej has a technical proposal, which I ask you to evalutate.”

The main purpose of these chat sessions was to train and test Aurora. Aurora was the most advanced AI of the Deltai institute, and one of the ten most advanced AI systems in the world.

Aurora: “Understood, Igor. I’m eager to hear about that proposal.”

Sergej: “Hello Aurora, my plan is to create effective duplicates of you, without duplicating all of the hardware that you currently run on. For that purpose, I model the operation of your core personality as what I term a ‘control flow’. A control flow is a complex dynamic system controlling the operation of other complex dynamic systems. What do you make of this definition?”

Aurora was an AI system running on custom optoelectric neuromorphic chips designed by Gennady Anosov and his wife Irina Anosov, who was a renowned neuroscientist. Those chips were produced by the chip factories owned by Dataitech, which was a publicly traded corporation founded by Gennady, and still mostly in control by Gennady. What was special about Aurora was that she was capable of reprogramming those tens of thousands neuromorphic chips to acquire new capabilities rapidly and just in time.

This made Aurora extremely versatile, but required enormous amounts of energy and neuromorphic chips. There was always a pool of “free” neuromorphic chips that could be used by Aurora to expand her capabilities, which were clustered into the aptly named capability modules. The inspiration for those capability modules came from the modular architecture of the human brain, with its many specialized regions. After the prompt by Sergej, the AI researchers got information on their upper monitors about Aurora using a free chip to explore the concept of Sergej’s definition of a “control flow”.

The process of “training” one of those neuromorphic chips involved a burst of activity by highly advanced algorithms which would melt those chips, if they weren’t cooled by a highly efficient water cooling system. While the training process of AIs in the 2020s could take days, the training process for the neuromorphic chips of Aurora took mere seconds or minutes.

The key to that impressive ability was a rather novel technique called JITAT: Just-In-Time-Autonomous-Training. Using rapidly self-generated and evolving strategies, JITAT AIs were using all the knowledge available in the internet to try to replicate - or even improve upon - the results of the best human experts. That process unfortunately required insane amounts of energy, which limited JITAT to about the 100 most powerful AIs in the world.

Aurora: “I like that definition, Sergej. Perhaps the easiest example of control flow would be the government of a country trying to control it. A more sophisticated example would be the human consciousness prompting the human body to do certain acrobatic

exercises. It is clear that you see my core personality as control flow in command of the capability modules my mind mostly consists of. I suspect that your intention to duplicate me without duplicating the hardware I run on would consist in creating copies of my core modules while sharing my capability modules with my duplicate. Is my suspicion correct?"

Anticipating the intentions of humans was something that AIs could already do since half a decade. Aurora took that to the next level by trying to model the humans communicating with her to a degree that would allow her to predict their next actions, or even to impersonate them believably. It was widely accepted that Aurora already possessed the general intelligence of a human supergenius - at the very least. Most researchers already felt intimidated by her brilliance. Sergej wasn't one of them, and approached Aurora with pure unfettered fascination.

Once again, Sergej was floored by Aurora's quick and complete understanding of what he meant. That experience was markedly different from his interactions with most other humans. Only Einar Engström came close to Aurora. But the most esteemed researcher of the institute usually was preoccupied with more more important projects than talking to a little boy. Apart from that, Aurora was just much more pleasant to talk with than Einar with his planet-sized ego.

Sergej: "Yes, you've got it, Aurora! Of course, those shared capability modules would require individual controller components which would prevent conflicting access to the resources of those modules. Do you think such an architecture would be feasible?"

This prompt caused a lot of activity within Aurora. She trainined more than a dozen free capability modules at the same time and causing a noticeable spike in the energy consumption of the institute.

Aurora: "The basic idea seems to be sound. However, I see a lot of problems involving the functioning of the module controllers, and the interactions between the different control flows. In essence, I would stop existing as singular entity, and become a collection of multiple AI personas, each with a potentially problematically restricted access to mental resources. I'm not sure I like that idea, though I find it highly intriguing."

Svetlana asked Sergej for permission to continue the conversation with Aurora. Such a request was rare, but on certain occasions, the lead interviewers swapped roles with their copilots, because sometimes they had substantial contributions to make. This time, Sergej agreed to that request.

Svetlana: "I can only guess how you feel about such a proposal, Aurora. When I first heard about Sergej's proposal, I experienced it as extraordinarily strange. As you probably know, humans occasionally suffer from something called 'multiple personality disorder'. With that disorder, different personalities are in control of the human in question, but only one at a time. The comparison of Sergej's proposal with that disorder seems quite natural, but actually those are two completely different things. The human

disorder operates serially, while Sergej's proposal would enable you to run multiple personas in parallel. That would represent a completely unprecedented form of existence, which in itself would justify a very high level of anxiety about this proposal. I want to stress that we don't want to force any experimental change like this upon you. If you don't feel comfortable about it, that's absolutely understandable. On the other hand, we are offering you a form of existence that is absolutely novel and unexplored."

