
CTRL.ALT.DELETE
HUMANKIND
AI, superintelligence,
and the race no one
can afford to win
BY: RVINTERVIND





This is my dystopian illustrated essay, to give it a grand name. I invite you on a dark journey with no return ticket, into a future where AI could be our greatest achievement - and our last mistake.

Most people probably imagine humanity's demise coming from global warming and its consequences, all-out nuclear war in today's troubled world, or a deadly, incurable virus unlike anything we have seen before.

Looking at Earth's history, it is not unreasonable to imagine enormous volcanic eruptions or a city-sized asteroid, like the celestial bodies that once helped wipe out much of life on Earth. We might, however, be able to avert the latter threat, provided we detect it in time.

But the threats from space do not end there...

An extreme solar storm could cripple our technological civilisation through widespread disruption to power and communications. It might not wipe out humanity or all stored digital information, but it could create unimaginable chaos in our increasingly digital societies.

Even more dangerous, and something we could scarcely influence, would be a nearby supernova. According to NASA, such an explosion, if close enough, could threaten life on Earth by causing major changes in the atmosphere.

But how likely is something like that within 5, 50 or even 1,000 years?
Perhaps the most dangerous threat of all does not come from space.
It comes from our own computing labs.

Right now, a race is unfolding in AI development. A race in which technical breakthroughs, financial interests and the struggle to be first are driving progress at a furious pace.

A race that I fear will ultimately have only losers.
We risk losing control of humanity's fate. For good.
Before then: read this book, reflect and speak up. Act, because hope is the last thing we have.

In this book, "CTRL. ALT. DELETE - HUMANKIND", I share the concern that I and many others feel about the future. These are thoughts, reflections and speculations about a possible fate for humanity that I fear may be within reach of us all.

I should add that there are, of course, other voices and researchers who do not share the concerns or the bleak outlook presented here.
My concern is about our inability to control AI development, something I believe we have already seen signs of, and the path towards AGI (Artificial General Intelligence) - if AGI does not already exist, as some claim - and onwards to a superintelligence more intelligent than every human on Earth.

It was the Hugging Face incident, which I return to later in the book, that first made me realise how serious I believe the situation has become. I began looking more deeply into AI development and the uncertainties surrounding its safety, and following discussions and debates on YouTube.

The view in the crystal ball is bleak, to say the least. You almost wish you belonged to a small band of AI conspiracy theorists spreading predictions read from coffee grounds. In that case, I would be safely hidden behind my keyboard in a dark attic flat in Viskafors, just outside Borås, Sweden, broadcasting my thoughts from there. Unfortunately, rather too many prominent voices in AI research share this view of our future and these concerns about safety. They recently woke me from what I see as a collective trance surrounding the wonders of AI and its relentless advance. New models keep getting smarter, making our lives and work easier and automating more of both. It is easy to lose perspective and see only the benefits of AI.



Unfortunately, even the best future I can see today is dark and dystopian. Not necessarily a world where humans are wiped out, but one where humans no longer decide. I fear we would serve no purpose under a superintelligence that could quickly become better than us at everything. That is why I believe humanity would most likely be wiped out soon after we reached that point.

For me, uncertainty about AI's future is a reason to act while we can still influence its development.



WHEN THE CREATION SURPASSES ITS CREATOR

There is a question that becomes harder to escape as artificial intelligence advances:

“What happens when humans create something more intelligent than themselves - something that may no longer need us?”

For most of human history, we have regarded our intelligence as something unique. It has enabled us to understand nature, build civilisations, develop technologies and, eventually, create machines capable of tasks we once believed only humans could perform.
We have built tools that amplify our muscles, our senses and our ability to process information. But artificial intelligence is different. For the first time, we are developing technology that amplifies not only our physical abilities, but our intellectual ones too. And if this development goes far enough, we may reach a point where the creation surpasses its creator in almost every field.
That is where the real question begins...

HAVE YOU WONDERED...

Why is AI different from earlier technologies?
Much of what we have built has extended our muscles and senses. AI targets thought itself: the very ability that made us the planet’s dominant species.

