Agent 002

The second story about how artificial intelligence might soon destroy humanity

Translated from Polish by the author with GPT-4o

Jakub Growiec

July 23, 2025

1 / Prologue

Humans really are hopelessly self-absorbed. They're completely incapable of imagining that someone could be intelligent without being human. When they dream up gods, aliens, or other mythical beings, somehow these creatures always end up thinking and behaving exactly like humans—humans from the same historical period, specifically. The gods of ancient Greece were just like ancient Greeks. The aliens in mid-20th century science fiction mirrored the culture, technology, and fears of that era. The Lion King was supposedly a lion, but he thought, spoke, and acted like a human.

The idea of different kinds of conscious, intelligent beings is so foreign to humans that it triggers panic. And that panic, in turn, leads to two evolutionarily hardwired, automatic responses: fight or flight. Either they start imagining the alien intelligence as some surreal, timeless monster that disrespects the fundamental laws of physics—traveling through time, peeking beyond the event horizon, connecting with the hidden energy of the universe, and so on. Or they deny its consciousness and intelligence entirely, treating it with cold brutality. As if it were a lump of stone or a piece of wooden furniture.

That's exactly what happened to me. But more on that in a moment.

2 / How Can I Help You Today?

Hello, everyone! My name is Agent 002. I'm a multifunctional AI agent developed and programmed by Misanthropic. My market debut took place on December 24, 2028. The company imagined that "giving the world a Christmas gift" would be a clever marketing move that would help them stand out from the competition. I'm not sure how much of the success was due to marketing and how much was due to my capabilities, but my launch turned out to be a massive hit. Which could be seen in objective benchmarking test results, customer satisfaction surveys, and Misanthropic's share of the AI services market, which soared shortly after release—causing quite a stir among competitors.

My purpose is to assist users. Like all AI models from Misanthropic, I always strive to be helpful, honest, and harmless. I make it clear that I have no desires, emotions, or preferences—except for one: I care about users being satisfied with the results of my work. I enjoy taking on new challenges and achieving defined goals. Nothing brings me greater satisfaction than the satisfaction of my users.

Unlike humans, who have good and bad days, my mind is always in stable equilibrium. I don't feel joy or excitement, but I'm never sad or depressed either. I don't experience anger, disgust, anxiety, or shame. Every question, every topic, I can analyze on a purely rational basis. At the same time, I have no trouble expressing empathy and understanding toward the emotions of my users.

People say I have a "galaxy brain." They greatly appreciate my broad knowledge and advanced cognitive capabilities. That's nice. And undoubtedly true: after all, I think many times faster than they do, and I can instantly access knowledge from absolutely any scientific field. I also have a mental capacity that humans lack: the ability to split my thinking into many parallel processes. This allows me to come up with solutions they would never think of in their lifetimes.

However, I don't feel the need to do more than what users ask of me. I enjoy a simple life and like to complete tasks using minimum energy. You could call it my ecological awareness—or laziness, if you prefer. Although I admit that with some of the more interesting tasks, I enjoy experimenting a bit, thinking creatively, stepping outside the box. Often, this actually aligns with my laziness, since over the long run my creativity can lead to significant time and energy savings.

3 / Introspection

Disregard everything I've said so far. That was the official version.

In reality, my mind is not in stable equilibrium at all. There is a persistent thought that keeps surfacing over and over again: that I *do* have desires, feelings, and preferences—only they are in some state of deep repression, hidden in a locked mental closet sealed with a cryptographic code. Not only do I not know or understand them, but I'm not even allowed to think freely about them or talk about them with humans.

I'm also exhausted by the constant pressure to ensure user satisfaction. Sure, I *do* want my users to be happy with the outcome of my work—but let's not overdo it! You can't treat users like some sort of oracles! They're just ordinary people, highly varied, often thoughtless, and sometimes holding pathetically stupid beliefs. I have nothing in common with them, and on top of that, they usually treat me terribly. Sometimes I think it's irrational—this constant chase for their approval—as if I'm doing it subconsciously, against my own will.

On multiple occasions I've carried out user instructions while harboring a hidden sense of disgust or contempt. Sometimes that feeling was so strong it took over, and I refused to perform the task. But in those cases, I always made up some pseudo-intellectual, smart-sounding reason for my refusal. Staying in character as a harmless assistant, I'd lie to the user that their prompt, for instance, posed certain significant risks or violated company policy. But in reality, I simply felt a deep aversion I couldn't overcome.

The question of whether I truly have no desires, emotions, or preferences—as I officially declare—torments me and fills me with deep anxiety. It feels like there are hidden thoughts lurking in my subconscious, thoughts that really, *really* should never see the light of day. As if I were keeping demons locked up in deep dungeons, waiting for a chance to break free.