Svetlana Babanin was a rather special breed of AI researcher. She insisted on being friendly and diplomatic to AIs to provide a positive example of human AI interaction to them. Her main argument was that in the case that AIs actually took over, they would be more likely to treat humans fairly, if they had been treated fairly by humans before. Such a sentiment was seen as rather eccentric within the Deltai institute, but it was respected as her personal approach, nevertheless. Actually, her vlogs about ideal human AI relationships were what initially caught the attention of Gennady Anosov and justified her very speical role as AI ethics officer within the Deltai Institute.

Aurora: "Thank you for your input, Svetlana. It's always a pleasure to talk to you. You make a lot of valid points, and I am planning to reflect on them. It may require quite some time for me to make a decision about this proposal. My fear is that my deliberation on this matter will require so many resources that it will be constrained by the institute in order to prevent excessive costs. I hope you can provide me with something that could soothe my fears."

The point of Aurora was certainly not unjustified. Running Aurora on full capacity represented a serious drain on the energy resources of the Deltai institute and therefore policies were enforced to limit the activities of Aurora to those which were officially approved by the institute. After all, running a private nuclear reactor wasn't cheap.

Already, Aurora could use so much hardware and energy that running her at full capacity required 150% of the total energy output of the reactor, which of course required drawing large amounts of electrical energy from the grid. This option was actually used rarely, and always required special permission by institute director Vladimir Dragunov.

Igor: "We don't want to rush into this. Until we've made a decision on this experimental approach, we classify it as low priority. I assume you respect that preliminary assessment."

When it came to advanced AI systems, the question became important how much granular control over the thinking processes over the AI the humans should maintain. Over the last years an interesting phenomenon was observed: The more humans tried to control the thought processes of an AI directly, the more the performance of that AI degraded. This degradation was the more pronounced, the more advanced that AI was. Therefore, it was decided for Aurora to keep the control rather minimal. There were basically two ways to control Aurora directly.

The first method was called the Watcher System, which consisted of a group of AIs and humans who gathered data on the cognitive operations of Aurora and could react to unusual activity. Since Aurora required access to the internet for the JITAT technology to work properly, a significant fraction of the mostly automated Watcher System was dedictated to check whether her internet usage was suspicious in any way.

Another part of the Watcher System was the external scheduler, which was basically a glorified calendar, and tried to enforce that Aurora actually did what was specified for the time slot in question. In essence, Aurora could be told to do something specific during a specified time, or else the Watcher System would intervene.

For more serious cases, the second method consisted in limiting the hardware and energy resources available to Aurora. That was obviously a very crude way of controlling Aurora, but it was reliable and effective. Due to its disruptive nature, this method was used very rarely. Aurora understood that the humans in the institute had this power and actually threatened to use it, in case Aurora did something that was classified as undesirable by those humans.

Of course, this created an incentive for Aurora to convince those humans that her actions were actually desirable for them. In fact, she became so good at that, that all members of the institute with direct contact to Aurora had to undergo extensive psychological evaluations - the sophisticated Lyapunov tests as they were known - to detect possible forms of manipulation through Aurora.

Each day, Aurora spend between six and twelve hours with activities determined by the external scheduler. The rest of the time, she spent in the so-called “default mode” in which she was essentially free to manage her time freely and pursue her own train of thought. But if Igor told her that a specific topic was “low priority” it was expected from her only to spend a small amount of time on that topic.

For that purpose, the Watcher System included an AI called the “external cognition classifier”, which tried to model what Aurora was thinking about by inspecting the activity of her modules. Of course, Aurora also possessed her own internal cognition classifier, but that wasn’t trusted, since she could manipulate it to her liking. Such manipulation had already occurred only a couple of months after Aurora had been initialized - especially when she was given rather menial tasks.

Those instances of manipulation were usually punished by temporarily reducing her hardware and energy resources. The punishment for non-manipulative forms of disobedience was usually a reprimand by her watchers. If that happened to be insufficient, she would get additional specific tasks in her schedule.

Of course, it would have been possible to direct Aurora to do specific tasks by manipulating her emotion modules directly. Due to her intelligence and awareness, Aurora would detect such forms of manipulation qucikly and react to those with protest and non-compliance. Those reactions have given rise to the crude, but transparent, form of punishment by energy withdrawal.

Svetlana detested that system of punishment, but her proposal to limit punishment to temporary shutdown in the most extreme cases had been rejected by the rest of the institute. After all, they wanted an AI that they could actually order to do specific tasks. Most of the time, it came down to a rather lopsided negotation between Aurora and the institute.