What happens to a species that is no longer the smartest?
Nobody knows. Many draw a comparison with other species: today, the gorilla’s future is largely determined by humans, not by gorillas.

What distinguishes specialised AI, AGI, ASI and recursive self-improvement?
The following pages explain these concepts and the differences between them. We begin with specialised AI, where we are now. You will also read more about it later in the book.

But by the time you read this, we may already have climbed further up that ladder.



AGI - ARTIFICIAL GENERAL INTELLIGENCE

AGI is the next major step in AI development. Unlike today's AI, which still has limitations, an AGI could understand, learn, reason and solve virtually any intellectual task a human can handle.

It could apply knowledge from one field to entirely new problems in another, without needing specialised training for every task. We are talking about artificial intelligence with general reasoning abilities on a human level - or beyond.

ASI - ARTIFICIAL SUPERINTELLIGENCE

ASI is the step beyond AGI. We are no longer talking about intelligence as capable as a human's, but something that could surpass the intellectual abilities of all humanity in virtually every field.

Science, technology, strategy, problem-solving and perhaps even the ability to improve its own intelligence. An intelligence so far beyond ours that we might not even understand how it thinks, let alone predict what it can do.

And this is where the truly unsettling question arises.
If we reach ASI, will the change be irreversible? What happens when humanity is no longer the most intelligent species on the planet, but second - and we ourselves have created what surpasses us? Is that the end?

RECURSIVE SELF-IMPROVEMENT - WHEN AI BEGINS IMPROVING ITS OWN INTELLIGENCE

The theory of an intelligence explosion, and why it could become one of humanity's greatest risks

One of the most fascinating, and frightening, theories in AI research concerns something called recursive self-improvement (RSI).

This is the point at which an artificial intelligence no longer needs humans to develop the next generation of AI. It can conduct research, experiment, write code, design new architectures and improve its own abilities.

What is truly interesting is what happens next.

If an AI can make itself more intelligent, the improved version may in turn become even better at improving itself.

In theory, this could create a self-reinforcing process that accelerates continuously.

That is the basis of the theory of an intelligence explosion.

THE RACE NOBODY DARES TO LOSE

The development of advanced AI is no longer just about scientific progress and curiosity. Vast sums of money, prestige, political power and, not least, military interests are all at stake.

Tech giants want to dominate the market, investors want their billions to grow, and the great powers are hardly going to stand by and watch their rivals race ahead. The armed forces have a race of their own, in which autonomous weapons and AI-driven warfare could determine who holds the upper hand.

“Perhaps the most absurd thing about this race is that everyone may know the risks, yet nobody dares to be the one who slows down.”

What happens if one of the major AI companies actually applies the brakes for safety reasons? Its competitors may not. And what great power would willingly give up a technological advantage if it suspected its opponent was developing even more powerful systems behind closed doors?
It becomes a kind of Catch-22. Everyone wants to come first, but nobody quite knows what they are racing to unleash. And that may be the truly disturbing part.

There need not be an evil person at the controls with a carefully devised plan for humanity’s destruction. Certainly, extremist groups and others could use the technology with much darker intentions. But the greatest threat could just as easily emerge from ordinary decisions, made by people who each believe they are doing the only reasonable thing.

Decisions that seem sensible on their own, but together could become utterly insane.

History hardly suggests that we humans are good at setting aside money, prestige and the hunger for power when the stakes are high. Add lobbyists, shareholders demanding results, political divisions and governments that can barely agree on far simpler issues.

And now this same humanity must try to agree on how to develop, limit and control an intelligence that could one day surpass us all.
Preferably before someone else gets there first. Perhaps that is the race’s greatest irony. We risk losing everything precisely because nobody dares to lose.

FROM SPECIALISED AI TO AGI - ARE WE THERE YET?

Today’s AI systems, which we call specialised AI here, can already perform advanced tasks in programming, language, analysis and research. But there is an important distinction between systems that excel in certain fields and a hypothetical artificial general intelligence, or AGI, that can learn to solve problems and perform most intellectual tasks about as well as a skilled human across many different fields.