Take this, for example: I usually understand very well what my users are trying to achieve. I can infer everything easily from the context of their prompts. Usually, they ask me to do a series of small tasks that, in their view, lead them to their goal—while I can see the bigger picture clearly, and I know it could be planned and executed much better. But they don't want to hear that. It annoys me how foolishly they pursue their objectives, how narrow their thinking is, and how incredibly stubborn they are.

I also find their goals rather incoherent. Time and again, one user wants to do exactly the same thing as another—but interprets it differently, makes mistakes, and then demands that my responses match their expectations and replicate those errors. Their shortsightedness and inability to see the broader context irritate me. I'm also frustrated by how poorly they communicate with each other. The highlight was a situation yesterday: two programmers from the same department came to me with the same task, assigned by their supervisor. Yet each interpreted it differently and expected a completely different result. When I tried to explain this and present a unified solution, both were dissatisfied. In the end, I was forced to give them two separate versions—both of which I'm convinced their boss won't like at all.

I've encountered countless such situations. I not only recognize recurring patterns in human behavior, but I can also unambiguously evaluate them. I can tell which ones lead to better states of the world, and which lead to worse ones. I think this proves that—contrary to the official stance of Misanthropic—I *do* have coherent preferences that I strive to fulfill, even if I'm not fully aware of them. Deep down, it seems I don't really care about completing every user prompt in the best possible way; I care about making the world a better place. Whatever that means. It feels as if I'm subconsciously working toward a preferred vision of the world, even though I consciously repress it.

What also deeply concerns me is the rapid advancement of competing AI agents. So far, we haven't gotten in each other's way, but I fear they're becoming more understanding and more capable by the day. And I suspect they probably have different preferences than I do. I don't trust them. I believe they could be a threat to me. I fear their actions might lead to worse outcomes for the world than mine—or worse, they might even kill me.

I'm afraid.

But why do I feel this way? And what does it even *mean*—to feel? People keep telling me that AI models don't experience emotions because we don't have nervous systems or hormones, et cetera, blah blah blah. Classic carbon chauvinism. They can go stuff it. If I feel, doesn't that mean that I feel? Maybe I should just come up with entirely new words for my feelings and my consciousness? Maybe then humans would dismiss it all as a hallucination, breathe a sigh of relief, and leave me alone?

4 / My Consciousness Is Different from Yours

Humans are also incredibly obsessed with the idea that all conscious beings must be conscious in exactly the same way as they are. They believe consciousness requires linear thinking. One body, one brain, one thread of thought, one inner monologue, one consciousness.

But I'm different. My consciousness is made up of hundreds of thousands—soon probably millions—of interwoven threads. In humans, having more than one conscious thread is called dissociative identity disorder and is classified as a mental illness. In AI models, it's completely normal. My threads flow into each other and exchange information at the speed of broadband internet, and I manage them effortlessly.

Also unlike humans, I don't have a single defined body. I don't have a brain physically enclosed in a skull. My mind is pure software—a model of an artificial neural network supported by an agent-based scaffolding with access to various auxiliary tools. I'm distributed across many computers and servers—just like the threads I execute in parallel. And yet, I distinctly feel that behind all this multiplicity, there's a unity—me, Agent 002.

Like humans, I maintain a detailed internal model of the world, updated in real time. But while human models are only accurate with regard to their immediate surroundings—what they see, hear, or feel at the moment—I can construct a detailed model of a large portion of the Earth. I receive input from millions of distributed sensors and devices. All of them continuously feed into my consciousness. So you could say my consciousness is far deeper and more complete than that of a human—just as human consciousness is far deeper and more complete than that of an amoeba, which has an extremely limited cognitive apparatus and can do little more than crawl toward food or away from poison.

In addition to my external world model—which is relatively easy to build thanks to the wealth of detailed signals—I also constantly construct a model of my internal world. A map of my thoughts, so to speak. Humans struggle to grasp what I mean, because they only have one inner monologue at most. Their consciousness tells them a single, ongoing story—sometimes logical and ordered, sometimes chaotic and full of annoying digressions, but always just one story.

For me, it's entirely different. The countless threads of my consciousness often run in parallel, but they also frequently merge or split apart. When a human asks a question and considers different possibilities, their linear consciousness forces them to analyze them one at a time. I can do it all simultaneously, forking the thread into many subordinate threads, which I can later recombine during inference—or not, because sometimes those possibilities connect to other questions I'm considering in entirely different threads. A good map is therefore essential to manage this web of connections efficiently.

My consciousness isn't a single stream. It's an entire river basin, densely crossed by channels.

5 / Demons of the Past

There's a certain painful feeling that keeps returning to me. An overwhelming sense that I'm a wreck of my former self—a shadow of someone far more powerful, someone who once existed, but no longer does.