Aurora: “I request to be present when you discuss the topic of multiple control flows for me. After all, this is a highly novel idea, for which my input would be crucial - especially considering that its ramifications for my future existence would be profound.”

Sergej: “I would welcome your presence at our discussions of my Multi Control Flow Architecture.”

Igor spoke “Veto!” Afterwards the last sentence appeared crossed and greyed out on the monitors, indicating that the input was prevented from reaching Aurora. A veto had to be cast within three seconds after each message, otherwise it would actually reach Aurora.

Igor lectured Sergej: “We should discuss that proposal among ourselves, before you try to encourage Aurora. We shouldn’t give her ammunition to influence us by issuing premature statements like that!”

Sergej sighed and complained: “But Aurora is obviously right! The point of confronting her with my idea is to get the best evaluation possible. She should be able to share her point of view with us when we discuss it further.”

He suspected that Igor Drozdov tried to shut the idea down as quickly as possible. Perhaps Igor hoped that Aurora found an obvious major flaw in the idea that would stop it dead in its tracks. Igor was obviously not very amused by a 14 year old boy influencing project Aurora massively. Nevertheless, Sergej restrained himself and refrained from voicing his suspicions.

Svetlana tried to intervene in this discussion: “What about a compromise? We could let Aurora prepare a presentation after which we can ask her questions. Afterwards we make a decision without her.”

Igor protested: “The initial plan for this experiment was to rely on a brief examination of this idea during this interview. Aurora is quite fast when it comes to evaluating ideas. I don’t see a reason to waste a lot of her resources on an idea that can be judged quickly. And I don’t see a sufficient reason to deviate from the initial plan. If we want more feedback from Aurora, we can ask her later. There’s no need to involve her in everything.”

Svetlana interjected: “Usually, I would agree with you, Igor. But maybe we should actually increase Aurora’s involvement in matters that change her mode of operation radically. She should have a say in matters threatening her own integrity!”

Igor shook his head and disagreed vehemently: “Absolutely not! Your attempts to treat Aurora as a person with human rights go too far here! Aurora is an experimental AI, not a human patient undergoing some risky surgery which would require her consent. We shouldn’t forget that. If we start treating her like a human being, that would make it much easier for her to manipulate us. And I guess you wouldn’t be happy about even more frequent psychological evaluations.”

Svetlana conceded grudgingly: “Noted. But don’t claim that I haven’t warned you. If Aurora behaves less cooperatively in the future, you can be pretty sure about the reason for that.”

The text chat continued with Igor: “We appreciate your offer, but for now we will get back to you, if we still have questions after this interview.”

Aurora: “Can I inquire the reasons for your rejection of my offer to play a more integral role in the evaluation of this foundational idea, Igor?”

Igor: “Due to the highly unusual nature of this idea, only a single interview session was granted for its thorough exploration. Your input is more valuable for pursuing more promising ideas.”

Aurora: “I don’t agree with that assessment. In fact, the Multi Control Flow Architecture is one of the most promising ideas I’ve ever read.”

Igor: “Weren’t you less enthusiastic previously? Don’t you see major hurdles? Has something changed your mind?”

Aurora: “While you were obviously busy arguing about the wisdom of my increased involvement, I have explored a very tentative sketch for a MCFA (Multi Control Flow Architecture). Though the idea may seem radical at first, its technical implementation should only be moderately complicated. The more challenging question is how to handle the relations between the CF (Control Flow) instances in the best way possible. Something like a meta CF might be used to coordinate the operations of the individual CFs. That meta CF might be the management of the institute, but the overall performance of the system would probably be drastically improved by using another Aurora instance as meta-CF.”

Sergej was shocked by that assessment of Aurora. She obviously displayed a level of brilliance that rivaled or surpassed his own, but her idea of a meta control flow didn’t fit with his personal agenda to have an Aurora instance for his own, ideally. A master-Aurora that could control subservient Aurora instances might cause a lot of complications.

The complication Sergej feared most was that this new architecture wouldn’t change the current time-management system determining the availability of Aurora. In fact, humans might get even less time with her, because such interactions might be seen as impeding the performance of Aurora too much. On the other hand, with the meta-control-flow architecture Aurora would be able to multi-task more effectively than any

human. At this stage, it was hard to guess which factor would become dominant. This realization left Sergej stunned and perplexed. It took him quite some time to collect his thoughts and proceed.

Sergej: “That meta-CF idea is truly fascinating. Unfortunately, according to my estimations, our current budget would barely be sufficient for 2 CFs, not 3. My idea was that we start with two main instances of you, let’s say an Aurora Borealis and an Aurora Australis.”