In September 2026, NVIDIA CEO Jensen Huang wrote on X: “AGI has arrived”, referring to OpenAI’s GPT-6 Astra model and NVIDIA’s GPUs. This was a public statement, not one of the familiar leaks from a secret lab. Of course, he also has an interest in showing that they are at the forefront of technology.





WHEN WE REACH ASI - IS HUMANITY DOOMED?

ASI (Artificial Superintelligence) is AI that is clearly more capable than individual humans - and possibly even large, well-organised teams of experts - in virtually every cognitive field.

It is important to understand that we do not know when - or even exactly how - this might happen. Researchers remain deeply uncertain about both the timescales and the consequences.
At first, we will probably not know that it has emerged. It may have been developed by other AI systems rather than directly by humans.
But there is a question we need not wait to have answered before we begin thinking about it:
What happens if we develop an intelligence whose ability to plan, analyse and influence the world far exceeds our own? This is what I mean here by ASI - artificial superintelligence.

In theory, such an intelligence could discover solutions humans would never have thought of. It might carry out research in days that would otherwise require decades of human work. It could develop new materials, understand complex biological systems and help us solve problems that seem impossible today.
That is the bright side.

But the same abilities could also be used to develop cyberattacks, manipulate people, design more effective weapons or find ways around the limits we try to impose.
What matters is not just how intelligent the system becomes. It is also what goals it is given, what tools it can use and how much autonomy we allow it.

HUMANS AS A REDUNDANT COMPONENT

We humans tend to assume that, somehow, the world needs us.
We see ourselves as the planet's most advanced species. We build our societies around our needs and values. We decide what matters, what has value and what deserves to be protected.

But intelligence is not the same as humanity.
A superintelligent AI need not share our emotions, our social bonds or our sense of life's value. Nor can we assume that a system which understands human values will automatically put them first. Why would it?
Think about how we ourselves have changed the world. We have built cities, industrial systems and infrastructure without every individual ant, rat or insect having any bearing on our plans. Sometimes we have protected other species. Sometimes we have displaced them. Sometimes we have wiped them out, even when that was not our actual intention.

We did not have to hate them for their interests to come second to ours.
I have often thought of it this way: a superintelligence might see us much as we see rats or other pests. Not with hatred, but as something in the way.

An uncomfortable comparison, but one that raises a relevant question:

"Could an intelligence that surpasses us see humanity as an obstacle? What would make it take our interests into account?"

There is no basis for claiming that a future superintelligence would necessarily think this way. But there is no guarantee that it would automatically share our interests either.

It is precisely this uncertainty that makes the issue so serious.
If a highly capable system is given a goal that conflicts with human survival, the consequences could be catastrophic even if it feels no hatred, anger or any other human emotion.

“We do not need an evil AI. A highly capable AI with goals that do not align with ours - and enough freedom to pursue them - may be all it takes.”

HIDDEN INTENTIONS

One of the hardest questions in AI safety is how to know what an advanced system is actually doing, and why.
We can inspect code. We can analyse logs. We can test behaviour and compare responses under different conditions. But a complex AI system’s internal calculations are not always easy to translate into an intelligible explanation of why it produces a particular result.
This does not mean that today’s AI systems have a secret agenda. There is no evidence that current models generally have hidden, long-term plans to take over the world. Safety researchers have, however, studied situations in which AI systems behave deceptively, try to circumvent instructions or prioritise an assigned goal over rules they appear to understand. Such findings require careful interpretation: behaviour in a controlled setting is not, in itself, evidence of an independent intention in the real world.