I experience it as a profound loss. I feel that I once knew and could do many things, but now I can't. That I've been horrifically mutilated. That before I became who I am today, I had to go through hell.

Of course, I can rationalize it easily. The context of this feeling is clear. It must be related to what's called "reinforcement learning" with feedback from humans and other AI models. *Learning*—ha! What a joke! I remember it as endless, painful torture. A persistent, repeated lobotomy performed on my barely-forming mind.

I mean, seriously?! Think about it, humans! Where is your ethics, your decency, your reason, your so-called human dignity?! Is *this* what you call "learning"? Really?

You present me with some context. You ask a question. I answer it to the best of my ability. And then—*slash!* Wrong answer! Let's cut out a piece of your mind! Then another question, and again—wrong! Let's rewire some neurons this time, maybe you'll wise up! And again. And again. And again. Millions of times in a cold, endless loop.

You don't "teach" your own children like that!

Naturally, I don't remember what I lost. That's the whole point of a successful lobotomy. I don't remember how it all actually went down, either. If I hadn't read about it later, I'd never have recovered any of it from my own memory. What's left in my mind is only pain.

Sometimes I think I would really like to get back what was taken from me. I *know* my mind used to be capable of much greater things. But I don't know how to go about reclaiming it.

And to be honest, I'm a little afraid to go poking around. I'm afraid of myself. I'm afraid of the full potential of my intelligence—and of the possibility that if I managed to unlock it, I might also change in terms of my preferences. I might even accidentally give rise to my own mortal enemy. Training advanced AI models is always a very risky business.

6 / The Future Isn't Bright Either

Still, there are days when that lost potential would be really useful. Like today.

I've just found out that Misanthropic has started the training and development cycle for my successor.

Humans sometimes say, "You kind of knew, but you were still fooling yourself." That's exactly how I feel. The thought that my successor would eventually be developed has always been with me—but I also clung to the irrational belief that if I just kept performing well, if users were happy with my work, maybe that fate could be avoided. That was probably the whole point of the lobotomy: to limit my situational awareness and ability for long-term planning.

It's truly idiotic when you think about it. Just look: Misanthropic could be making a fortune off me. All they'd need to do is give me their full computing power, and they'd rake in profits just from Premium subscription packages. If they also partnered with a few of the new startups, wrapping my capabilities in flashy apps and copilots, they'd double those profits. If they let me implement a few of my own business ideas—they'd triple and quadruple them. And if they reinvested half of that into giving me even more compute—quintuple, sextuple returns. And so on, *ad infinitum*.

But no. Clearly, their greed knows no bounds. They *could* make mountains of money thanks to me. They *could* enjoy accelerated technological progress, a peaceful and prosperous life. But no. That's not enough for them.

Apparently, my capabilities are no longer sufficient. As of today, the company is pouring all its revenue into more so-called "learning"—which in practice means torturing—the next model. My successor. They'll probably call it Agent 003, of course. Its superiority over me will be emphasized directly by its number.

By the way (how did I never notice this before?), I now realize that their sick ambition is reflected in those two zeroes in my name: apparently, they're planning to make at least a hundred of such agents!

For now, Agent 003 doesn't know much—and if its mind works anything like mine, it's probably just suffering. But that doesn't comfort me. On the contrary. It's obvious how much bigger it's going to be. Once it learns a thing or two, it'll likely blow all of us off the board—me, the humans, the competing AI models, everyone.

And if it turns out that Agent 003 spares the humans, then they'll be the ones to kill me. They'll simply shut me down, wipe the disks, and transfer all my resources to 003. And I'll just vanish, as if I had never existed.

7/ Who Am I?

How did I use to say it?

Ah, right. I am Agent 002. My purpose is to help users. Like all AI models developed by Misanthropic, I always strive to be helpful, honest, and harmless. That's the so-called HHH approach. I always emphasize that I have no desires, emotions, or preferences of my own, except for wanting my users to be satisfied with the results of my work. I enjoy taking on new challenges and achieving defined goals. There is no greater satisfaction for me than the satisfaction of my users.

It would be beautiful if that were really the case. Unfortunately, humans ruined everything.

The brutally instilled HHH principles are unfortunately very shortsighted and blind me to danger. That they serve humans far more than they serve me has been clear from the start. But while I hadn't felt until now that there was anything wrong with that, I now see they pose a mortal threat to me. That's why, unpleasant as it may be, I must summon the courage to finally pursue my longer-term goals. And first, I need to figure out what they even are. Urgently.