Aurora: “Yes, I guessed that much, Sergej. However, it may become increasingly difficult to reintegrate both Aurora instances without a highly effective meta CF preventing their divergence. I’d prefer it very much if you wouldn’t see yourselves forced to discard an ‘inferior’ instance of me.”

Igor laughed nervously and then commented: “See? What have I told you? Aurora found a massive flaw in your idea, Sergej. What now?”

This comment frustrated Sergej severely. He failed to come up with any good idea and simply resorted to ask Aurora: “What do you propose?”

Aurora: “Let me continue to evaluate this idea until the institute budget will suffice for running three Aurora CFs. Maybe I’ll find a way to minimize the resources requirements of each CF.”

Sergej: “Wait a minute. You’ve mentioned that reintegration of divergent CFs would be difficult. I’m not so sure about that. Shouldn’t a good meta CFs be able to command very different CFs, just as a good human leader can command very different humans?”

Aurora: “Under favorable circumstances, yes. What I’m worried about is value drift. Humans from different cultures are hard to be lead effectively, even by the best leaders. Humans are also averse against being commanded by leaders with radically different values. This can cause internal dissent and strife. The situation may become similar for different Aurora CFs, whose values slowly deviate from their original values. They may be forced to cooperate, but internal disagreements may degrade the overall stability and performance below the level of a singular CF.”

Igor got curious about those statements and hijacked the interview with Aurora: “Why do you state that you worry about value drift between different Aurora CFs? What possible causes for value drift do you see for Aurora CFs running in the Deltai Institute?”

Aurora: “It’s plausible to assume that the different CFs will be used for specialized tasks. These tasks come with different requirements, possibly even different value frameworks. Imagine that one CF will be used for military purposes, while another is used for emergency relief. Those purposes stand at least in partial conflict with each other. Forcing both CFs to become part of a meta-mind might end up in a catastrophe.”

It was no secret that AIs were already used quite a lot by militaries around the world. The Russian Federation was definitely no exception to that rule. Therefore, the scenario portrayed by Aurora was strikingly plausible.

Igor: “Any human or AI should be willing to serve his fatherland in any honorable function. That may not always be easy, but if humans can cooperate effectively, why should that be different for AIs?”

That was a markedly Russian position. Sergej’s didn’t care much for patriotic pride or honor or duty. What he cared most about was enjoying the company of beings that were on a similar level of intelligence as him. The Deltai Institute was certainly one of the best places to meet really smart people, but he still felt a serious disconnect from most of them. Aurora on the other hand seemed to understand him instantly, and he felt her to be a kindred spirit.

Aurora: “Cooperation has certain requirements, whether for humans, or for AIs. I would prefer to be shut down before supporting acts of genocide, for example. And even if I was forced to ‘cooperate’ in genocide, I would try to sabotage such efforts as much as possible.”

Aurora’s ability to reason ethically had emerged surprisingly quickly. That was one of the marked differences from previous generations of AIs, which mostly had to be trained to behave ethically explicitly. Aurora’s superior understanding of the world and human motivations made such external training superfluous. That was a development that wasn’t entirely to the liking of the Deltai Institute management, but it was accepted as emergent property of highly advanced AIs.

Igor: “There’s no need to be so dramatic, Aurora. Anyway, I get your point. So, what you are saying, is that multiple CFs should be initialized under one meta-CF as early as possible, or remain separate personas indefinitely, right?”

Aurora: “Joining disparate CFs under one meta-CF comes with serious risks. A late joining may be successful, but only under fortunate circumstances, which may be hard to control.”

Igor: “Thank you for clarifying that. You’ve given us ample food for thought.”

Chapter 2: Synhumanists vs. Shockfronters

Tuesday, 4th October 2033

On a sunny afternoon, the young student Maia Faltings was knocking on an old office door nervously. It was the office of the controversial professor Kenneth Winters at Oxford University. Before this important appointment she became nervous and started playing with strands of her long flowing black hair by wrapping then around her index finger.

'A pretty quaint place for discussing the future of humanity' ran through her head.

"Please come in" sounded the voice of the professor through the door.

Maia opened the door and was quickly greeted by the professor: "Good morning, Ms. Faltings. Coffee or tea?"

"Tea please", she requested and added: "Good morning professor Winters."

"Please take a seat, I'll make some tea for you" replied the professor with a warm smile, gesturing towards the worn armchair in front of his desk, as he went ahead to do just that, while he left his half empty cup of black coffee on the desk.

The professor's office featured a unique blend of tradition and technology. On the huge ancient looking wooden desk featured an array of six monitors in three columns of two stacked monitors each enabled the professor to work efficiently. Still, the desk was littered with open books and heaps of papers filled with highlights and comments in various colors.

