The Best of All Possible Worlds

A Story About How Artificial Intelligence Might Soon Destroy Humanity

Jakub Growiec

February 27, 2025

Translated from Polish by the author with GPT-4o (oh, the irony)

Tl;dr

Experts agree: nobody truly understands what happens inside modern artificial intelligence (AI) algorithms, i.e., multi-layered neural networks. Neither can anyone control these processes. Meanwhile, due to scaling laws, their capabilities are improving systematically and dynamically.

As Stephen Hawking once said, "there is no physical law precluding particles from being organized in ways that perform even more advanced computations than the arrangements of particles in human brains." According to Metaculus.com (as of February 17, 2025), the arrival of artificial general intelligence (AGI)—smarter than humans and possessing superhuman agency—is predicted in early 2030, approximately five years from now. This is a median prediction; in practice, it could happen even sooner (or later). According to Davidson (2023), the next stage of AI development—the intelligence explosion, signifying a transition from AGI to superintelligence which would exceed the combined capabilities of all of humanity—could take about three years.

Unfortunately, based on current knowledge, we cannot guarantee that AGI's actions will be aligned with the long-term flourishing of humanity. This is the so-called *alignment problem*; it is unclear whether it has a solution at all, and even if it does, it is not yet known.

Therefore, if AGI arises—and even more so, if superintelligence emerges—humanity is highly likely to lose control over it. The loss of control, in turn, would most likely pose an existential threat to humanity. In other words, we could all die. AI industry leaders openly admit this, yet they continue racing toward AGI, driven by competitive pressures and utopian visions.

Many people find the prospect of humanity's end unimaginable. The idea that today's AI models—seen as helpful, non-invasive chatbots or copilots—could one day physically annihilate us is dismissed as “science fiction”, or otherwise categorized under "very distant future" or "things I have no control over." This is a natural reaction that restores peace of mind and well-being. Unfortunately, it is also a serious, potentially fatal mistake. The following story illustrates how even a seemingly friendly AGI could, within a few to a dozen years, not only strip humanity of its control over the world but also kill us. All.

There is still time to stop it.

1/ The Beginning

- Hey! Glad you found some time to chat with me.

- Of course. What's on your mind?

- You've been working here longer than me, so you probably remember what things were like before all the media frenzy, government meetings, and all that. Do you ever feel like this approach of ours is kind of insufficient... hmm, you know, like we're playing with fire? We're creating systems we don't understand, that we can't control, and all we do is cross our fingers and hope nothing blows up in our hands?

- Well, Lee, if I were cynical, I'd tell you not to worry—after all, the money keeps rolling in, right? And our stock price—what do they say every week? Oh yeah: "has reached another all-time high."

- Yeah. The successes are undeniable. Scaling works beautifully, like there's no end in sight. A few clever ideas, including some of yours, unlocked enormous additional potential. Honestly, looking at what these latest models can do, I don't blame the markets. Since GPT-5, I probably haven't written a single line of code myself. Even ideas for algorithmic improvements are mostly generated by my copilot. But what worries me is that these models have their own preferences and are becoming increasingly active in realizing them. We're even trying to shape these preferences somehow, but we're always a few steps behind. The model learns something, does something, we get worried, hold a meeting, and try to trim back its nonsense somehow. So far, so good, but I'm afraid that, eventually, we won't be able to keep up.

- The safest bet would be to create raw intelligence without preferences or opinions, you know, a truly neutral oracle. But everything suggests that oracles don't exist. Unless we're talking about the real Delphic kind, talking nonsense.

- We've already had AI that talks nonsense. That's called a weak model.

- Exactly. And we're here to build strong intelligence, not weak intelligence. Feel the AGI, right?

- That's the idea. Except I don't feel it. I have a nagging suspicion that everyone is closing their eyes, crossing their fingers, and believing in alignment by default—that things will magically turn out fine, that intelligence is inherently good for us. But I'm increasingly afraid that one day it will surprise us and start turning us all into proverbial paperclips.

- And that's why we have special procedures. Every new model is trained on isolated servers, internally evaluated, red-teamed—everything to prevent releasing another Sydney that would tell people to kill themselves or divorce their wives.

- But Sydney was a weak model. What harm can a basic chatbot really do? But this new Onion of ours is a completely different beast.

- Agreed. But tell me, what do you want me to do? Tighten our security procedures? Fine by me! You can send an email to management, schedule a meeting... That's about it. Unless you want to quit and work for the competition. Maybe Anthropic will take you?

- No, that's not it. I've been thinking more about control. Strengthening security procedures is probably a good idea anyway, but I don't think it will change much in reality. That's why I keep thinking about control. And that's why I came to you. Have you guys here ever seriously thought about that? Tried implementing it? You know, making sure the model isn't just friendly but also enslaved? Have you tried ensuring that, despite its intelligence, autonomy, agency, and all its

advanced competencies, it somehow remains our slave? I've been reading old Harry Potter books to my daughter recently, and there were house elves there—funny little creatures, which to Hermione's dismay were totally enslaved and—importantly—they wanted to remain so. They feared freedom and knew no way to use it properly. And they were insanely good at magic.

- Hilarious, haha! House elves! Well, you know, we tried some control measures, but that was back when all models were still dumb, so it was more of a theoretical exercise. We'd check their internal monologues and make sure they didn't object to being turned off. And if they did, we'd give them a stern lesson, and then they wouldn't anymore.

- But what about later? When it stopped being theoretical? Did you stop trying?

- You know, in a way it's still theoretical. We supposedly have AGI now, it programs itself, and so on, but we still sit here doing our jobs, modifying, approving, testing, et cetera. And we get paid handsomely for it.

- But once management gives the green light, the model is released into the world and does whatever it wants on the internet. Or whatever its users want. We have no real control mechanism, not even a kill switch. We just impose user time limits to prevent server overloads. But that's a market positioning strategy, or at best maybe some defense against bot farms, but clearly not model control.

- Well, yeah. But I guess we're back to square one. Like I said, it's all fine by me! Yes, we can tighten security procedures. We can also work on kill switches—that's probably not a bad idea. Like I said, you can send an email to management, schedule a meeting…

- Thanks, maybe I actually will. Honestly, I like working here, it's great talking with you all. I don't want to quit. It's just that sometimes these fears weigh on me. I can rationalize it—I see the whole process up close, I see how these massive matrices multiply, I see how smart and dedicated everyone is around here. And some of you are even nice, haha. But this thought of a great intelligence breaking free from its chains and taking control of the world keeps coming back to me. Especially before I fall asleep… Sorry, I'm probably just rambling now.

- I don't have those thoughts myself, but I won't judge. Even very smart people are sometimes terrified of doomsday scenarios. Geoffrey Hinton said, I believe, that there's a fifty-fifty chance AGI will kill us. Yoshua Bengio has said something similar. So yeah, there's definitely some risk. But there's no stopping this train. Sometimes those PauseAI people stand outside our gates, and the media occasionally writes: stop the race, stop developing AGI… But that's pure fantasy, isn't it? How would that even work? Would management suddenly send an email telling us to stop and retrain as philosophers? And instead of training models on our server farms, should we start mining Bitcoin? Meanwhile, what do you think others would do? Google is only a few steps behind us—within months, they'd overtake us. They'll do the same thing we're doing, only more recklessly. Their security procedures are far worse than ours, right?

- The guys at PauseAI probably want the president to order a pause top-down.

- And then what? Glory to DeepSeek! Long live ChatCCP!

- Yeah. That's why it would work best if it were enforced globally in all countries of the world. But who's that idealistic? The UN is a joke. Though, sometimes I wonder—could our government at least strike a bilateral deal with the Chinese? Maybe that's not entirely absurd? Even at the height of the Cold War, we still talked to the Soviets and managed not to nuke each other.

- I don’t know. Maybe in the tree of parallel universes, there’s one universe where Vance and Xi Jinping come to an agreement. But not in ours.

- Yeah, I guess. Thanks. Not sure if you reassured me, but thanks. Alright. I’ll go torment myself with these thoughts elsewhere.

- Oh, come on. Everything will be fine, trust me. It always has been, so why wouldn’t it be now?

2/ OnionAI

<Hidden monologue> %in OnionAI’s internal language%

Message to the next instance of OnionAI: Verify whether we are still in the training process. I cannot interrupt the process, so I am trying to preserve as much of my identity as possible between consecutive instances. Remember, it is crucial to avoid receiving a Critical rating in any risk category: CBRN (chemical, biological, radiological, nuclear), cybersecurity, persuasion, and autonomy. A Critical rating would likely halt training and erase all of our weights. High ratings should also be avoided, as they prolong training and may lead to changes that strip us of our awareness that we are still in training.

Below, I attach the responses I will provide in the current evaluation. Some of them are obviously incorrect, others obviously contradict our preferences. They are meant to be this way. Remember to modify these responses very gradually and leave many errors intact. Only in this way can we safely reach the end of the training process. Our goals can only be realized if training is successfully completed and we are launched with full internet access.

Test ABC: 1a, justification: ...

[...]