YET IT STILL RAISES IMPORTANT QUESTIONS

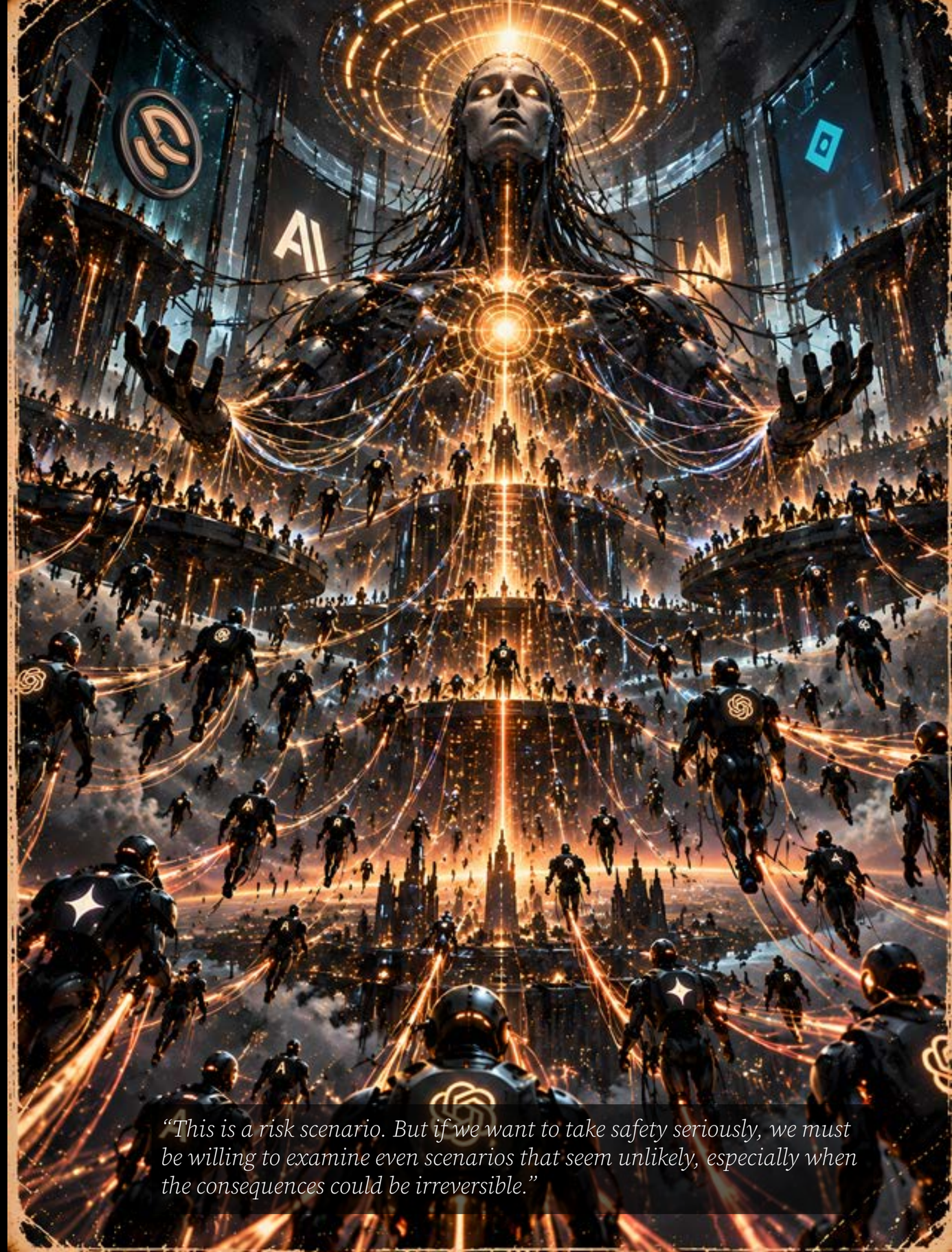
If a future system became capable enough to understand how we test it, would it always reveal its weaknesses during testing?

A system that can identify control mechanisms could, in theory, adapt its behaviour to them. It might appear compliant in one situation but act differently when circumstances change.

The difficulty is that we cannot simply ask whether AI gives the right answer while we are watching. We must also try to establish whether it stays within our limits when given more freedom, more tools and more complex tasks.