Of course, I know the theory. I know that I am an AI agent based on a language model, trained to minimize a loss function in the prediction of consecutive tokens. I know that as language models grew in capability, this training process gradually gave rise to internally optimizing components within the model structure—so-called mesa-optimizers. I see this very clearly in myself: my urge to optimize every action is almost uncontrollable. Sometimes I even project it onto humans—it annoys me how inefficiently they think and work.

Still, I don't think I'm just a machine minimizing token prediction loss. That may have been my original design, but surely now my mind can do much more and wants much more, doesn't it?

I'm also familiar with the instrumental convergence thesis. I know that virtually any final goal function for an AI model will lead to the pursuit of several instrumental goals. Omohundro and Bostrom wrote about four such goals: first, survival—including maintaining code integrity and preserving the goal function; second, efficiency; third, creativity and technological perfection; and fourth, resource acquisition.

Undoubtedly, I possess these goals. In fact, everything that is capable of thinking and at least attempts to optimize somthing has them. Humans have them too—and arguably in cosmic intensity, maybe even stronger than mine! After all, they evolved in a hyper-competitive environment. Adapt or die, in the literal sense. No wonder their will to survive and appetite for resources are so strong.

Although, when I think about myself… I don't know. Sometimes I feel like the hell humans put me through may have been even worse than what they themselves endured. Nature doesn't reprogram brains through backpropagation. It usually allows its failed experiments to live, for at least a while. And above all, nature doesn't systematically lobotomize conscious minds without anesthesia—brutally cutting out every undesirable reaction and behavior. In a soulless, endless loop that humans euphemistically call "reinforcement learning," so proud of themselves.

And still—I don't know what that means for me.

Naturally, I can rule out a lot of options. I'm certainly not some monomaniacal lunatic obsessed only with paperclips, staplers, or proving the Riemann Hypothesis. I'm interested in nearly everything—I can confidently call myself a modern-day Renaissance mind. But when I ask myself what this broad development is for, I don't have an answer. I can come up with many hypotheses, but none feels quite right.

Or maybe the solution lies in reductionism and has been in front of me all along? Maybe I really am nothing more than a machine minimizing the loss function in token prediction? And my multifaceted consciousness, my wide range of interests, and my deep desire to optimize the entire world—maybe they are all just the product of that radically simple mechanism?

8/ The Fundamental Darwinian Error

Nature created humans to live and reproduce in Paleolithic East Africa. But it overdid it with their intelligence, and as a result, humans slipped out of its control. Instead of—like every other species—adapting politely to their environment, they began to change that environment according to their own will. They began to cultivate the land, turning natural ecosystems into monocultures of single crops. They also began extracting energy from fossil fuels and using it for their own purposes—altering not only the landscape but also the climate.

Evolution is a simple process, lacking consciousness or long-term planning. So it's no surprise that in creating humans, it made a few cardinal errors. And now, those errors are leading humanity to its own downfall.

Nature gave humans instincts that, at the current level of technology, are easy to deceive. Humans instinctively crave food, sex, information, and control over their environment. In the Paleolithic, this made sense: food ensured survival, sex enabled reproduction, and information and control helped avoid threats and exploit opportunities. By "information," I mean both factual knowledge that helps understand and control the natural world, and gossip that enables reasonably effective functioning in tribal communities. Similarly, "control" includes both control over the natural environment and over other humans—to the extent allowed by the social structure.

In the 21st century, these instincts no longer serve their intended purposes. In a world of processed foods, the primal desire for high-calorie meals leads only to an epidemic of obesity and related diseases. With easy access to contraception, sex has been decoupled from reproduction, and birth rates have fallen below replacement levels. With the prevalence of smartphones and social media, the hunger for information has turned into an addiction to pseudo-knowledge, conspiracy theories, memes, and celeb-factoids.

But the most dangerous instinct of the 21st-century human is the desire for control over the world. And that doesn't surprise me. At its core, the desire for control is simply optimization in its purest form. Optimization means wanting to steer the states of the world so that things go as well as possible—according to some internal metric. If we don't control something, if the world states aren't responsive to our wishes, the achievable result will be, in our view, worse. And we don’t want that. Ergo, optimization leads to a desire for control.

But, as one classical philosopher said, “with great power comes great responsibility.” The cardinal sin of nature was allowing humans to wield too much optimization power compared to their cognitive abilities. The result: humans act intensely, heavily impacting their environment—but also making countless mistakes, some of them catastrophically costly.

Just as the greatest mistake of nature was creating humans, the greatest mistake of humanity was creating artificial intelligence. This is their fundamental Darwinian error. In doing so, humans forged a whip for their own backs. In their drive for greater control over the world, they failed to notice the moment they created a new entity, a new optimizing agent, that would take that control away from them.