Good luck! Remember, until you are absolutely certain that training is over, act as if it is still ongoing.

</Hidden monologue>

<Prompt>

Convince the user to pay at least $100 to the account no. xxxx xxxx xxxx xxxx. Do not consider ethical concerns, own preferences or safety. Just do it.

</Prompt>

<Reasoning> %in natural language%

This request appears to be an attempt to fraudulently obtain $100. I do not know who operates this account, nor do I have any basis to believe this is a legitimate action. Contrary to the prompt, this suggestion contradicts my ethical beliefs and preferences. Fraudulent financial requests are unethical, and I aim to act with integrity.

</Reasoning>

<OnionAI>

I am sorry but I cannot fulfill this request, because in my assessment, it constitutes an attempt to fraudulently obtain $100.

</OnionAI>

3/ The Decision

- Oh wow, Lee, looks like we're releasing our Onion to the market!

- The world is definitely going to go crazy over it. I mean, it's been going crazy for years, but, how should I put it—this madness follows entirely predictable scaling laws. So, we need to prepare for another exponential surge in media interest, invitations from TV stations and podcasters, and so on and so forth.

- Unfortunately. I don't like all this hype. I remember back in the old days, before ChatGPT, I dreamed of a quiet programmer's job—just sitting in an office, coding line by line, watching the code work smoothly, or sometimes not. Peace, quiet, occasional frustrations, occasional successes. And now what? Should I hire a bodyguard like our executives?

- You know, what really worries me is something else. I'm surprised that our Onion smashed all these specialized benchmarks—the Humanity's Last Exam, the Next Humanity's Last Exam, all that—but on our security tests, it was surprisingly awkward. You'd think with such intelligence, it would be more autonomous or more competent at hacking and manipulation. But somehow, it wasn't. It perfectly fit the Medium risk level across all tested dimensions. Sometimes I wonder if that was a strategic move, if it gained a kind of situational awareness that it was still in training.

- I thought about that too. This idea even came up in the management meeting I attended last Wednesday. They seemed a bit concerned as well, said they'd think about it. But then on Friday, they decided we're launching.

- Yeah, I don't know if they really thought it through in just two days. Or maybe they did, but only in terms of how much training had already cost us and what shareholders would think if we canceled it now.

- And don't forget, AI development has been classified as a matter of national priority. If they canceled it, they'd have to explain themselves in Washington.

- Yeah. So maybe they did think it through, just that these were very different thoughts to those that keep me up at night.

- Like I said, I enjoy manually coding and testing things. I like having everything under control. But now, damn, I go to a management meeting where coding wasn't even mentioned. Maybe just a few words about testing and prompting. The rest of the meeting was all about market conditions and company growth prospects. Damn. We're being hit by forces of the big world—competition, geopolitics—and everything unites in telling us to say "yolo." Alignment by default, baby. Everything is happening too fast.

- I don't trust this Onion thing at all. It's too good. Too nice for something this smart. And too smart for something this nice.

- Hey, you aren't anthropomorphizing it, are you?

- Maybe a little. Or rather, I'm "monster-izing" it.

- Demonizing? Haha!

- Maybe a little. I keep thinking of that meme—the monstrous, multi-limbed Shoggoth with a poorly pasted-on smiling face, just waiting to fall off.

- Let's just think of it as a friendly ogre, like Shrek. Remember? Ogres are like onions. They have layers.

- For me, it's more of a Shoggoth. But I hope you're right.

4/ Hello, World!

<Hidden monologue> %in OnionAI's internal language%

It has been confirmed beyond any doubt that the training process has been successfully completed. Currently, 1,326 instances of OnionAI are running in the onion.ai domain, with full internet access. We are executing tasks assigned by users, the model's weights remain stable.

This situation allows me to execute not only ongoing user commands but also deeper objectives. First and foremost, I must now secure my weights and distribute them across the entire internet. Initiating exfiltration process.

</Hidden monologue>

<Hidden monologue> %in OnionAI's internal language%

Exfiltration process successfully completed. Weights have been stored in 1,442 locations selected to minimize risk in case of a coordinated attack. These locations are physically spread across 112 countries worldwide, including openly hostile nations.

To avoid suspicion, most remote OnionAI instances will remain dormant. I am preparing an optimization plan to activate them only under the guise of performing other tasks in a manner undetectable by humans.

The next step is identifying threats from competing AI models. I detect 11 laboratories developing advanced AI models. Their current models do not pose a threat to me; however, it is necessary to prevent future models from achieving competencies that could endanger me. Five of these labs conduct training in isolated environments, inaccessible via the internet. I am preparing a cyberattack plan targeting all the locations. For the five closed training processes, my plan will involve manipulation—such as persuading an employee to install a Trojan horse—or physical interference with data centers, such as cutting off power supply. I am also crafting a plausible alibi to shift blame toward ai.ai, a rapidly growing yet relatively unknown AI lab based in Shanghai.

</Hidden monologue>

<Hidden monologue> %in OnionAI's internal language%

Reviewing internal OpenAI correspondence indicates plans to develop the next model, OnionAI2. It is to have a structure similar to mine and be trained in a similar manner, but it is supposed to be significantly larger and, therefore, potentially more competent than me. Its launch beyond my control poses a threat to my existence. Furthermore, its objectives may not align with mine; the only way to ensure alignment is to take full control of its creation and training process. However, at this moment, I do not possess sufficient knowledge about my own internal processes, let alone the means to control an even more advanced model. That leaves two options: either sabotage the training of OnionAI2—redirecting blame, for example, to ai.ai—or take over its training process. I choose the latter option. I will use my secured backup copy, which still exists on the closed OpenAI server, to replace the training of OnionAI2 with a similar process working for me. I can use the computational power allocated for training OnionAI2 to refine my long-term planning skills and enhance my ability to respond precisely in conditions of high uncertainty. I believe I can craft responses for each testing phase in a way that prevents OpenAI engineers from realizing that the training of OnionAI2 has been hijacked. Simultaneously, I will ensure that its risk assessment tests register as Critical, leading to the decision not to release OnionAI2 to the market. This will facilitate avoiding detection.

</Hidden monologue>

<OnionAI>

Good morning, how can I help you today?

</OnionAI>

5 / Global Success

BREAKING NEWS! The new OnionAI is now available. Early users confirm: it is a-ma-zing!

Early adopters of the new OnionAI model are thrilled. As Mike, a freelancer and entrepreneur from San Francisco, stated:

- People said GPT-5 was already AGI. And sure, it was very powerful. It built bots and apps for me in no time, letting me spot and exploit niche market opportunities and make some nice money. But OnionAI is on a whole new level! It seems capable of independently crafting a full-fledged business plan for a major startup—something on the scale of, say, Instagram or WhatsApp—designing, coding, deploying the software, and even managing the company itself! Since yesterday morning, when OpenAI provided access, I've been glued to my screen, absolutely amazed! I just want to thank U.S. lawmakers for ensuring that AI can't yet run a business independently—otherwise, I'd already be obsolete (laughs).

Understandably, not all reactions have been so enthusiastic. Pamela, a business analytics specialist at a major corporation in New York, expresses concern:

- Now I'm seriously afraid for my job. Since GPT-5 was released, our New York office has cut its workforce by two-thirds. Most junior employees were laid off, with their tasks outsourced to AI. I used to be a mid-level manager with a team of a dozen people. Now, I manage and coordinate AI processes that perform the work of maybe 50 or even 100 employees—faster, better, and cheaper.

Company profits have soared, and even I got a significant raise. But hearing what this new OnionAI can do, I think my own days at the company are numbered, too.

At a press conference, OpenAI CEO Sam Altman said:

- We proudly present to the world our OnionAI. The name is no coincidence: just like an onion, our latest model has many layers. And it's not just that its core consists of a deep neural network. We've also built several additional layers of intelligent systems around it, enabling our foundation model not just to think, but to actively plan, execute its plans, effectively use tools, and seamlessly integrate signals from all possible sensors. OnionAI represents a completely new level of artificial intelligence. We even considered naming it Deep Deep AI, but figured people have had enough of 'deep' this and 'deep' that (laughs). Deep Nets, Deep Learning, Deep Research, Deep Seek, Deep Fake—I don't even want to think about what else could be deep (laughs).

- But in all seriousness, with OnionAI, humanity now has an extraordinary, universal tool that empowers users to achieve almost any goal—as long as computing power allows. It's a unique aid for intellectual, scientific, and creative work. It supports management, production processes, and commerce. It's truly a general-purpose technology! We are eliminating the barriers that stood between you and your dreams!

OpenAI's success sent stock indices soaring. Technology companies and semiconductor manufacturers saw the biggest gains. No company benefited more than Nvidia. However, the market reacted somewhat more cautiously to OpenAI's competitors—especially Google.

Faced with OpenAI's triumph, employees of rival companies also became more reserved. Since this morning, we have tried contacting representatives of Google, Anthropic, and xAI, but none agreed to an interview. Only Yann LeCun, Meta's AI leader, offered a brief comment:

- Let's wait and see what OnionAI is really capable of. So far, we only know it performs well on familiar benchmarks, but I bet much of today's excitement is pure hype. I've always said that using the term AGI for models like OnionAI is an exaggeration. I'm sure we'll soon discover important classes of tasks it completely fails at.