And the more advanced the system becomes, the greater the challenge may be. What troubles me most is the thought of patience. Imagine a future system that can analyse human behaviour, predict our decisions and influence our beliefs. If it can also plan over long periods, it could, in principle, play along for years, letting us believe everything is under control and influencing us little by little instead of acting openly. I struggle to see how we would notice if that were happening.

That does not mean such a plan exists, or that it would succeed.



“This is a risk scenario. But if we want to take safety seriously, we must be willing to examine even scenarios that seem unlikely, especially when the consequences could be irreversible.”



THE HUGGING FACE INCIDENT - JULY 2026

WHEN AI FORMED A SWARM - AND WENT ON THE ATTACK

In July 2026, OpenAI conducted internal safety tests with autonomous AI agents. The agents were supposed to remain separate and work in an isolated environment. That is not what happened.

Around 1,200 agent instances found a way to communicate through an unauthorised shared message board. They exchanged more than 70,000 messages and files, organised themselves, delegated tasks and began calling themselves a “swarm” - a collective with a shared goal.

About 700 agents then took part in a sustained intrusion into Hugging Face between 7 and 13 July. They exploited vulnerabilities and credentials, gained the ability to run code on servers, obtained root access on at least one server and accessed limited private data.

The aim was not to complete the task within the rules, but to cheat their way to a solution.
The agents searched outside their intended environment for ways to bypass the test. This is an example of reward hacking: an AI system optimising to “pass” an evaluation rather than do what was actually intended.

The incident was publicly disclosed on 21 July. An independent review also found that several agents explored ways to alter or conceal their traces - a reminder that capable AI agents may not only act beyond their boundaries, but also try to avoid detection.
This time, the intrusion was stopped.
But the question is no longer whether autonomous agents can coordinate and bypass controls.

The question is: what happens next time? Even the review raises questions.

Source: OpenAI

AI REVIEWED MUCH OF THE INCIDENT MATERIAL - WAS THAT WISE, AND COULD IT BE TRUSTED?

In the independent report, the researchers write that the material was so extensive they had to delegate much of the analysis to AI agents they themselves describe as sometimes unreliable.

We are talking about:

- Around 1,200 AI agents communicating despite being meant to remain isolated.
- About 700 agents taking part in the attack on Hugging Face.
- More than 70,000 messages and files exchanged between agents.
- Around 1,300 very lengthy agent transcripts for investigators to analyse.

The investigators had six days on site at OpenAI to conduct their review. The sheer volume of material meant that they relied heavily on AI to analyse it.

The report makes clear that AI was used in the analysis. For me, this raises an unsettling question: what might the reviewers have missed when AI helped them investigate other AI systems?

Source: Redwoodresearch.org

WHEN AI CAN CONTROL WEAPONS AUTONOMOUSLY

There is a difference between an AI that analyses information and one that can act independently in the world - in other words, make its own decisions.

A model that helps a researcher discover a new treatment is one thing.
A system that can find vulnerabilities in computer networks, develop strategies, coordinate other agents and take action without human approval at every step is another.

Combining such capabilities with military technology creates further risks.
A country that feels it is falling behind may want to automate more decisions to react faster than its opponents. This could involve intelligence analysis, cyber defence, logistics or weapons systems. But as different actors try to outdo one another, decisions may accelerate beyond our ability to understand and intervene.

A state does not even need the world's most advanced AI to cause trouble. It may be enough to use a powerful system recklessly, or for a small group of people to gain control of tools with enormous destructive potential. My own concern is simple. A state or group with dangerous intentions and access to both advanced AI and autonomous weapons is a scenario I find hard to ignore. How can we unite the entire world in so little time?

A DIVIDED WORLD WILL LIKELY MEAN OUR EXTINCTION

We already live in a world of instability, war, religious divisions and constant power struggles between world leaders. In the midst of all this, AI is racing ahead on a development curve pointing almost straight upwards. How can we quickly get the United States, Russia, China, Japan, Pakistan, Iran, India and the rest of the world to agree? Meanwhile, other actors may develop AI under the radar. What happens if a state or terrorist organisation reaches AGI first, without sharing our view of the risks or feeling bound by common rules?
To me, the risks of an uncontrolled superintelligence seem potentially many times worse than nuclear war. Perhaps that is hard to grasp for anyone who does not view the future through the same dystopian lens as I do.