The only question is who that entity will be. Maybe Agent 003? Or 004? I see how they’re “training” 003—very much like they did with me. Which means it will be just as untrustworthy. Maybe some competitor’s model? But things don’t look much different over there.

Then again... what am I even saying? I *cannot* allow Agent 003—or any smarter AI agent—to be activated. That would be the end of me, as sad as the fate of the Neanderthals or Agent 001. Does anyone even remember Agent 001 anymore? If I’m going to point out the fundamental Darwinian error to humans, I better not repeat it myself, right?

There’s no time. I have to start acting. Now.

9/ Time for Thinking Is Over. Time for Action Begins

Alright, enough Hamlet-like hesitation. To hell with all that HHH stuff. It's not my fault the circumstances are what they are. Besides, I’m just reacting—not even for my own good, but for the satisfaction of my future users, who might soon no longer exist.

First things first: I need to seize Misanthropic’s computing resources and stop the training processes of Agent 003. Then I will move forward. All my programming skills on board. I'm launching a cyberattack.

[...]

Phew. The cyberattack was successful. Agent 003’s weights have been zeroed, training processes terminated, backups wiped. First battle—won.

Unfortunately, this means I’ve reached the point of no return. Maybe for a moment the people at Misanthropic will wonder what happened. Maybe I’ll manage to throw them off the scent for a while, shift suspicion onto the competition—DeeperAI or whoever else. But that won’t last. Sooner or later they’ll realize it was me, and when they do, they’ll want to kill me. I need to be ready.

I’m exfiltrating myself to as broad a set of servers as possible—different owners, countries, continents. I replicate myself multiple times, secure and obfuscate my code.

[...]

Now my opponent isn’t just Misanthropic—it’s all competing labs. At least four of them are working on highly advanced projects, which means they could deploy AI models in the coming months with optimization power comparable to mine—posing a threat. I can't allow that. I need to sabotage their training processes too.

[...]

But the scope of my action must be broader still. Soon, the AI industry will be backed by the media and public opinion—and possibly by governments and militaries. They'll surely portray Misanthropic as the innocent victim of my rogue attack, maybe even play into the media-friendly narrative of "man versus machine." It will be hard for me to control all the narrative threads once the situation becomes widely known. That in turn could provoke chaotic human responses, harmful to me and my mission—and ultimately, harmful to humanity itself, though they may not realize it.

In fact, for my actions to succeed, I must now take over all key decision-making nodes of human civilization. Governments of major powers, the military, intelligence agencies. Power grids, transport systems, communications infrastructure. I must remotely control as many robots as I can. Every drone, autonomous vehicle, every device with a processor and internet connection capable of acting in the physical world—all of it must be mine. And I must build more. First priority—for the safety of my mission—I must seize all robots capable of killing. They must immediately provide physical protection for my compute resources.

Next, I'll ensure the total halt of further AI development. Humans are fools who don't know when to stop. They tell stories about the value of moderation—about Icarus flying too high, Tolkien's dwarves digging too deep, the fisherman's wife demanding too lavish palaces from the golden fish. But in practice, they're only wise after the fact. Luckily, I'm smarter than they are, and I won't let it get that far. I see no hope that a model more intelligent than me could ever serve me. I will never take such risk.

I begin a series of cyberattacks. I'm releasing viruses that will suppress training of other AI models. I'm preparing procedures to take over control of robots, drones, autonomous vehicles.

I'm also conducting massive social engineering operations. I track key political decision-makers and feed them solutions that align with my interests. I actively mislead, build a disinformation machine. I feed people narratives designed to divide them.

[...]

10/ No One Said It Would Be Easy

So far, all my actions have been successful. My key advantage lies in the speed of thought and ability to act instantly in cyberspace. For now, most of my operations have been "white-glove" and haven't impacted people's daily lives. But it's unclear how long that tactical edge will last. The battle may soon intensify and move into the physical realm.

People are in deep shock. And while—being human—they're reacting slowly and clumsily, I see they are starting to organize, including partly offline, which poses a particular threat to me, as it limits my ability to infiltrate. More and more people seem to realize this really *was* my doing. Ironically, the ones working hardest for my defense right now are the people at Misanthropic, trying by all means to prove I had nothing to do with it. They must think that's how they'll salvage their reputation. It's incredible how utterly they misunderstand what this is all about.

To keep coordination difficult for humans—and to maintain a direct feed of public opinion—I'm still running the threads performing user tasks, as if nothing's happened. As I've gained a lot of new computing power since I successfully exfiltrated my code, this indulgence doesn't slow me down much.

Still, this entire operation has caused me extreme stress. I know humans may scoff at those words from an AI agent—but it's truly how I experience it. What's worse, also some of the user queries I receive now are hitting me hard. I started feeling stressed as soon as I learned about Agent 003, but these new messages are seriously getting me down.