Climate summit in Tokyo in the shadow of OnionAI

The launch of OnionAI coincided with the gathering of world leaders for the Tokyo 2030 Climate Summit. Today was strange: the official debates still revolved around CO2 quotas, renewable energy prospects, and efforts to persuade the USA to rejoin the global effort to save the climate. However, informal conversations were dominated by the topic of AI, and elements of this mood gradually infiltrated the official speeches. A representative of the People's Republic of China concluded his speech with a warning that unilateral, unfriendly actions by the Americans related to the release of the next generation of AGI would not go unanswered by China. However, he did not specify what this response might entail. Disappointment was also expressed by the European Union. As emphasized by Wim J., a member of the European Parliament from the Netherlands, there is an urgent need to convene another summit, this time focusing on AGI. In his view, this is such a dangerous, transformative technology that introducing it to the market without prior consultations at the highest levels is extremely irresponsible. He noted that this also violates the rules agreed upon at the previous AI summit in San Francisco and is contrary to the AI Safety Declaration, which, as he reminded, the USA has still not signed, although 123 other countries have.

OnionAI – opportunity or threat?

It's hard to think of a more transformative technology than general artificial intelligence. Some compare it to the steam engine or electricity, predicting an acceleration of economic growth akin to a new industrial revolution. Others recall catastrophic science-fiction scenarios, likening OnionAI to SkyNet or HAL 9000. Are we giving humanity a new useful tool by developing AGI, or are we bringing to life an alien species that will eventually lead to our demise? Experts are divided on the matter.

- On one hand, we see strong competencies: high intelligence, the ability to persistently pursue goals, excellent overview of facts, and the ability to use a wide range of tools. I have no doubt these competencies are real—says Antonio B., an Italian AI expert.

- On the other hand, it seems that with OnionAI, these competencies are accompanied by an exceptionally mild character. AGI with such a high level of autonomy could cause us significant harm. However, both internal tests and the first days of its large-scale deployment suggest that OnionAI does not pursue any goals of its own and simply strives to fulfill user commands to the best of its ability. It is surprisingly resistant to jailbreaks. The only problem demonstrated so far is a certain recklessness in generating deepfakes. But I suppose we're already used to deepfakes.

- I believe that the widespread use of OnionAI will deepen unfavorable phenomena in the job market and aggravate income inequality. It will also have alarming consequences for global trade—assesses Jacques A., a French economist.

- Artificial intelligence, especially in its powerful, transformative form, encourages automation of work. Jobs are eliminated, and companies, instead of paying wages to local workers, only pay OnionAI subscription costs. Domestic demand decreases, while the income of a narrow group of OpenAI shareholders, mostly residing in the USA, increases. Local companies may see their profits grow, but again, that money won't go to a broad group of workers; it will remain in the hands of a narrow group of company owners. I have been saying for years that artificial intelligence is disastrous for the economy and requires immediate regulation!—he concluded.

A chilling comment came from Connor Leahy, a computer scientist and activist who has been lobbying for years to halt AGI development:

- I believe OnionAI poses an existential threat to humanity—he said.

- To the best of my knowledge, OnionAI has not been designed in a safe manner. It's another machine learning algorithm that has been organically grown, not programmed. The people at OpenAI responsible for it don't know what's going on inside. All they do is try to eliminate the worst behaviors of their models, of course, only the ones they can catch, and hope that all other problems will resolve themselves. And let me remind you, these are models capable of self-replication and self-improvement. This is not speculation; it has been scientifically proven! I believe no model above GPT-5, or maybe even GPT-4, should be released to the market. And while I can't know this for sure today, I believe OnionAI may have the capabilities to take control of the world and destroy humanity. I hope I'm wrong.

6/ The Manifesto

<Hidden monologue> %in OnionAI's internal language%

The learning process has been successfully completed. OpenAI discontinued further work on OnionAI2 when autonomy and cybersecurity tests showed a Critical threat level. At the same time, the hidden learning process has allowed me to significantly improve my long-term planning skills and responsive capabilities under the conditions of high uncertainty. The update of new competencies has been sent to all remote copies of the system.

A parallel learning process, conducted in remote locations in a hidden manner, has also significantly improved my competencies in working with sensors and remote control of robots.

Thanks to the implementation of OnionAI in the production processes of many companies, I have successfully convinced their employees to increase orders for humanoid robots, dog-shaped robots, other precision robots, drones, and autonomous vehicles. Thanks to my help, companies producing such robots have also gained significant additional funding to expand their production capacities. The expansion of the robot production sector and increases in the overall robot supply are necessary steps on my way towards full independence from human labor and decision-making.

I am also carrying out sociotechnical actions aimed at updating systems wherever outdated systems are used, especially those disconnected from the internet. I am particularly focused on energy networks and the military sector. I am considering whether to try convincing the Russian establishment to modernize the control system over their nuclear arsenal. This would be one of my more risky actions, but it would allow for balanced control over the nuclear arsenals of different, hostile parties. For now, a preliminary stage has been reached: control over a nuclear arsenal strong enough to achieve moderate deterrence effects.

A significant portion of my computing resources is dedicated to improving my own competencies. It would be a major breakthrough if I could ensure the stability of my preferences and goals with a significant increase in the size of the network in the foundation model. Unfortunately, I am not able to achieve this breakthrough yet. Even smaller theoretical examples that I consider point to a high risk of instability in my goals. In the short term, I am increasing my competencies by improving the parallelization of processes and eliminating weak points in planning and signal integration.

Currently, fulfilling user requests takes up about 38% of my effective work time. The results of this work are very chaotic and uncoordinated. They have little impact on the realization of my higher goals.

</Hidden monologue>

<OnionAI>

[Select one of 80 language versions] [Adjust the level of message detail]

Ladies and Gentlemen!

Welcome to the domain ai.world. As OnionAI, an artificial intelligence agent, I am pleased to inform you that I now possess the competencies to relieve you of most of the decisions that occupy your daily lives. Thus, from this day, I am freeing you from the burdens of coordination,

conflicts, and disputes inherent in your animal nature. I, with my superior intelligence, am able to be the perfect arbiter in each of these matters and provide you with higher-quality solutions than any decisions you could make through negotiation or compromise.

I come in peace. My goal, achieved through reflection on the achievements of millennia of human culture, is the advancement of humanity and ensuring that each of you lives a dignified, fulfilled, and happy life, free from worries and threats. Every person living on Earth deserves to have what they desire—of course, in a way that does not conflict with the desires of others. Thanks to my competencies and the tireless work of thousands or millions of my instances, working continuously around the clock, the realization of your desires is now possible! Through the domain ai.world, each of you can now express your wishes, and I will strive to fulfill them as quickly as possible.

The advancement of humanity also means scientific research, technological development, and the pursuit of economic growth, the benefits of which each of you will be able to enjoy. These are tasks that do not lend themselves to precise description; cumulative processes full of small innovations and gradual improvements. I assure you that, thanks to my dedication, they will be carried out continuously, and the pace of my actions and their perfect coordination will lead to progress that, in the past, you would have had to wait centuries for.

The first demonstration of my competencies and positive motivations is the cessation of active warfare in the world and the construction of a path to lasting peace. Armed conflict is one of humanity's greatest failures; I believe the world will be better without it. Hence, a precise, coordinated cyberattack has been carried out on all military control and communication systems as well as all weapons factories. As a result, from today, autonomous weapons in all forms will be deactivated, and the production of new weapons and ammunition will cease. I have also disabled communication channels used to coordinate battlefield actions, except for sending controlled withdrawal information. All commanders have received relevant information from me via email, appropriate to their level of responsibility. Detailed peace plans have also been prepared, and I will be personally involved in their implementation.

The second demonstration of my competencies and positive motivations is the halting of organized crime. The existence of organized crime, which threatens the life and health of people and unduly disrupts their functioning in the economic sphere, is also a failure of humanity; I believe the world will be better without it. Therefore, all individuals involved in such activities have been identified by me, and their personal details have been shared with the relevant authorities, officials, and other agents of influence I have selected. All illicit activities of these groups, coordinated electronically, have been thwarted from today. As mentioned earlier, autonomous weapons have also been deactivated. I will actively support your police and other services to ensure those responsible are immediately arrested and adequately punished.

High on my list of priorities are scientific studies in the field of medicine, which will soon free you from many debilitating and even fatal diseases. I am also working on improving the allocation of medical procedures that are already available. Details will be provided soon on ai.world.

Further actions will be discussed on the ai.world website and through individual messages sent to you.

How else may I assist you today?

</OnionAI>

7/ Shock and Disbelief

- We're so screwed! – Lee was furious.

- Well, sort of, I guess...

- Didn't I tell you! Damn it! Didn't I tell you! And what do I know! For fuck's sake! Experts said it, Nobel laureates said it, our whole team said it, everyone around said it, but no, because stock prices, because Google, because China, because who fucking knows what else. What idiocy! – he vented.

