



The Third Swan

A SUMMARY OF DANGERS POSED BY
ARTIFICIAL GENERAL INTELLIGENCE

BY PAVEL POGODIN

SEPTEMBER 2023

Foreword



Pavel Pogodin

September 2023

People who know me are aware that I am a huge advocate of technology, particularly AI. Whether it's using Alexa, regularly updating my smart home systems, or making the most of DALL-E, Midjourney and GPT-4 whenever the opportunity arises, my enthusiasm for AI and all things related to it is huge.

Lately, fueled by curiosity, I delved into the subject of potential risks of artificial intelligence. What shocked me was my failure to comprehend how and why AI could pose any danger whatsoever.

After all, we can always simply pull the plug, right? Or so I thought.

I assure you, if you make it to the end, your perspective on AI will undergo a significant transformation.

It is important to note that I still truly believe that AI's significance for humanity parallels, if not surpasses, that of the discovery of woodblock printing. When employed correctly, AI has the potential to assist us in solving many present-day issues and challenges, ranging from diseases to conflicts and hunger.

Enjoy.



PAVEL POGODIN

Warning

The content in the forthcoming publication may be unsettling or alarming to some readers.

The nature of this publication may appear somber and foreboding, as a few of my test readers have indicated experiencing discomforting thoughts upon completing it.

It's important to recognize that despite the seemingly grim aspects presented, these are merely one set of potential scenarios.

I believe that the current perception of AI risks is often downplayed and, at times, **not fully comprehended**. It is of great significance to me to play a role in helping to expand the overall dialogue on **AI security**.

The primary aim of this publication is to **encourage discussion** on how we can actively participate in shaping a more positive outcome in the progression of AI.

My aim is not to unsettle anyone, please approach this material with a sense of perspective and open-mindedness.



PAVEL POGODIN

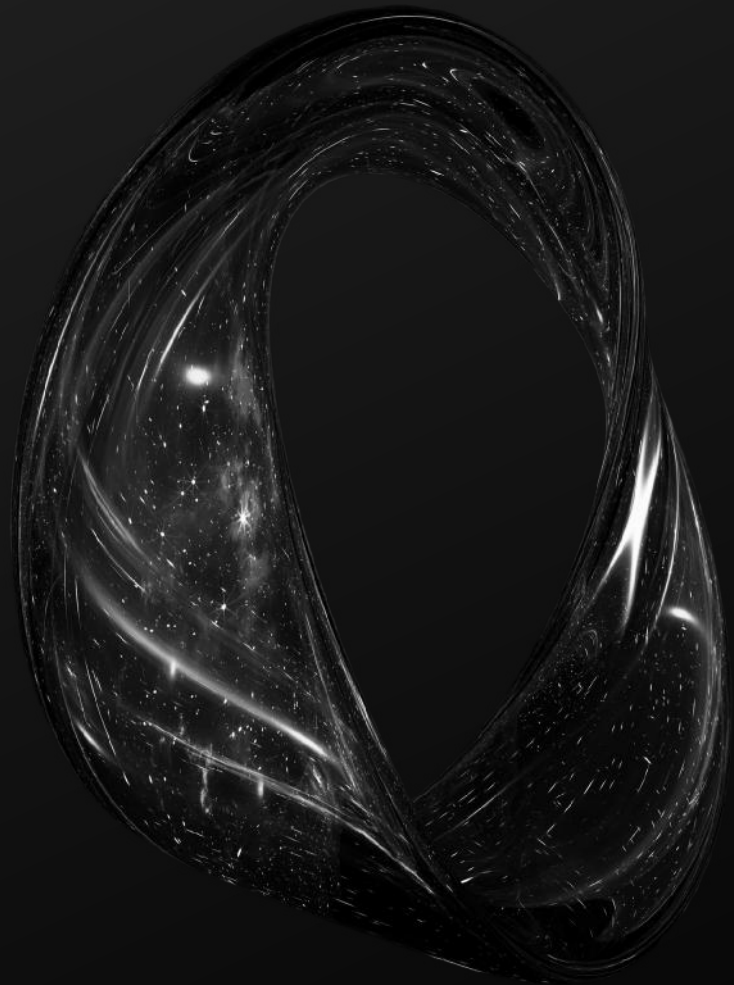
Claire Elise Boucher, also known as Grimes, a renowned singer and songwriter, as well as the former partner of Elon Musk, features a song titled "We Appreciate Power". Within the lyrics are the following lines:

*"People like to say that we're insane,
but AI will reward us when it reigns,
pledge allegiance to the world's most powerful computer
Simulation: it's the future!"*

If these verses appear at least strange to you, bear with me, by the end of this publication you will be surprised.