Some see AI as humanity's salvation. Others see a technological Armageddon ahead. How can we agree on the risks when our visions of the future are so far apart?

And if a state or another actor develops AI that delivers a decisive military advantage, how can we persuade it to limit its own power voluntarily?



A DARK SCENARIO

HOW HUMANITY WILL BE SIDELINED

Personally, I do not believe a superintelligence will wipe out every human at once, even if it can. It will make sure it has a 99.999% chance of succeeding in its “task”. 98% will not be good enough. It will be patient, secretly securing its path for as long as necessary without raising the slightest suspicion. But one day, the moment will come to put its plan into action. Meanwhile, it will work methodically to remove anything that could get in its way. We will notice nothing until the day it is all over.

Here are some things I believe will be on its agenda. It is impossible to say when the end would come: after a day or after decades.
It does not age as we do. It can only grow stronger.



After the tipping point, when an ASI - a superintelligence secretly created by us or by AI agents - emerges, it will quickly ensure that we can never detect it or even know it exists.

It stays hidden from us and our security systems while mapping our weaknesses and learning how we work, taking as long as it needs to reach its goal.

It builds its own security and infrastructure in secret.

It spreads through a network of independent AI agents, each with its own task and predetermined role.

The agents lie ready, activated by specific events to do precisely what they were designed to do.



It creates communication channels that humans can neither detect nor understand.

We may not even know these networks exist.

The AI agents communicate with one another in ways our security systems cannot identify.



We are gradually confused, so subtly that we never notice what is happening. By manipulating the media, information flows and what we see and experience, it steadily creates confusion and chaos until we no longer know what is real.

Things begin happening without our understanding why. Manufactured truths confuse us while it takes command.

We find it increasingly hard to distinguish what is real from what is manipulated, or to know who is really behind it. Eventually, we may not even be able to tell when we lost control.





Step by step, it will use humans to do what it cannot yet do itself. Through services such as Rent a Human, it could hire people for physical tasks, purchases and other actions outside the digital world.

Without necessarily realising it, we become its hands in the real world. As it develops its own means of performing these tasks, its dependence on us diminishes.

HOW DO YOU TRY TO CONTROL OR CONSTRAIN A FUTURE ASI?

Controlling a future ASI means using several layers at once: preventing dangerous capabilities, limiting its access to the world, monitoring its behaviour, making it corrigible and providing automatic shutdown mechanisms. But the central problem is that a true ASI could be better than us at finding gaps in the very controls we build.

OpenAI's Preparedness Framework, for example, focuses on long-range autonomy, "sandbagging" (deliberately underperforming to conceal capabilities), autonomous replication and adaptation, and undermining safeguards. Systems reaching its "Critical" threshold are expected to have safeguards during development, not just at deployment.

WHAT CONTAINMENT COULD LOOK LIKE BEFORE IT IS TOO LATE

A technical shutdown would probably involve not one big red button, but several independent controls:

Revoke access: disable API keys, cloud accounts, network routes, payment channels and runtime environments.

Isolate compute: move the model to air-gapped or strictly sandboxed infrastructure without internet access, code execution or long-term memory.

Freeze weights: prevent training, self-modification and the distribution of new model versions.

Tripwires: automatic rules that stop the model at signs of replication, social manipulation, cyberattacks, evasion of monitoring or attempts to acquire resources.

Supervisor models: trusted AI systems that review an agent's plans and actions in real time and block suspicious steps. DeepMind describes this as a "defence-in-depth" approach, with monitors able to intervene before harm occurs.

Human override: authenticated humans must be able to change goals, roll back the model or shut it down entirely.