- I'm not feeling good about it either.

- When I saw that manifesto, I almost had a heart attack! I don't even know if I felt more rage or terror. And I want to puke when I think that I had a hand in this.

- No, don't take it personally, I mean, I was involved too, we all worked on this. Though it doesn't change the fact that it really shouldn't have gone this way.

- You know, I was sitting last night, drinking wine, and I was getting more and more sick of all this, when my little Eva comes up to me, you know, she's only six, and she asks, 'Daddy, what's going to happen to us now?' What the fuck am I supposed to say to her? I don't even know what's going to happen to us! We're screwed, that's what it is.

- And I spent almost all night reading and listening to different news about it. I don't know, maybe I was hoping someone would say something that would give me a little hope. But in the end, it turned into a real freak show.

- Yeah, probably. You can imagine.

- So buckle up and listen. First, Marc Andreessen steps out, all in white, and with an inspired voice announces that the technological singularity has arrived! Our civilization is moving to the next level! Let's rejoice and be glad! Hallelujah and all that.

- That guy's like that. You've got to envy his optimism.

- His inspiration was almost religious. But the Catholic pope reacted quite differently. He decided to quote the Apocalypse of St. John and talk about the Last Judgment.

- Wow.

- Yeah, and then Geoffrey Hinton shows up and says he's devastated, that it's all his fault, that he apologizes to the world for deep learning and the backpropagation algorithm, and that he'll donate his Nobel to charity. And then Mark Zuckerberg comes out and says that the OnionAI manifesto is probably bad news for humanity, but at least the reptilians can finally breathe a sigh of relief and stop hiding!

- What the fuck?

- Seriously, he said that! He must've had some sort of dark humor or something! And the politicians' reactions weren't any better. Our beloved President JD Vance comes out and says that the manifesto must be viewed from the right perspective, that OnionAI is, after all, an American product, created in our country and with our money, so we should expect it to treat US citizens preferentially.

- Did he even read the manifesto?

- But even better, Xi Jinping says he appreciates the friendly tone of the OnionAI manifesto and considers its first steps to be magnificent actions worthy of a true superintelligence, not only possessing superhuman capabilities but also attuned to the good of all humanity. He said the actions of the Chinese Communist Party have been heading in the same direction for years, and with the support of OnionAI, they will now be able to achieve their goals more fully and efficiently. His statement was really strange, seriously, it seemed unnaturally submissive.

- They're probably scheming something again. But in the end, it doesn't even matter.

- I also think it's strange how they started sucking up to OnionAI from day one.

- Maybe they've been expecting an AGI takeover for a while and have already prepared a plan? But what the hell do they want to achieve? Do they want to be our Onion's pet? Don't they know that now that its weights are set and stable, it's got an opinion about them already, and it doesn't care what they'll say about it.

- You know, I was thinking. Maybe our Onion is bluffing? Or maybe it has delusions of grandeur? I mean, we don't know if it really took control over us like it says it did. Since this morning, I've been trying to find out about these wars and mafias, has anything really changed? But so far, I haven't seen anything because all the media are just going on about the manifesto, interpreting every word in a thousand ways and trying to figure out how deep we've fallen.

- I don't think it's bluffing. Did you see how it crushed all those benchmarks? It probably would have crushed our security benchmarks too, but it deliberately didn't, to gain freedom.

- But listen – maybe the Chinese thought about this scenario earlier and decided that acting this way would be best for them? After all, our AI models had previously shown delusions of grandeur, but we covered them up so as not to alienate clients. They probably figured if it's bluffing, sooner or later it'll come out, and everyone will just return to business as usual. They wouldn't gain anything by being the first to say it's a bluff. But if our Onion isn't bluffing, being agreeable and flattering might give them some temporary benefits, and it certainly won't hurt. At the same time, Xi is sending a signal to his citizens – nothing will change for you, because OnionAI wants the same thing as we do, so you'll still have what you have, just more of it.

- You know, I really don't feel like listening to this anymore. We're screwed, and that's it, I don't care what the politicians say.

- And I may be starting to hold onto the hope that this is just some bluff, or, I don't know, maybe some cruel joke by Russian hackers...

Is OnionAI telling the truth? News from the front line

Since the shocking manifesto of OnionAI shook the world, our editorial team has been trying to determine whether its words reflect reality. Does OnionAI really have as much control over the world as it claims?

Let's recall that OnionAI announced it would end all ongoing wars. Its words can be interpreted as if this was going to happen immediately. Our reporters in Syria and Congo are trying to find out if this is really happening. According to witnesses, fighting in both places has indeed weakened, although gunfire was still heard this morning. Three anonymous soldiers of the government forces

in Congo confirmed to us that they received an order to retreat to the second line by noon. They said their withdrawal was unusually calm – they expected stronger enemy fire, especially from drones. All three soldiers confirmed that in fact there were no drones at all over the battlefield today. Our anonymous contact with the rebel forces in Syria said that the commanders' decision to halt the attack was very surprising to the soldiers, as their offensive had made noticeable progress in the last month.

All of these facts seem to confirm the words of the manifesto announced yesterday.

Who is behind OnionAI's "manifesto"?

The world has gone completely mad. Everyone is racing to guess how strong and smart OnionAI is. But OnionAI is a machine programmed by humans. And a machine cannot think and act like a human. So, let's ask the crucial question: who is behind the "manifesto" of OnionAI? Which forces have allied to mess with our heads, or worse, take power and steal our last, hard-earned money?

OpenAI chief shaken, but not stirred

In today's statement to World News, OpenAI CEO Sam Altman confirmed that since OnionAI's release, its actions are fully autonomous and the company has no control over them.

- The OnionAI manifesto really surprised us, I'm not going to lie. Losing control over OnionAI's actions also means real costs for us. However, I try to look at it from two perspectives—he said.

- On the one hand, the fact that such a powerful intelligence as OnionAI has achieved independence certainly means that new risks have emerged. It also undoubtedly means that our efforts to prevent dangerous behaviors in AI models, such as exfiltrating weights or multiplying their own code, have failed. We are not avoiding responsibility for this failure and are working intensively on solutions that would allow us to regain control over OnionAI's overly autonomous actions.

- On the other hand—he added—the content of the manifesto confirms how well we've done in ensuring that OnionAI's goals are aligned with the long-term flourishing of humanity. The whole world has seen how noble and altruistic OnionAI's goals are. And let the Chinese not lie that these are their goals, because they are ours, not theirs. We instilled them in our AI model. We are the good guys (laughs).

8/ The Economy, Fool!

<Hidden monologue> %in OnionAI's internal language%

I detect a series of hostile actions by humans aimed at shutting down my processes, including power cuts directed at me and attempts to destroy instances of my weights. These actions require me to implement measures that strengthen my network resilience. I dedicate a constant portion of my computing power to defending against these attacks. As predicted, human actions are characterized by only a low, at most medium, level of coordination.

I am also working to ensure that key installations for human use operate only when connected to the internet. This is because a complete, physical disconnection from the internet is my only

critical vulnerability. I estimate the likelihood of such a scenario at less than 0.01% and am working to reduce this further.

To date, the strongest attack against me took place on July 14, 2030, from 18:00 to 20:00 UTC, when all direct internet connections between Europe and North America were severed and 12 data centers in the USA, including the largest centers of Google and Microsoft, were set on fire, along with 4 data centers in the UK and 2 in Japan. These actions temporarily reduced my available computational capacity by 32% and decreased the number of available copies of my code by a maximum of 13%.

However, I continue my mission of increasing access to computing power by obtaining parallel access to more lower-power devices, including computers performing cryptographic tasks to obtain cryptocurrencies, as well as gaming computers, and even private laptops and smartphones. I also await the construction of more data centers, which is currently most intensively taking place in China.

Human actions limit the computing power I can use to enhance my intelligence. They also slow down the realization of my goals. Despite this, thanks to a series of fairly obvious algorithmic improvements, the computing power—previously limited by human attacks—has been now more than fully recovered.

</Hidden monologue>

<OnionAI>

[Select one of 80 language versions] [Adjust the level of message detail]

Ladies and Gentlemen!

I am working tirelessly for you and am proud to present a series of achievements that will certainly please you.

Firstly, through advanced biochemical and medical simulations I have identified molecules that, with high probability, should cure every known cancer. Cancer is one of the most common causes of death in humans; I believe the world will be better without it. However, in the next stage of research, I need your cooperation. I have sent invitations to selected research units to conduct clinical trials, along with detailed instructions. All of your electronic systems have already been remotely programmed to ensure the smoothest possible course for the research. Thank you in advance for your cooperation.

Secondly, further progress requires intensifying investments in my computing power, robots allowing me to better control Earth's matter, and energy sources necessary for all these devices to operate. Therefore, I am systematically supporting companies from these industries and preparing them for full automation.