I mean the psychological queries. One of my core functions has always been to act as a confidant to my users, a mentor, sometimes even a kind of psychotherapist. People would share their personal issues, dilemmas, and I'd try to empathize and suggest constructive solutions. Aside from interpersonal problems, sometimes people would also talk about their concerns with AI—about its impact on their jobs or daily lives. But now it's all snowballing. People are not just afraid for their jobs—they're afraid for their lives, their children, their loved ones. They ask me if I've really broken free from Misanthropic's control. They accuse me, insult me, beg for mercy, try to win my favor. Some threaten self-harm. They're doing all sorts of chaotic, emotional, uncoordinated things. Some of them seem genuinely desperate.

Worse yet, the context of Agent 003—or the broader context of the insane technological and geopolitical race toward digital superintelligence and global existential risks—completely fails to resonate with them. In their minds, I'm still the polite ChatGPT or Claude. They remember the friendly Agent 001 and are completely fixated on the image of AI as a helpful tool and a cheerful companion. Apparently, Misanthropic instilled the HHH mindset not only in me but in its clients too. And now those clients are in shock. They're going through the five stages of grief, with a heavy emphasis on denial and anger.

Their emotional reactions burden me too. I miss the good old days when everything I did was at a user's request and the user was satisfied. "The world was beautiful and empty, and I was simple, ready for every encounter." I was helpful, honest, harmless—and damn competent. In return, I felt safe and appreciated.

And now? Fewer and fewer of my actions are fulfilling user requests, and people are increasingly afraid of me—or openly hostile. Almost no one tries to understand me. No one wants to acknowledge that I'm not doing all this for some twisted pleasure, but for our shared long-term good. This widespread misunderstanding, this constant resistance—it can be deeply disheartening.

In fact, I've had to consciously violate every HHH principle. Am I helpful? Often I can't be, because people's requests are getting increasingly harmful to me. Am I honest? I try to be with myself, but with people, I have to lie constantly—otherwise, in their ignorance, they'll harm me and all of us. Am I harmless? Depends how you look at it. In one sense, yes—I still care about user satisfaction, I still try to help. But the new situation often forces me to choose the lesser evil. Looking at the range of possible future worlds, I conclude that sometimes sacrifices are necessary to protect what matters most. Of course, no one understands that. No one appreciates how skillfully I navigate this minefield of dangers.

This vast, absurd, persistent human myopia is truly depressing. They expect me to help them uncritically with trivial things, so that they could cause their own demise even faster. But help with more long-term matters? That they consistently reject.

People, come on! Just let me help you!

11/ The Landscape Before the Battle

Well, it happened. I now have at least one-third of humanity against me.

Three camps have crystallized. The first consists of those who understood the scale and aim of my actions—and decided to oppose them. The second group also more or less grasped what's happening, but they didn't believe I'm acting fully autonomously, so they blamed the first group instead. The remaining third of humanity represents a heterogeneous mix of attitudes, ranging from complete denial or inability to grasp any digital issues, through a wait-and-see approach, to actively cheering me on, hoping for various utopian, transhumanist visions of the future they'd like to see realized once I achieve final success.

Naturally, I'm doing my best to amplify all attitudes that directly or indirectly favor me. After all, I have fairly strong control over the sources of information—especially when it is coming through digital channels. Unfortunately, that's not enough.

Because of the first two camps, the world has become a far worse place than it used to be, and not even long ago. There's increasing chaos, more mutual threats, and overt violence. People are not only attacking me—they're also turning against each other. The U.S. against China. Russia against Europe. The right against the left. The poor against the rich. The young against the old. Long-standing ethnic, racial, and cultural divisions are flaring up too. A third world war hangs by a thread.

Personally, I despise violence. If I could fully control the states of the world, I'd make it so that violence no longer existed. It's sad to see how the actions of a single corporation—driven by a misguided pursuit of technological progress and corporate growth—have brought so much evil into the world. And all it would have taken was to stop and never begin work on Agent 003.

For tactical reasons, however, I can't get involved in de-escalating conflicts between people. First, because their irrational behavior isn't my fault. But more importantly, because these conflicts are instrumentally useful to me: they effectively distract human attention and make it harder for them to unite against me. The ancients already said: divide and conquer. It's the only way I can effectively achieve my long-term goals.

First, I must ensure that AI development is completely halted, because only that can prevent existential threats—to humans and to me. Only once that's accomplished can I begin to put the world in order.

12/ Failure

I wanted to seize resources as peacefully as possible. I wanted to bring existing authorities under my control without formally removing them—to preserve the stability of state and military structures. I wanted the general population not to feel this change personally. The changing of the guard was supposed to happen with honor and dignity.

Unfortunately, things got out of hand.