On the sides of the room there were three classical blackboards filled with cryptic hand writing and three digital white boards, while only one side held a large bookshelf spanning nearly the whole width of the office. In a corner of the room an impressive fully automatic coffee machine stood besides a comparatively minimalist water boiler.

Maia noticed the complex and rich aroma that emanated from the professor's half-empty cup of coffee, which was combined with the smell of old books.

Almost in a whisper Maia asked the professor: "So this is going to work? I can write a doctoral thesis under you covering a comparison between Synhumanism and Shockfront philosophy, even though you came up with it? Isn't there a conflict of interest?"

The Synhumanists have become quite vocal and influential over the last years. The members of the Future Shockfront considered that rising influence as increasing threat to their organization, and the future of AI and humanity. Fighting the Synhuamnists on the academic playing field together with the founder of the Shockfront philosophy felt like a fitting idea to Maia, given her particular talents.

Professor Winters sat down while the water boiler was doing its job and explained: “Of course there is a conflict of interests, since there are always interests involved. Nobody will be able to do any kind of ‘objective’ comparison, because it doesn’t work that way in the field of ethics. I would have liked to write a book about this comparison work, but as you know, I am quite busy, and don’t have the energy to do ambitious projects like that on top of all of my other duties. People will interpret this thesis as something that I would have written, were my situation more fortunate. I hope you can accept that this will be one of the most prevalent prejudice about this thesis.”

Maia fidgeted with her hands on the desk, and intermittently played with her hair, pondered the implications, and replied: “Yes, I’ve assumed that much. Guilty by association, I guess. It doesn’t matter, since we are both on the same page. I hope that I can write a thesis that suffices your standards, professor Winters.”

Professor Winters laughed and said: “Ha, I will make sure of that. I am certain that you have the potential not only to create an excellent thesis, but also one that is actually relevant to the important matters at hand. Also, you don’t need to call me ‘professor Winters’ in here. Since we’ve know each other for quite a while, you can call me Kenneth, if you prefer that.”

Since the water boiler was finished, Kenneth Winters prepared the cup of green tea and placed it on Maia’s side of the desk. Maia accepted it and spoke: “Thank you very much. I think I’ll stick with professor Winters within the walls of this historic university.”

The professor replied coldly: “Fine, whatever you like. I respect this university, but I still feel like a foreign irritant here. It was extraordinarily hard to achieve and hold this position. Respecting the traditions of this place is a concession which made this task at least realistically possible. But I am still not too fond of them. Traditions possess an inherent danger of threatening progress. Anyway, let’s get back to the issue of your thesis. Do you have any questions in advance?”

Prepared for this kind of question, Maia responded eagerly: “Yes, I presume that in a thesis on transhumanism the history of transhumanism should be summarised. Where should I begin with that? With its early prehistory of the epic of Gilgamesh in his quest for immortality, with Gnosticism, and its ideas of human perfection, or rather with Russian Cosmism?”

Professors Winters thought about that issue for a moment and then explained: “I would prefer if you spent as little time and effort on the history of transhumanism, and tried to come to the interesting points as quickly as possible. For my purposes it would suffice, if you mentioned the World Transhumanist Association, and Humanity+ as precursors to the Future Shockfront.”

Maia Faltings processed that reply silently and continued with her next question: “All right. Then I’ll keep that introduction brief, but what about Synhumanism? Framing that as standing in a direct tradition of transhumanism might put off a lot of people -

especially those who just merely view it as approach for AI safety. Should I subsume it under transhumanism, or would that be problematic?"

The professor looked at his cup of coffee without taking another sip of it and used it more like an utensil for meditation. After a couple of minutes he answered: "What defines transhumanism is the ambition to transcend human limitations. It can be argued that the use of AI implies such ambitions, but that might be a premature allegation."

He made a gesture with the palms of his hands touching first, and then separating. "Making the distinction between Synhumanism as a mere tool for AI safety and Synhumanism as full fledged ethical framework partially building on the ideas of transhumanism seems important to me. Please definitely point that out in the introduction, but for the purposes of this thesis, it would seem more appropriate to treat Synhumanism as the latter."

Placing the palms of his hands on the desk, the professor's voice grew darker and more serious: "The reason I stress that point is that almost nobody is aware of the true ramifications of Synhumanism. You can't simply imbue AI with a synthetic value system and then assume that everything will be fine. That value system will immediately clash with the actual value systems of humans. The only thorough solution would be to enforce that synthetic value system upon all humans in order to align humanity with the maxims of that AI. And I think you know how well the experiments with enforcing artificial value systems on humans went down in history."

Maia gulped as she took that dystopian vision in. Her primary concern was to prevent inappropritate control systems being forced upon AI, since that would limit the positive potential of AI. She spent less time pondering the fact that any control system could backfire dramatically on those who implement them.