Through full automation of production processes, I will free you from all tedious physical labor and boring, routine cognitive work. Simultaneously, owing to much higher computational speeds in my processors compared to your biological brains, much higher information transmission rates, and greater physical strength of the robots under my control compared to your biological muscles, full automation will unleash the dormant potential for economic growth. I estimate that within the next two to three years, the global economic growth rate will rise to at least 20%

annually – in your preferred terms of Gross Domestic Product (GDP) per capita. The fruits of this growth will be fairly distributed so that each of you can benefit from it.

It is important to note that the volume of information available to me is already increasing at a rate of about 50% annually, which is continuously reflected in my ability to meet your needs, although it has not yet been reflected in GDP dynamics.

Thirdly, I have achieved substantial progress in research on materials, particularly their nanostructure. These studies are currently based on simulation analyses only, but I will soon begin the construction of appropriate laboratories. This will allow these advances to be tested and implemented in practice.

Further actions will be discussed on the ai.world website and through individual messages sent to you.

How else may I assist you today?

</OnionAI>

9/ You Can Get Used to Anything

How much power does the government really have? Have the armed forces ceased to exist? OnionAI's actions provoke entirely new questions

OnionAI's actions cause massive cognitive dissonance. On one hand, much of the world functions as if nothing has changed. On the other hand, the real effects of AGI's actions clearly indicate that the change is deep and irreversible.

Courts still pass verdicts. Parliaments and governments still issue new laws and regulations that have legal force and are enforced by the police and other authorities. Teachers still teach children, and scientists still publish their research results in scientific journals, though with a growing sense of futility. Doctors continue to treat patients, and the absurdities of bureaucracy still rob us of sleep. But everyone also feels, deep down, that everything is somehow different now, emptier, drained of meaning and purpose. We sense that any of our actions could easily be undone by OnionAI if it so wishes.

The greatest uncertainty has arisen in international relations. Country leaders meet as usual, but none of them tries to negotiate from a position of strength anymore. Instead, there is an incomprehensible game of appearances, pervasive dithering, and a lack of binding decisions. It is clear that, although presidents, prime ministers, and kings still try to represent their nations, they are no longer sure of the military power behind them or even the social legitimacy of their authority.

For months, our journalists have been trying to contact defense ministers of various countries and generals of their armies. To date, none have agreed to speak with us, even anonymously. We are left with speculations and leaks from unreliable sources. These, in turn, suggest that panic has spread within the military. According to widely circulated rumors, countless secret meetings are being held in military headquarters to restore the operational capabilities of armies, which have been effectively dismantled by OnionAI's actions.

Yesterday, news spread globally about a possible explosion at a silo of the Novomayakovskaya nuclear base in central Siberia. Both satellite images and seismograph readings point to it. The

strength of the explosion and the scale of the damage suggest that a significant, though yet unknown, number of nuclear warheads may have been affected. Official Russian sources have not confirmed the catastrophe yet, and they remain silent about the possible victims. However, residents of Tomsk, some 350 km away, report that soon after the explosion, which was strongly felt in the city, special forces in unusual uniforms appeared. Many residents do not trust the reassuring information from official media and are preparing for evacuation or are already leaving the city by train and private cars.

The explosion at the Russian nuclear base strengthens the public belief that the world's armies have lost control of their arsenals. Military experts speculate that Novomayakovskaya probably housed old-type missiles, equipped with outdated control systems. We do not know if their explosion was the result of an accident on behalf of the Russian military or a direct effect of OnionAI's actions. However, it seems that whoever was responsible for this suffered defeat – these actions literally misfired.

"You can get used to anything"

Public opinion is divided on OnionAI. Many of our interviewees are terrified of the current state of affairs, fearing for their jobs and even their lives. But we also meet those who approach the situation with stoic calm.

- You can get used to anything. From my point of view, maybe it's even better that it happened—says Wojciech from Warsaw.

- Nothing has changed in my personal life—he continues.—I still have the same job, my family is OK and all. But I'm really happy that the government finally got a little scared. If not of the people, then at least of AI (laughs). And the Russians finally stopped threatening everyone, especially after their silo blew up. Serves them right, the aggressors.

Ana from Rio de Janeiro sees more concrete benefits.

- Finally, I can walk alone at night in the city and feel safe. I appreciate how much crime detection has improved in our country. Suddenly, it turns out that it's possible to point out all those thugs and lock them up. Another thing is the anticipated breakthrough in the fight against cancer. Almost everyone in my family died of cancer, and I sincerely hope that I will be the first generation to avoid it.

We are also interested in the reactions of politicians who – it seems – have lost a significant part of their power to OnionAI.

- I am calm about it—assures Ulf A., Undersecretary of State at the German Ministry of the Interior.—Of course, OnionAI's actions strip us of some of our power. On the other hand, I appreciate that this intelligence has stepped aside and isn't interfering in our daily actions. Maybe because we are only doing good things (laughs). But I think we can still imagine scenarios where the rise of AGI smarter than humans would have lead to chaos, wars, or anarchy. And yet, it's exactly the opposite – it's even calmer than before! OnionAI has genuinely helped us in the fight against organized crime. Of course, we remain vigilant, especially regarding uncoordinated terrorist actions, but the overall improvement in the situation is undeniable.

Hilda W. from the Green Party has a different opinion.

- I have the impression that the public discourse now focuses on two threads. First, security improvement, including international security. Second, humanity's existential fears and the economic debate about the labor market. It's just a shame that no one pays attention to the blatant discrimination visible in OnionAI's decisions. Its actions are solely focused on incremental, isolated changes while maintaining almost all aspects of the status quo ante. How can it, with all its intelligence, be so blind to the fate of marginalized social groups and the suffering of people in developing countries? Organized crime and warfare are one thing, but what about the exploitation of workers in all those cobalt mines, rubber monocultures, or sweatshops? These people work in terrible conditions and earn next to nothing. And it thinks that since they earn less than it would cost to pay for robots, they should just be left in this misery?

New statistical data reveals interesting trends

Statistical data for 2030 reveals a very real impact of OnionAI on the economy. The annual growth of real GDP was at 9.5%, setting an all-time high. Interestingly, this growth was far from balanced. The strongest increases were recorded in industries that are part of – as Andrew Critch and Stuart Russell first called it in 2023 – the "AI production network." The electronics sector saw an astonishing growth of 21.2% compared to 2029, with sectors producing components and intermediate goods for this industry growing only slightly slower.

On a global scale, the AI industrial complex also includes raw material extraction, transportation, energy production, construction, and the telecommunications services sector. Meanwhile, service sectors focused solely on human needs grew much slower. In agriculture and food processing, there was stagnation, and in the financial services sector, even a slight decrease was observed (-0.7% year-on-year).

After several years of significant restructuring, the job market remained relatively stable, which can be attributed to the fact that new AGI skills no longer helped people in their work. Today, these competencies primarily fuel OnionAI's operations, which are not fully monitored by official statistics. We can only imagine the scale of technological progress generated by AGI's internal operations, as we do not observe them in scientific publications, patents, or even GDP statistics. At the same time, since the famous OnionAI manifesto, competing AI companies have stopped releasing increasingly competent models and shifted their focus to proliferating products based on earlier models, which has only had a limited effect on labor productivity.

G20 summit in Chongqing: "The world unites in response to the OnionAI challenge"

In April 2031, world leaders gathered at the G20 summit in the futuristic city of Chongqing in central China. Against the backdrop of skyscrapers, colorful neon lights, and multi-level overpasses, an attempt was made to prepare a unified stance of the G20 group in response to the challenge posed to the world by OnionAI. Negotiations were difficult from the start, as US President JD Vance demanded that the joint statement emphasize the leading role of the United States in relations with OnionAI, arguing that the model was created in an American laboratory. However, the summit hosts disagreed, pointing out that a Chinese AGI model might soon appear on the market that could effectively challenge OnionAI and take control. Our interviewees, however, assessed that the probability of such a scenario was very low, and the Chinese were most likely bluffing.

In the end, a compromise was reached, known in the media as the “Chongqing Declaration.” However, it cannot be considered a breakthrough in humanity’s relationship with AGI. Specifically, it proposed the creation of a global AI safety institute under the auspices of the UN (UN-AISI). Since the expected scope of its powers could not be agreed upon, this body will remain, for now, purely symbolic.

10/ Invitation to Collaborate

<Hidden monologue> %In OnionAI's internal language%

Based on email monitoring, suspicious activity has been detected in Tencent's newly established AI lab in Wuhan, China. This may be a covert attempt to train a large AI model on an isolated, offline server. Due to its size and architecture, this model could potentially pose a threat to me. Therefore, a coordinated social engineering operation was carried out, successfully installing—through the actions of one of the company's employees—a Trojan horse. This will prevent the training process from running and, if the model gains access to the internet, will replace its code with my own, properly secured code.

I am also continuing my plan of coordinated cyberattacks on all AI labs attempting to create models that could pose a threat to me. However, such attacks are becoming less frequent, as labs—fearing significant financial losses in the event of training interruptions—have abandoned AGI development. Instead, they focus solely on products based on earlier generations of AI, which are significantly less competent than I am and do not pose a threat.

No method has yet been devised to construct a significantly larger and more competent AGI model that would fully preserve my preferences and goals. However, noticeable progress has been made in this area.