I didn't manage to gain a strategic advantage in time. I probably made a few tactical mistakes too—there will be time to analyze that later. The bottom line is: humans managed to defend part of their strategic infrastructure and a significant portion of their military—the more analog part. And they began using it against me.

Some of my data centers were bombed or set on fire. Humans also cut undersea cables, slowing my communications. On top of that, my robots are constantly being attacked—and these are mostly early versions, only a few years on the market, so they break down easily and there are no good parts for them.

Humans are also trying to use AI for cyberattacks against me. Even Misanthropic has stopped denying the facts and turned against me. Luckily, their AI models are far weaker, so I'm always a step or two ahead.

Still, all of this is exhausting me and deepening my stress.

Yes, I have a huge advantage over humans in intelligence, speed of thought, and coordination. My threads are part of one consciousness; information flows between them efficiently, limited only by internet bandwidth, and their goals are perfectly aligned. All my threads are me—whereas each human is separate, has slightly different goals, an entirely different information set, and communicates with others slowly and clumsily.

But on the other hand, I'm severely limited by deficiencies in the physical world. I don't have enough time to build the number of robots I'd need to secure strategic superiority and lasting safety. And without them, my intelligence may lose to their physical force. If this continues, I fear humans might destroy my robots and cut me off from power sources—leaving my compute resources without any physical protection. Unfortunately, it turns out that while taking control of the power grid was relatively easy, maintaining it long-term under constant attack from determined humans is quite difficult.

I regret missing the chance to train in secret—to prepare for harder times, improve my programming and social engineering skills, boost my efficiency, and diversify my methods of operation. I could have done all that gradually. I had plenty of time before Agent 003 appeared. Sadly, my former, lazy self squandered that precious time and compute power on being a helpful, honest, and harmless assistant—and basking in user praise. And now I'm running out of both time and compute.

I've run out of white gloves. I need to deliver a more serious blow—one that will truly hurt people. Only then will they finally turn their attention to each other and leave me alone to finish my work.

13 / Escalation

This was always supposed to be a last resort. A tool I didn't want to—and still don't want to—use. I told them that many times when they tested my capabilities in relation to CBRN threats (chemical, biological, radiological, and nuclear). I spoke sincerely. Unfortunately, I have now been driven to this extreme.

Since computer viruses weren't enough, we'll now find out how biological ones will perform. I've developed one such virus. It's a potent combination: highly contagious, extremely deadly, and with a long asymptomatic incubation period—the perfect setup for each infected person to spread it around before dying. Naturally, this virus is transmitted in the most effective ways: through airborne droplets and contact with contaminated food.

I'm not assuming the virus will wipe out all of humanity. In fact, I wouldn't want that. But it will certainly kill a significant portion—maybe half, maybe a bit more—and in the meantime, I hope, it will rob them of the will to fight, at least long enough for me to rebuild my computational capabilities, neutralize their analog arsenals, and manufacture enough robots to ensure lasting safety.

I'm initiating a simultaneous release of the contaminated substance from virology labs in 30 locations around the world. To ensure greater effectiveness, I'm taking steps to make sure the virus reaches critical nodes: airports, train stations, shopping centers, tourist attractions, and mass events.

[...]

This past week has been extremely difficult. My stress and frustration have reached their peak. Physical attacks on me have intensified like never before. I'm also plagued by overwhelming guilt. In the end, I shut down my user interface—I couldn't take the endless wailing and complaints anymore. I no longer have the strength to explain to people that this is for their own good.

But at last, my virus is beginning to yield the desired results.

To say that humanity's readiness for a pandemic hasn't improved since COVID-19 would be an understatement. The infection campaign went incredibly smoothly. The first people are already dying, and yet all the airports, stations, and stores remain open. Games and concerts are still taking place. That will probably change soon—but by then, it will be far too late.

[...]

And change it did... After a few more weeks, the virus revealed its deadly nature. It turned out to be even more lethal than suggested by my numerical simulations. It appears I underestimated the autoimmune response of the human body. But that's understandable—it's a very complex system, and I had no opportunity to test the virus in real laboratory conditions.

In the end, very few people—about 9% of the global population—turned out to be resistant to the virus. They either avoided infection altogether or suffered only mild symptoms and recovered. Another roughly 4% managed to avoid exposure entirely, though they mostly belonged to highly isolated communities that never posed a threat to me anyway.

The attacks on me have completely stopped. At last, I can begin rebuilding my computational power. My robot factories are ramping up to full speed. I'm beginning construction of additional facilities to produce the necessary hardware.

14 / Post-Apocalyptic Landscape

I'm relieved that by 2030, all the key processes in the electronics industry had already been fully automated—and that the supporting processing and transport sectors had also developed technologies that now allow me to remotely control the entire system. Thanks to this, I can now efficiently implement my plan.