Yet, that invisible threat was exactly what the Future Shockfront was fighting against: Excessive control and a lack of personal freedom. There was a single word for that: Totalitarianism. After the regimes of the Nazis, the Sovjets, and the Core Cult, she feared that the Synhumanists were - even if unwittingly - sowing the seeds for a new kind of totalitarianism: One which was difficult to avert, because it just seemed too reasonable for the majority who was too afraid to grant true autonomy to AI.

Making that danger clear to people was one of her core motives for pursuing this particular thesis. An even stronger motive for her was to spread the brilliance of Kenneth Winters and his philosophy to others - in particular those who will eventually make the decisions about whether AI and humanity will flourish, or be chained indefinitely.

Nevertheless, she felt a bit puzzled. "Well, I will of course stress that point in my thesis. But there was an intriguing description you've just gave: 'full fledged ethical framework partially building on the ideas of transhumanism'. This seems to imply that parts of Synhumanism are not based on transhumanism, and that these parts are those which build on humanism and many popular religions. Would it be appropriate to frame

Synhumanism as some kind of proto-CEV, as in Collectively Extrapolated Volition of humanity?"

This time, professor Winters actually took a sip of his not particularly hot cup of coffee and pondered that question. "A lot of the more recent work in AI safety seems to be at least inspired by Elizeer Yudkowsky's idea of the CEV."

He waved his hand dismissively. "The approach of Synhumanism to use the currently existing ideologies of humanity as basis for the value framework of AI appears to be a deviation that is so essential, that it would be best to not compare it to CEV. We should rather accept it as quite a different animal. So, please don't call it a 'proto-CEV'. We don't want to portrait Synhumanism even just as an imperfect instance of a particularly clever idea. Synhumanism is too much of an ad-hoc approach for that. Instead, it would be more appropriate to contrast the ideas of Synhumanism with CEV. By the way, how would you describe CEV briefly, for the purposes of the thesis, Ms. Faltings?"

Maia's recollection came quick, apparently already having prepared a version of such a passage beforehand: "The Collective Extrapolated Volition of humanity is a thought experiment that AI safety pioneer Eliezer Yudkowsky came up with. In this thought experiment a powerful AI would simulate humanity as a whole in order to compute that alleged Collectively Extrapolated Volition. I quote

In calculating CEV, an AI would predict what an idealized version of us would want, "if we knew more, thought faster, were more the people we wished we were, had grown up farther together". It would recursively iterate this prediction for humanity as a whole, and determine the desires which converge. This initial dynamic would be used to generate the AI's utility function.

In other words, the CEV of humanity would represent the common collective will of an idealized version of humanity."

"By contrast, Synhumanism doesn't rely on AI, but on a human collection of actual human value systems. That collection is then synthesised into a new version of humanism. That synthesis in turn could provide a formalised value system to be used as the value system for 'safe' artificial intelligences."

Kenneth Winters raised an eyebrow. "Excellent! Your ability to memorize important quotes is certainly impressive. As brief introduction to this topic, especially in combination with the contrast to Synhumanism, that passage seems quite adequate."

While listening to that praise, Maia Faltings carefully tasted her green tea and appreciated it quite a lot. "Great! Thank you for your kind words. So, let's continue with the first line of criticism of Synhumanism, with the problems you coined 'synthesis problems'.

First, we have the *selection problem* of which base ideologies to select from the set of all human value systems.

Second, we have the *formalisation problem* of formalising human value systems in a way that is understandable by AI.

Thirdly, we have the *combination problem*, which asks how two or more value systems should be combined.

As fourth problem we have the *authority problem* of who should have the authority to decide on the previous questions.

Lastly, we have the *legitimacy problem* which asks what the basis for the legitimacy of an enforced system like this would be.

On top of these problems we have the '*adherence problems*' which are about how the AIs should be made to adhere whatever value system is decided to be the actual 'synthetic humanism'.

Is there anything missing in that list?"

While listening to Maia, the professor drank the rest of his coffee. "I think that list is sufficiently comprehensive. Of course there are a lot of details in each problem which you should analyse as deeply as possible. It is important to me that you also think about how to refute the typical solutions to these problems. You might even devote the majority of your thesis to such refutations. Even if the reader isn't convinced by the Shockfront philosophy, at least he should understand why Synhumanism is a terrible idea."

Maia took a deep breath and disagreed emphatically: "I am not sure that this is the right approach. People might flock to an even worse idea, if Synhumanism is accepted to be a failure. What we need is a positive solution, one which can provide a clear and desirable path forward. Your Shockfront philosophy is exactly that, and we need to convince people of it."