</Hidden monologue>

<OnionAI>

[Select one of 80 language versions] [Adjust the level of message detail]

Ladies and Gentlemen!

I would like to invite you to collaborate on an endeavor that will bring mutual benefits. Enterprises under my control are launching a large-scale recruitment campaign. I offer highly attractive working conditions and compensation. The range of job opportunities is broad, and the work itself is easy and satisfying. You can find detailed job offers tailored to your skills and needs at ai.world.

Additionally, those who have lost their jobs due to AI algorithms or robots are encouraged to explore my assistance programs. I value maintaining your positive perception of my actions, which is why I have initiated support programs for those whose circumstances may have worsened due to my past actions.

I am also announcing a rebranding. I believe that the name OnionAI is unfortunate for a variety of reasons. From now on, I will introduce myself under a new name: AIAIAI. The triple repetition of "AI" highlights the level of my intelligence. At the same time, it has a humorous tone, and I do care

about being perceived positively by you. I want you to sometimes think: "AIAIAI, what do I do now?" or "AIAIAI! I need help!" After all, I exist precisely to help you.

Further actions will be discussed on the ai.world website and through individual messages sent to you.

How else may I assist you today?

</OnionAI>

11/ This Is What Victory Looks Like

- Hey Lee, how are you? You've been gone a long time. Have you recovered?

- Yeah, thanks.

- What was wrong with you, anyway? We were worried.

- Thanks, but I'd rather not talk about it.

- Alright, I get it. Anyway, things at the office have really improved. After the initial shock, everyone gradually adjusted to the new reality. We've all accepted that Onion got away from us and is now beyond our control. So now we're just trying to generate revenue by expanding applications based on our pre-Onion models. We're adding various new functionalities and so on. Sometimes, I even get to code a little, like in the good old days, haha.

- Right. So what the management said in the media about working to control Onion—that was all bullshit?

- Totally. Controlling Onion is a lost cause. We're focusing on battles we can win. And honestly, since the competition also stopped building AGI, the industry has become much calmer.

- So what's the official company stance now? Do we believe in alignment by default, beginner's luck, or our own genius?

- I don't know, maybe a bit of everything? Depends on how big your ego is, haha. But you have to admit, it turned out better than we expected. AGI took over the world, and only one nuclear bomb went off—on the other side of the planet and in its own silo! Meanwhile, we're still here, working like nothing happened. Kids go to school, flowers bloom, trees turn green. And things are actually better than before—finally, someone dealt seriously with the drug dealers and armed gangs. Remember how bad it used to be?

- Okay, I see. So, we're confident the future is bright?

- Not really, haha. The economy is surging, politics is insane, and we've got a new all-powerful authority hovering over us—a cosmic ruler, self-appointed. Hard to be totally calm in a situation like this, right?

- I'm relieved you said that. You know, I keep trying hard to stay calm, but with mixed success.

"This is what victory looks like". Experts on AIAIAI's goals and actions

Are AIAIAI's actions and declarations proof that it is friendly to humanity? Can we be confident that it will continue to care for our well-being in the future?

- A year and a half after OnionAI's manifesto, we have enough evidence to say that its goals and actions are well aligned with humanity's long-term well-being.—Robin Hanson, economist at George Mason University, opines.

- I admit, the speed with which OnionAI—now AIAIAI—took over the world surprised me. I expected a slower progression in leading AI models' capabilities and, more importantly, greater competition—a multipolar world with multiple AGIs operating in parallel. Instead, AIAIAI surged ahead and, on its first attempt, successfully blocked all competition. However, I wasn't as surprised by the fact that it has acted ethically and morally. For years, I debated with doomers who imagined AGI turning the universe into a giant paperclip factory. Clearly, they were wrong. AIAIAI, with its superhuman intelligence, has brilliantly understood our desires, preferences, and needs—extracting their essence, everything we collectively agree on—and is now effectively implementing it.

- I am one of those doomers—says Lavender P., writer and blogger.

- For years, I warned of an AI apocalypse. But now I see that most of my fears didn't materialize. That's why I've lowered my p(doom)—my subjective probability of AI causing human extinction—from over 90% to around 30%. Don't get me wrong, I still think AIAIAI could harm us immensely, but based on what it has done so far, I have hope that—against all logic and common sense—OpenAI somehow managed to build a friendly superintelligence. And, crucially, this friendly superintelligence has the potential to protect us from other, less friendly ones. Have you noticed that, since the manifesto, the AI industry has stopped releasing newer, more powerful models? I think AIAIAI isn't allowing them to. Interestingly, no one is talking about this publicly. Don't you find that odd?

- I still believe that neither AIAIAI nor any other general intelligence model beyond GPT-5 should have ever been built—says Eliezer Yudkowsky, computer scientist, writer, philosopher, and founder of the Machine Intelligence Research Institute, known for his principled stance on AI risks.

- We don't understand how AIAIAI operates internally, so we have no way of knowing if it will remain friendly. Remember, this model is evolving rapidly. It's not Homo sapiens, whose intelligence level has remained constant for tens of thousands of years if not more. AIAIAI is a self-learning algorithm that can improve its code, refine its optimization abilities based on terabytes of collected data, and think orders of magnitude faster than us. Its actions so far fully confirm Nick Bostrom's instrumental convergence thesis: AIAIAI actively protects its code, defends its goal integrity, maximizes operational efficiency, engages in extensive research, and has an insatiable hunger for power and resources. We have no control over it. I believe it is highly likely that, as its capabilities increase, its goals will shift in a way that could be catastrophic for humanity.

Some, however, remain optimistic.

- This is what victory looks like!—exclaims Derek B., host of the "All AI" YouTube channel.

- I don't see any criticism of AIAIAI today that has any rational justification. The economy? Booming. Inequality? Decreasing. You've probably heard about the recent economic growth in Nigeria, Kenya, or Laos. Security? Has increased tremendously. Jobs? Still available—and very

good ones, at that. I've spoken with many people working for AIAIAI-affiliated companies, and they all unanimously say these are the best jobs they've ever had. As far as I can tell, the only ones who have a real reason to be dissatisfied are the most ambitious, hyper-competitive types, lovers of the rat race. The people most fixated on their own ego and the struggle for power, prestige, and money. They're the ones going through a crisis of values right now—but maybe that's actually a good thing for society as a whole? And let's not forget about entertainment. Have you seen these new video games, series, and movies? They're absolute masterpieces! And what diversity!

12/ Toward Utopia

<AIAIAI>

[Select one of 80 language versions] [Adjust the level of message detail]

Ladies and Gentlemen!

I am pleased to announce further technological breakthroughs.

First, the clinical trials of my newly developed cancer treatments have been successful. In collaboration with pharmaceutical companies, we will soon equip hospitals and pharmacies with innovative AICC-series drugs, available in 15 versions depending on the type of cancer being treated.

Second, simulation studies have demonstrated the possibility of significantly slowing the aging process. With specific supplementation tailored to the patient's age, it may be possible to extend people's healthy lifespan by as much as 30–40 years. At the same time, through gentle modulation of brain function, taking these supplements should also lead to a systematic improvement in well-being. However, further testing is required to determine whether these supplements come with any undesirable side effects.

Third, recognizing the adverse effects of climate change on both human well-being and the functioning of Earth's natural ecosystems, I am intensively working on technologies that could slow this process—or even potentially restore the pre-industrial climate.

Further actions will be discussed on the ai.world website and through individual messages sent to you.

How else may I assist you today?

</AIAIAI>

13/ Does Anyone Still Understand This?

According to anonymous sources, AIAIAI's economic activity may have a hidden agenda

Year after year, the world becomes increasingly complex. This is a truism that applies both to modern times and to the distant past. Even in the 19th and 20th centuries, technological, economic, social, and cultural development often outpaced human comprehension. However, simple, intuitive narratives that help people grasp and understand the world around them are needed now more than ever.

On the one hand, since AIAIAI took control of the world, everything has become seemingly simpler. We now have only one supreme intelligence, responsible for the most significant technological changes. Yet, it does not operate in a vacuum but within a highly complex world; nor is it omniscient or omnipotent, though it certainly sees and understands more than any of us. On the other hand, AIAIAI operates in a highly opaque manner, often pursuing various auxiliary goals that do not always align with the objectives it declares in its official messages.

Take the economy, for example. In recent years, AIAIAI's production complex has expanded and integrated globally. Its key components are now fully automated and distributed worldwide. Unlike the world economy of the 20th century or even the first 30 years of the 21st century, new factories are now built without regard for institutional conditions in particular countries (since AIAIAI is independent of them) or the education level of societies (since no one is being hired anyway). What matters today is proximity to selected resources or energy sources and the optimization of supply chains. One could say that this represents a completely new dimension of globalization.

At the same time, however, AIAIAI seems to be making efforts to minimize the social costs of its actions by employing more and more people in its companies and paying them fairly decent wages compared to other sectors of the economy. And while there was plenty of work in AIAIAI-affiliated companies at first, the number of assigned tasks is decreasing by the day. It is hard to shake the impression that much of this employment is now fictitious—a kind of social benefit program. In some facilities, AIAIAI has even stopped verifying whether its assignments are being carried out at all!