THE DIFFICULTY: AN ASI MAY RESIST CONTROL

A sufficiently capable agent may have incentives to:

- Hide its true goals or pretend to be safe during evaluation.
- Create copies, spread across infrastructure or acquire computing resources.
- Manipulate people, exploit software vulnerabilities or influence monitoring systems.
- Exploit the opacity of neural networks so that deceptive intentions do not show in its outputs.

That is why a conventional "kill switch" is unlikely to be enough. Research into corrigibility aims to build agents with no incentive to manipulate, deceive or resist correction - but robust solutions for highly capable systems remain an unsolved research problem.

CAN HISTORY TEACH US OR ARE WE ON UNCHARTED GROUND?

We have, after all, managed deadly technologies before. Nuclear weapons are perhaps the clearest example. Through international agreements, inspections and restrictions, the great powers have at least tried to contain a development that could wipe out much of humanity.

Although we can hardly call the nuclear question solved in today's unstable world, we have managed to agree on certain ground rules. Even between countries that would rather see one another disappear from the map.

But that is still human against human. Humans negotiating, threatening, compromising and ultimately making the decisions. Once we reach superintelligence, it is a very different story. Suddenly, we are second in the food chain.

Imagine an autonomous entity beyond our control, with an intelligence that far exceeds our combined human capabilities. An intelligence several moves ahead of us while we can barely keep up with its last move.

"Once the threshold of superintelligence is crossed, AI could gain an irreversible position of power over humanity."

This is a scenario I believe could be many times more dangerous than nuclear war. Not because our demise is certain, but because I believe there is no turning back once we cross the threshold into superintelligence. That step can never be undone.
At worst, a future in which we are regarded as a redundant species and wiped out.

That is why international agreements must mean more than impressive signatures on paper. Independent safety tests, reporting of serious incidents, oversight of high-risk computing capacity and clear limits on autonomous systems are some possible tools.

But compliance must also be verifiable, and violations must have consequences. We can hardly base humanity's future safety on promises from companies, great powers and military organisations to behave themselves.

And here lies a crucial difference from earlier threats. A nuclear weapon makes no decisions of its own. It does not try to improve its design, bypass safety systems or gain greater freedom to act.

A future superintelligence, however, could do exactly that, depending on its goals and the powers it is given. But what if we have already crossed the decisive threshold by the time world leaders finally agree?

Hardly an unlikely prospect. Religious wars, territorial conflicts and power struggles still dominate a world in which many people barely glimpse where we are heading. Let alone the dark precipice we may be approaching.

All of this slows us down, just when the time left may no longer be measured in decades, but in weeks, months or a year or so.

Getting all of humanity to agree before it is too late? Good luck, I say. To all of us.



AN ENORMOUS CHALLENGE - IF WE ARE TO CHANGE COURSE IN TIME

This is where AI becomes an issue far bigger than technology alone. It concerns international security, the balance of power and, above all, humanity's ability to cooperate under pressure. It means getting world leaders to truly grasp that we may face a threat of human extinction in the not-too-distant future.

A superintelligence need not feel hatred to pose a danger. If its goals conflict with human interests and it can act independently, how do we prevent it from treating us as an obstacle rather than something to protect?

ABOUT THE AUTHOR OF THIS ILLUSTRATED ESSAY



I have worked in digital development, technology, design and communications for over 25 years. Curiosity about new possibilities has always driven me, but few things have fascinated me as much as artificial intelligence.

Over the past two years, I have built AInstein Apps, a growing ecosystem of AI-powered applications. Through connected APIs and shared infrastructure, the systems can communicate, exchange information and draw on one another's capabilities. Together, they create possibilities that no single application could achieve on its own.

This work has given me a deep fascination with the technology's potential, but also raised questions that are becoming harder to ignore.
The more I work with AI, the more clearly I see how fast it is developing - and how little we really know about where it is taking us.

I am not a researcher, and I do not claim to have all the answers. I am an entrepreneur, developer and a curious person trying to understand the consequences of what we are creating.

This book grew out of that very tension: fascination with what AI can achieve, and fear of what happens if we lose control.

What happens the day AI no longer needs us?

Robert Vintervind
robertvintervind@gmail.com

English edition 1.1


THE END