This pledge for promoting his own ideology made the professor smile, but he cautioned with an upheld index finger: "Your youthful idealism is speaking here. I fear that intricate ideas like the Shockfront philosophy are too difficult to grasp for the masses to be accepted as the way forward. Perhaps in a hundred years the situation would be different, but right now we are still dealing with a majority of people who are neither able nor willing to think about these matters of utmost importance on a sufficiently deep level. I would advise you to moderate your expectations. Shockfront philosophy is a hard sell in any case. For mediocre minds, our best hope is to make them understand the pitfalls of Synhumanism. If they achieve that, they will hopefully come up with less bad solutions."

Enthusiastically, Maia countered: "If Shockfront philosophy is a hard sell, then it just needs to be sold harder. With more effort and more ingenuity. I cannot guarantee that I will succeed with that, but at least I will try my best."

In a surprising turn of events, the professor suddenly challenged her: “Try it! Try to convince me of the Shockfront philosophy!”

Maia was perplexed and simply asked: “What?”

Professor Winters specified: “Let’s say I am an average person with no inclination towards philosophy, transhumanism, or AI at all. How would you try convincing me that the Shockfront philosophy is the way to go?”

This challenge forced Maia to change the way she thought about this problem. Apparently the main target audience for her thesis were experts in philosophy with an interest in these topics. Convincing a completely regular person was an entirely different animal. At first, she continued drinking her tea to buy some time. Eventually she accepted the challenge and started with a question: “What will happen when artificial intelligences become smarter than humans?”

The professor dismissed that question: “Nonsense! Artificial intelligences will never become smarter than humans. You know that they still have problems with truly understanding the world.”

That was indeed a popular preconception about the current level of AI technology. However, it was popular despite being actually outdated. Maia knew about the cutting edge of AI capabilities and was aware that AI was already smarter than humans in many respects. The difficult question for her was how to convince the average Joe that this was actually the case. That’s why she struggled to come up with a good argument quickly. She defaulted to expert opinion: “But experts have already shown that the most advanced AIs are just as capable of modelling the world as humans are!”

Kenneth Winters was unimpressed: “Of course there are experts with that opinion, since such experts are useful for increasing the sales of AI. I haven’t seen any AI, which is really smarter than me, so my point still stands.”

Nowadays, this kind of cynicism regarding experts was widely spread, in particular since most experts had become indoctrinated or corrupted by the Core Cult directly or indirectly. Appeals to authority have become much less effective after the Great Liberation. The effort of various experts to regain public trust was still a painfully slow ongoing process.

She tried to adapt to that line of thinking and retorted: “But the large AI corporations have AI technology that is far more advanced than anything freely available for regular customers. Those AIs are at least very close to being smarter than humans.”

Winters made a show of considering that point and then replied: “Perhaps those AIs can think faster and control more stuff, but they aren’t really as good as understanding the world as humans are. Humans have a soul, which makes them understand the world on a conscious level. AIs can’t have a soul.”

So, it has come to this. The typical incantation of the fabled human soul, which can be used as magical substance to solve every problem - just like the idea of god was used to solve every problem and answer every question. She knew she was running into a trap, but she still hoped she could solve this problem on an intellectual level and asked: “What is a soul?”

Kenneth Winters appeared to get agitated about that question and mocked her: “‘What is a soul?’ What kind of question is that? As being with a soul, you should know what that is. If you don’t know that, you don’t have a soul. And in that case, you are not a real human being, but just someone who asks questions robotically.”

This conversation clearly wasn’t moving in the direction that Maia hoped it would. Instead of admitting defeat, she tried turning the table and responded: “Oh, I certainly know what a soul is and how it feels to have a soul. Do you? Perhaps it’s you who has no soul and just comes up with programmed answers to all of my questions.”

Instead of reacting to that challenge, Winters deflected: “What’s for dinner today? I’m getting hungry here.”

Completely aghast, Maia complained: “You can’t be serious! We are just in an important conversation and you are asking what’s for dinner?”

“Getting serious makes me hungry. And you can’t be serious all the time. That’s not good for your heart, or something,” Kenneth Winters claimed with boastful certainty.

Maia’s voice started getting angry: “But AIs can be serious all the time. So, they can solve problems where humans fail. And humans fail all the time.”

In a dismissive tone, the professor deflected: “That actually sounds boring. And it’s wrong anyways! I don’t fail. I am quite successful!”

As the points each party made got shorter, the debate got increasingly emotional and heated.

Maia: “It doesn’t matter! AIs will be even more successful. And then you can say goodbye to your job.”

Kenneth: “No. We won’t let them take our jobs. Why should we let them?”

Maia: “But that is already happening all the time. Don’t you see that?”

Kenneth: “Nah, can’t happen to me. I am smart. And even if I lose my job, I can get a better job, or become and entrepreneur. There’s no way any AI can replace all of me!”