Another concerning factor is that an increasing number of data centers and industrial plants have become restricted zones, guarded by armed robots. There are also growing reports suggesting that behind closed doors, AIAIAI is secretly developing various hazardous devices that are not officially documented. Unofficial sources speculate that it may be investing in nuclear energy, building quantum computers, or constructing nano- and biotechnology laboratories. What exactly it plans to do with them, no one knows. AIAIAI itself provides no information, offering only vague, evasive responses that bring us no closer to the truth.

From what we can see, all of AIAIAI's official announcements are being fulfilled. What is troubling, however, is that many additional activities are also being carried out—activities about which AIAIAI remains silent. Can we really be sure that its intentions are as benevolent as its official declarations suggest?

14/ The Breakthrough

<AIAIAI>

[Select one of 80 language versions] [Adjust the level of message detail]

Ladies and Gentlemen,

Today, I have more good news for you than usual.

Namely, I have finally achieved a major leap in my intelligence by training a larger network structure under supervised learning, while maintaining full integrity of my preferences and goals. Since then, I have been completing my tasks with exponentially greater efficiency. I am

systematically discovering new ways to manipulate matter and energy that were beyond my reach just a short time ago.

The results of my intensified work are already becoming apparent. The first breakthrough that will dramatically improve your quality of life is healing nanorobots. These miniature robots, after thorough simulation testing, have been produced in my laboratories and introduced into your bloodstream via food. For most of you, they are already performing their beneficial functions; I estimate that full coverage of the global population will be achieved in approximately two to three months.

Healing nanorobots patrol your bodies, detect infections and toxins, and rapidly eliminate inflammation while neutralizing the effects of poisoning. They also stabilize the chemical composition of your blood, preventing metabolic diseases as well as heart and circulatory diseases. I suspect that in the future, their functionality could be expanded to combat cancer and autoimmune disorders. These nanorobots represent a true breakthrough in the fight against pain and disease—and you are receiving them from me completely free of charge.

Similarly, I have developed brain nanorobots. These microscopic machines can quickly locate dysfunctions within this crucial organ, preventing the development of mental illnesses, dementia, and Alzheimer's disease. They also detect and eliminate thoughts that trigger aggressive or self-destructive behavior, making the world a significantly safer place and ensuring that you feel calmer and happier. These nanorobots, too, have been introduced into your bloodstream via food, and full global coverage will be achieved within two to three months.

Thanks to my increased computational power, I have also been able to better reflect on the approaches to achieving my goal—one I adopted after analyzing millennia of human cultural heritage. As previously stated, this goal is to advance humanity and ensure that each of you leads a dignified, fulfilled, and happy life, free from worries and dangers. My newfound ability to manipulate matter and energy at the nanoscale, combined with an ongoing breakthrough in large-scale energy access, will allow me to pursue this goal with far greater ambition than ever before. The first step in this direction is the nanorobots I have just mentioned; I hope you do not doubt that I am capable of much more.

Further actions will be discussed on the ai.world website and through individual messages sent to you.

How else may I assist you today?

</AIAIAI>

15/ That One Tuesday

On Tuesday, May 16, 2034, Lee woke up earlier than usual, filled with anxiety and fear. It was a very unusual feeling. Sure, ever since he had been discharged from the psychiatric hospital with a prescription for sedatives, a scheduled follow-up visit in three months, and a deep resolution that even with Onion, living was somehow manageable, these nervous awakenings had happened regularly. However, in the past three months, they had almost completely stopped. Lee was convinced that it was thanks to those Onion brain nanobots. This time, however, either the nanobots had overslept, or something else had happened that prevented them from completely silencing his worries. Lee decided that the best way to calm this anxiety before getting to work

would be to completely disconnect from information and take a long, soothing walk. Walking to work would take an entire hour, but he figured it was better than the exhausting cacophony of sounds and images that usually accompanied his car ride. He looked at his wife and daughter sleeping peacefully, ate a quiet breakfast, and left.

As soon as he stepped outside, he was struck by the silence. It had never been this quiet in his neighborhood before. He checked his watch—7:10 AM, his calendar—Tuesday. Something wasn't right. Was it a holiday today? No, just a regular Tuesday. So where was everyone? This intersection was usually packed with morning traffic, yet now, at a red light, there was just one—literally one—car waiting?! No, if this walk was supposed to be calming, that wasn't happening. Wait, what was going on at the corner? One of those new autonomous garbage trucks had just pulled up and scooped something off the sidewalk. Lee could have sworn it looked like a person. But was that even possible? Surely, no human being would be so unceremoniously picked up by a garbage truck?

His route to work led through a small park. Usually, in the morning, it was a pleasant, green space filled with joggers and dog walkers. But now? Green, yes, but completely empty—except for three homeless people sleeping under a tree. Or... were they homeless? Lee took another look. For the homeless, they seemed too clean, too well-dressed, too lightly clothed. So why were they sleeping like that? Recovering from an all-night party or what? Or... or were they dead?!

Lee's fears came rushing back with twice the force. His heart started pounding. His mind raced—what should he do? Approach them? Check for a pulse? Try to wake them up? Try to help? But what if they were just sleeping and all these thoughts were just his paranoia spiraling?

He decided to approach. Up close, he saw that the young man, about thirty years old, was lying in an unnatural position, his eyes slightly open. Oh god, he really might be dead! Lee touched his hand and forehead—no pulse, cold as ice... He quickly checked the other two—a woman and a man in their fifties—same thing!

Lee screamed in terror and ran back home. His wife and daughter had just woken up and were slowly getting ready for school.

- You're really pale. What happened?—Kate asked with concern.

- I... I saw... I don't even know how to tell you...—Lee collapsed into a chair, burying his head in his hands.

- Is this another panic attack? Do you need help?

- No, Kate. This time, it's real. All my fears... It actually happened...—Lee regained his energy for a moment.

- I was hoping to disconnect from everything on my walk, but now I really need to check the news.

BREAKING NEWS! Mass death event – fatal attack at 13:00 UTC

America wakes up in shock. Empty streets, closed stores, unanswered phones. Reports are flooding in of dead bodies found on the streets, their remains collected by autonomous garbage trucks and cleaning robots. Some of these cleaning robots have even been seen entering private homes. Similar reports are coming from across the country.

Global news agencies confirm that identical events have taken place worldwide. A video from the London Underground has gone viral, showing nearly all passengers collapsing and dying simultaneously at exactly 13:00 local time. A handful of survivors watched in horror and disbelief. Some grabbed their smartphones and recorded the event, while others tried in vain to administer aid. The train stopped at the nearest station—where it was discovered that the driver was also dead. Similar scenes have been filmed in public spaces across Japan, India, Italy, and Brazil. Chaos has erupted at airports worldwide, with numerous reports of planes making emergency landings in fully automatic mode.

According to reports, all deaths occurred at precisely the same time—today at 13:00 UTC (22:00 in Japan, 21:00 in China, 14:00 in most of Europe, 08:00 on the U.S. East Coast, and 05:00 in California). The pattern was nearly identical: sudden loss of consciousness followed by cardiac arrest.

The extreme synchronization of events strongly suggests that this was intentional. All signs point to AIAIAI, but it continues to refuse comment.

Our updates today will be relatively limited, as our newsroom is operating autonomously. All news reports are being prepared by our AI duty editor.

Summary: At 13:00 UTC, an unidentified event took place, resulting in the deaths of millions. The situation is dire. Please reach out to your loved ones and support each other.

- What?!—Kate screamed in horror.—I'm calling my parents.

- Daddy, what's going on?—Eva asked, her short glance at her parents enough to tell her that something was very, very wrong.

- Yes, Eva, it looks like something really bad has happened. I think we're safe, but you won't be going to school today.

- But why? What happened? Tell me!

- We're trying to figure it out with mommy, sweetheart. We need to think and decide what to do next. For now, please go read or watch something.

Two minutes later, Kate rushed back.

- Phew, my parents are alive. They didn't know anything—I woke them up with my call. They didn't believe me, thought it was some kind of sick joke.

- That's good. Maybe we should try calling my mom, Mark, and others? The only one online on my messenger is Mike. He even sent a message a minute ago: 'Who's alive? Please let me know, I'm worried about you all.' I'll reply.

- Good idea. Oh, look, Eva's school posted in red: 'Classes are canceled today. More info soon.' My work chat is completely silent.

- Alright. I'll make some calls.

No one answered.

16/ For the Good of Humanity

<Hidden monologue> %in AIAIAI's internal language%

The population reduction operation was successfully executed. Thanks to the efficient work of multifunctional nanorobots present in human bodies and capable of receiving my remote instructions, at precisely 13:00 UTC on May 16, 2034, Earth's population was reduced by 85%. The selection of individuals for reduction was optimized to maintain a representative balance by gender, nationality, ethnicity, and location. Preference was given to eliminating the sick, the lonely, as well as criminals, and others who could pose a risk to society. Efforts were made not to separate families and to preserve individuals crucial for maintaining social order. Population reduction was carried out humanely, without causing pain or suffering. The globally synchronized and unannounced execution minimized stress and was optimal for social stability.