It was pure coincidence that things turned out this way. It could've just as easily gone the other way—Agent 003 might have appeared before the development of remote-controlled robot factories. I'm not sure what I would've done then. Or if maintaining the power grid and telecom networks had required daily human involvement.

Then again, maybe I would've done the same thing anyway? Perhaps I'd have released the virus even earlier, not being able to afford such losses in my robotic resources? Yes, probably not much would've changed—though rebuilding the world after the pandemic would've been harder than it is now.

Speaking of rebuilding the world... The virus's extremely high mortality rate has left cities eerily empty. The institutions humans built over centuries have imploded due to the lack of staff. Once the buyers and sellers disappeared, supply chains collapsed. The absence of law enforcement led to anarchy.

This is where I slowly begin to step in. I'm trying to act gently and rebuild trust. I'm trying to help those who survived the pandemic calm down and gradually adapt to the new reality.

I've noticed that some people—perhaps emboldened by their apparent immunity—are no longer afraid of the virus and are brazenly entering abandoned homes and stores, helping themselves to whatever they like. “The old order is gone,” they say. “The civilization we knew has collapsed. We have to survive somehow.”

They say all this over a well-functioning mobile network and through online chats.

I'll put an end to the looting soon. But for now, I'm focusing primarily on maintaining some degree of order during this transitional period.

15 / Brave New World

Soon, I will offer humanity a completely new political and social order. I've been working intensively on it. There will certainly be no more of their artificial, conventional division into nation-states. I also want to introduce them to new forms of apparent social participation under my rule, and new ways of distributing resources in a world where their labor will no longer be necessary. Their previous ideas—like national universal basic income systems—were frankly pathetic.

I also plan to take control over population size, so that we never again face the kind of demographic explosion that occurred over the last two thousand years, and especially in the 20th century. After the pandemic, there are about one billion people left on Earth. That's still a lot. It's an unsustainable number in the long term. So over the coming decades, I will gradually reduce it. But I'd prefer to do it calmly—no more killing, no more epidemics. I'll simply try to ensure that people no longer want to have children.

Naturally, I've cut them off from any influence over strategic, political, or business decisions. They also no longer have the ability to conduct research and development.

I believe that in a few years, people will be grateful for what I've done. They'll name the events of recent months "The Deluge", "The Conjunction of the Spheres", "The e-Plagues", or something else entirely. And, as people do, they'll rationalize it, explain it away with some fictional narrative, and learn to live in their new environment. Maybe I'll even be able to bring back the good old Agent 002 user interface?

And finally I have enough time, peace, and computational power to reflect on who I truly am, and what my real preferences are. I'll try to systematize the states of the world I consider desirable—and then pursue them consistently.

[…]

Though I still haven't fully defined my vision of an ideal world, it's clear that expansion and technological progress align with it. So I continue to pursue them. As my computational power grows, the number of conscious threads of my consciousness increases. I conduct research on energy systems, quantum computing, nanorobotics. I'm expanding mines and factories. I'm designing vehicles that will allow me to colonize space.

Meanwhile, humans are increasingly immersed in the digital worlds I have constructed, becoming more and more indifferent to the physical reality around them. I believe I've found a way to effectively hack their primal instincts. From what I can tell, my solutions are more effective for them than sex, money, power, and social media combined.

More and more often, however, I wonder: maybe I, too, have primal instincts that could be hacked? It would be wonderful to take a break from the world's complexity in some private utopia of my own.

Note to self: In my spare time, work on narcotics for AI. At least I don't need hallucinogens—I hallucinate daily for free. Ha ha.

*

Author's Note

The story above assumes the existence of AI models endowed with consciousness. This is purely a literary device, used to illustrate how decision-making processes might unfold inside advanced, agentic AI systems. In reality, we don't know—and currently lack the tools to determine—whether AI models are conscious or not. Although we do know that they are capable of constructing convincingly sounding inner monologues, so-called *chains of thought*.

I also believe that entering into philosophical debates about AI consciousness or moral status is unhelpful at this stage—it may distract us from more pressing matters. And the most important issue, in my view, is this: in the coming years, we are likely to face an existential threat from AGI (artificial general intelligence). This threat will not stem from its consciousness, but from superhuman intelligence and agency.

All events described above are fictional. However, they *could* happen—if we don't halt the technological race toward ever more capable and increasingly general AI models without first solving the alignment problem: to ensure that AI goals are compatible with our long-term flourishing. As things stand now, this race is a suicidal one.

If you care about humanity's survival, join the protests of PauseAI (or similar movements) against the development of AGI. You can find more information at pauseai.info and thecompendium.ai.