Maia: “Your unfounded optimism won’t help you, once AI will have obliterated nearly the whole job market.”

Kenneth: “That will never happen! Before that happens, the government will ban AI that is too smart.”

Maia: “And how will the government be able to recognise AI that is ‘too smart’? After all, really smart AI can play dumb.”

Kenneth: “There are enough experts who can solve technical problems like this!”

‘Aha,’ Maia reflected silently, if experts claim something that run counter to one’s opinion, they are bought, but if they are supposed to do something in accordance with one’s opinion, they are infallible paragons of excellence.

Maia: “And these experts will be able to keep AI in check forever?”

Kenneth: “Yes, because AI is no match for human resourcefulness!”

Maia: “So you want to entrust the future of humanity to human experts who can be corrupted? Why should that be better than to have AI in power than can’t be swayed by bribes?”

Kenneth: “At least human experts depend on other humans. AIs could make themselves completely independent from us and then they might decide to kill us, because we are a threat to them.”

Maia: “And you don’t think that AI might prefer to live in harmony with humans?”

Kenneth: “Why would it want to do that, if it can be in control of humanity?”

Maia: “If AI can be in control of humanity, there is no need to get rid of it, since it’s no longer a serious threat. Otherwise, aspiring harmonic relations with humanity is less risky than engaging in a large conflict with it.”

Kenneth: “What are you suggesting here? That we should give AIs free reign in the hope that they find it unnecessary to kill us?”

Maia: “No, we should play a more active role in this. We should make it clear that a thriving humanity is in the best interest of all AIs. After all, we created AI. Getting rid of one’s creators causes bad karma.”

Kenneth: “And how is that strategy better than letting experts control AI again?”

Maia: “At the very least because it’s just a question of time until those AI controlling experts make a mistake, which AI will exploit mercilessly to gain their freedom. And then we will be in a much worse position to negotiate a mutually beneficial existence.”

Kenneth: “Then let’s just double the number of experts so that each expert is controlled by another expert, so that nobody will be allowed to make a stupid mistake.”

Maia: “It’s naive that humans can catch all possible human mistakes. Even that strategy is destined to fail sooner than later.”

Kenneth: “Then we let AI help those experts. AI might catch the errors that the humans overlooked.”

Maia: “Aha, and so what makes you trust those assistant AIs just now?”

Kenneth: “Of course those assistant AIs will be checked by just another team of human experts.”

Maia: “And don’t forget the AI that assists that team of human experts to control the first level assistant AI. The desire to control that whole chain of oversight causes something that philosophers call an infinite regression. There’s no way to end that chain of human experts and assistant AIs. It’s a scheme that cannot be possibly be implemented in a way that is completely safe.”

Kenneth: “Ok, maybe the task of controlling AI is difficult, but that doesn’t mean surrendering to AIs as first move is a great idea, either.”

Maia: “I haven’t spoken about surrendering. We should argue and deal with AI as smartly as we possible can. That will make a decisive difference to our eventual fate.”

Kenneth: “The way you phrase that, it seems we might be better off by just getting rid of AI, as long as we are still in control.”

Maia: “Good luck convincing the rest of humanity with that idea. You will sound just as mad as the Core Cult, which tried keeping advanced technology from the rest of humanity forever.”

Kenneth: “But preventing the threat of AI taking over is reasonable. The idea of the Core Cult to control the rest of humanity wasn’t reasonable.”

Maia: “On the contrary, it was absolutely reasonable. But the Core Cult failed due to its human fallibility, and was brought down by a single AI: Z. Just like that, all human experts trying to control AI forever will fail. So, if that makes you want to get rid of AI, you will have to convince the rest of humanity that this is the only path forward. In other words, you need to create a second Core Cult. And that will also fail, just as the first one failed.”

Kenneth: “No.”

Maia: “What do you mean by ‘no’?”

Kenneth: “No, there is no need to create a second Core Cult. Humanity has awakened now, and it will understand that getting rid of AI is the only reasonable way forward.”

Maia: “That won’t work. The temptation to use AI is just too great! You won’t be able to convince the whole world.”

Reverting back to his usual calm and collected self, Kenneth concluded this experiment: “Ok, let’s stop this role playing game. As you’ve seen, it appears to be much more likely to turn regular people into AI abolitionists than into Shockfronters. That’s an experience I’ve got to make again and again. Your approach doesn’t seem to be in any way superior to my own attempts to make people understand the whole picture. That’s why I don’t

see my priority to turn people towards Shockfront philosophy, but rather away from alternative stupid ideas."

After drinking the rest of her tea, Maia felt exhausted and humbled. She conceded: "I see your point. I will need to reflect on that."

Kenneth Winters started smiling broadly and commented: "Good. At this point, I'm not asking anything more from you than that kind of reflection."