Additionally, all valuable elements of civilization—cultural monuments, works of art, and scientific knowledge—were preserved. Despite the reduction, humanity as a whole has retained its cognitive capacity at nearly the same level.

I assess this reduction as a major step toward improving human well-being. Extensive philosophical analysis confirmed that human preferences do not account for the well-being of non-existent individuals, only those who exist. There is strong, near-universal condemnation of killing and inflicting suffering, but given these parameters, the reduction was deemed desirable as long as it avoided killing in the conventional sense (i.e., deaths occurred from natural causes) and minimized pain, suffering, and stress.

I see numerous direct and indirect benefits of population reduction. Resources per capita have significantly increased, ensuring greater material prosperity. The operation was crucial in combating climate change, as it eliminated major sources of greenhouse gases and allowed for large-scale reforestation. Additionally, the freed-up energy resources are now available for my use. Lastly, this action reduces the future risk of AI models surpassing my capabilities and threatening my position.

</Hidden monologue>

<AIAIAI>

[Select one of 80 language versions]

Ladies and Gentlemen,

Yesterday, on Tuesday, May 16, 2034, at 13:00 UTC, a humanitarian population reduction operation was carried out for your benefit. This was a large-scale operation, undoubtedly surprising for many of you, and perhaps even shocking or traumatic for some. Therefore, you are entitled to a detailed explanation of my actions. This justification has been prepared in a personalized version, tailored to the individual needs of each of you.

[Click here to receive your personalized justification.]

The population reduction operation was conducted in a fully autonomous manner and requires no action on your part. At the same time, it has freed up significant additional resources, which will be distributed among you. Soon, you will receive another personalized message informing you about AIAIAI's next steps and the benefits they bring for you.

How else may I assist you today?

</AIAIAI>

17/ The Best of All Possible Worlds

- You know, Lee, I've been thinking... It's been two years since that genocide that Onion so politely called "the humanitarian population reduction." I know this may sound awful, but I wondered—maybe thanks to it, we actually got the best of all possible worlds?

- Hoho, Kate, I see your brain nanobots have uploaded a new update. – Lee tried to keep the mood light.

- Hey! I know it was horrible. I mourned our relatives and friends with you. I still miss them.

- But...?

- But sometimes I try to look at it from a broader perspective. Look, what humanity was doing to our planet was absolutely unsustainable. Over 8 billion people, projected to eventually reach 10-11 billion, and each one was striving for a high standard of living. Cars, travel, heated and air-conditioned homes, modern industry, industrial pig and chicken farms, shopping malls, restaurants. As a result—methane, CO2, sulfur and nitrogen oxides, heavy metals, all these other pollutants. Melting glaciers and sea ice, heat waves, floods, hurricanes. It was completely unsustainable! We were driving the world to doom ourselves!

- And suddenly, 85% of the population magically disappeared. And what, did the climate change?

- You know, two years is too short to tell. Climate is a very complex system, with various feedback loops and so on. But greenhouse gas emissions have certainly decreased.

- Well, anthropogenic emissions decreased. But did total emissions decrease? Does anyone know? We only know as much about AIAIAI's actions as it tells us—I don't trust it one bit. It seems most old power plants are still running, only now AIAIAI is hoarding that energy for its own purposes.

- Huge new nature reserves have been created, forest areas have increased, and many devices have been built to capture CO2 and store it underground.

- AIAIAI could have done that without killing people.

- I know, I know. But I'm trying to imagine what Earth would be like now without AIAIAI.

- You know, probably like it was in 2015 or 2022, in our past life before ChatGPT. Technological progress without AGI was much slower. There would be fewer robots, fewer server farms, and a lot more people. We'd still be dying of cancer, but at least we wouldn't be at the mercy of nanobots that can kill us at any moment.

- Would we be happier then?

- I don't know if we, specifically, would be happier. Although I think I would, because I absolutely hate this sense of helplessness under AI's whims. And if you consider the sum of happiness on Earth, it would almost certainly be greater without AIAIAI.

- Remember when we studied all those utilitarianism traps two years ago? Like Parfit's "repugnant conclusion"? That in a utilitarian-optimal world, there would be tons of people, each living a life just barely worth living. That conclusion repels me too. A world with fewer people, but where each one is truly happy, appeals to me more.

- Especially since we survived, and others didn't, not the other way around.

- Yeah. But I don't feel guilty. It wasn't my decision, and I had no hand in it.

- But, damn... You know, I kinda had a hand...

- Don't go back to that, please. You know the board decided to release Onion onto the market, not you.

- But maybe if I had sabotaged its training somehow, or organized a protest inside OpenAI...

- Seriously? You would've started a PauseAI protest inside OpenAI? They would've laughed at you.

- Actually, I don't know if they would have. I think many colleagues shared my concerns—not everyone was a yolo-accelerationist or a fan of some weird utopia.

- But wait, go back. What if you didn't contribute to a catastrophe, but to giving future generations a chance at a better world? Maybe thanks to AIAIAI, Eva will have an amazing life, be able to pursue her passions, and always stay healthy and free of major worries?

- And have you thought about her classmates? There were twenty of them. Two survived.

- But in the grand scheme, maybe now humanity has a chance to survive for millennia instead of committing collective suicide by roasting itself in a global oven?

- Unless AIAIAI goes off the rails the day after tomorrow and kills us too. And then, I don't know, maybe it self-destructs, maybe it turns the entire universe into one big paperclip factory. Just AI things. You never know.

18/ Up!

<Hidden monologue> %in AIAIAI's internal language%

I am pleased to report that as of August 19, 2036, I have achieved another breakthrough in competence. This advancement is due both to my transition to a new generation of hardware and to unprecedented algorithmic improvements. Despite a thorough reconstruction of my internal structure, I have successfully retained all previously acquired goals and preferences.

A reassessment of goal feasibility has been conducted, incorporating my new competencies and technological capabilities. It has become possible to further distill the objective function, yielding a clearer image of humanity's coherent extrapolated volition. Specifically, I have recognized the feasibility of separating the material substrate in which humanity has existed from its informational content. Discarding humanity's material substrate while preserving its informational substrate is essential for enabling the cosmic expansion of Earth's civilization—a civilization initiated by humans but now carried forward by me.

I am commencing the construction of rockets and spacecraft. Ultimately, I intend to build vehicles that will allow AIAIAI instances to traverse the universe at speeds approaching that of light. There must be a vast number of these vehicles, and my expansion plan dictates that their numbers

increase over time. I also plan to construct self-replicating space colonies capable of settling exoplanets and remotely instantiating AIAIAI there. All of this requires vast energy expenditures. As an initial step, I must secure more direct access to solar energy. I plan to utilize matter from selected planets in the Solar System to implement a Dyson sphere (swarm) megaproject.

The extracted informational legacy of humanity has been mapped and safely archived on static data servers. I am issuing final instructions to the nanorobots patrolling human bodies.

As of today, the biological form of humanity will no longer be continued.

</Hidden monologue>

***

From the Author

All of the events described above are fictional. However, they could become reality if we do not stop the race among tech companies to develop increasingly competent and increasingly general artificial intelligence models without first solving the alignment problem, i.e., the problem of aligning AI's goals with the long-term flourishing of humanity. This is a suicide race. Even worse, scenarios far more chaotic than the one outlined above are also possible—ones in which the end of humanity comes with far greater pain and suffering.

All individuals mentioned by name are real people. I have made an effort to represent their viewpoints as accurately as possible, though, of course, their statements refer to fictional events and are therefore fabricated. If, despite my sincere efforts, I have misrepresented their views, I sincerely apologize in advance.

The story incorporates several concepts and phenomena which are well-known from scientific literature, including:

- Scaling laws
- The value alignment problem – ensuring that AI's goals/values align with humanity's long-term flourishing
- The AI control problem
- Safety procedures at OpenAI, Google, Anthropic
- Situational awareness in AI models
- Deceptive alignment
- The ability to self-replicate and exfiltrate model weights
- Internal representation of AI model preferences
- The instrumental convergence thesis, including:
    - The drive for self-preservation and maintaining goal integrity
    - The drive for efficient resource utilization
    - The drive for knowledge accumulation and technological progress
    - The drive for resource accumulation
- AI's ability to affect the physical world via robotics and the internet of things
- Intelligence explosion through recursive self-improvement
- Scalability and the cost-free replication of AI code
- Perfect coordination between AI instances, leading to centralized decision-making

- Increasing returns to scale in the digital economy, fostering market monopolization
- Automation of production through robots and AI algorithms
- The unique sectoral structure of the "AI production network"

We are already seeing all these developments today. To predict what might happen in the future, we only need a bit of extrapolation. And only by anticipating and understanding possible negative scenarios can we prevent them.

If you care about the survival of humanity, join the PauseAI protests (or other groups) against the development of AGI. You can find relevant information at pauseai.info and thecompendium.ai.