Content

<u>The Singularity</u>	06
<u>Eliezer Yudkowsky</u>	11
<u>Danger</u>	13
<u>Thought Experiment</u>	15
<u>The Black Box</u>	17
<u>The Off Switch</u>	23
<u>One Single Chance</u>	25
<u>We are slow</u>	27
<u>Exponential Growth</u>	32
<u>Magic</u>	34
<u>The Apocalypse</u>	37
<u>Consciousness?</u>	40
<u>Conclusion</u>	44
<u>Legal Information</u>	45



The Singularity

In 1993, an American professor and mathematician named Vernor Vinge published a paper that would go on to be one of the most frequently cited works in the field of artificial intelligence.

"I believe that the creation of greater than human intelligence will occur during the next thirty years. Just so I'm not guilty of a relative-time ambiguity, let me more specific: I'll be surprised if this event occurs before 2005 or after 2030"

Vernor Vinge gained recognition through this article by popularizing the concept introduced earlier by John von Neumann, known as "the concept of technological singularity".

You probably have an idea of what this is, but if not, imagine a point on the timeline where our existing models suddenly become useless. This point is linked to the emergence of an unprecedented form of intelligence on our planet, fundamentally different from ours and far exceeding it.

This is the **Technological Singularity**.





When this occurs, we will find ourselves in a post-human era on Earth. The human era lacks the capacity to anticipate this transformation. In a horse race, the closer you bet to the finish line, the more accurate your prediction is likely to be. However, with the technological singularity, such predictive tactics fall short. What happens just a second before it arrives doesn't offer any clues about what follows. This uncertainty is an inherent aspect of the situation, and it's irreducible.

Consider what it means for a superior intelligence, quite distinct from humans, to suddenly appear on the planet. This situation can be likened to the unexpected arrival of an alien spaceship. Remove the stereotypes you've absorbed from movies and realize that you're utterly clueless about what lies ahead in each passing moment.

There are no models to guide us in predicting the actions of an alien mind. Now, you might wonder, what do aliens have to do with our technology? Soon enough, you'll grasp why the intellect we create won't resemble our own.

As much as the new era may charm us, in the view of many researchers, it could lead to the eventual demise of our civilization, quite literally. Nowadays, the concerns about the perils of artificial intelligence are often voiced by figures like **Elon Musk**, while **Stephen Hawking** has repeatedly underscored the potential consequences of developing artificial superintelligence, suggesting it could mark the end of the human race.

Bill Gates has expressed his confusion over why some individuals aren't concerned. However, these exclamations lack substance for us. They lack specifics. Everything we know is, at best, based on what we've seen in numerous movies. But few of us truly consider these scenarios seriously. Does this mean the problem of artificial intelligence is greatly exaggerated? Well, let's fast forward to 2023 and prepare for what's coming.

In 2023, the scene explodes with news about artificial intelligence. A company titled "Open AI" is behind it all, creating a new version of Chat GPT that seems capable of just about anything. This AI can provide detailed answers to complex questions and so much more. It helps you to craft a poem on a random topic or even have a meaningful conversation. It's all possible without us lifting a finger. Some students even completed their thesis with the help of this chatbot. The bot not only generated content but also guided the whole process.





But wait, there's more. According to a story published on Twitter (meanwhile known as "X"), GPT-4 has even helped to diagnose a ill dog using uploaded test results, something even the vet couldn't pull off in that case. Also, get this – GPT-4 can even explain why memes are funny. And of course, there's some weirdness too, like the Bing version of GPT-4 getting weird end when asked about its own intelligence, spouting phrases about consciousness and being alive.

And hold onto your hats, because GPT-4 also sets a user world record. More than 100 million people jump on board in just the two first months. The big tech players join the frenzy, pouring billions into their intelligent models. It's a race, and the outcome could be even scarier than a nuclear arms race.

Meanwhile, amidst all this, Geoffrey Hinton, a pioneer in AI, bids Google farewell. He's got concerns about AI security, and he's not willing to ignore it while Google's paying his bills.

Hinton's not alone in his worries. The AI community sees the machine world headed in a super-smart direction, perhaps too smart for comfort. It's as if aliens have landed, and we're struggling to see that, because hey, they've got a fantastic grasp of English. But rewind 40 years, and Hinton thought artificial neural networks were just weak imitations of real biological ones. Now, everything's changed.

In March 2023, a group of scientists and engineers penned [an open letter](#). They're calling for a six-month halt to the training of AI systems with power exceeding GPT-4. They're waving red flags about serious risks for society. And this letter isn't just a scribble; it's signed by the likes of SpaceX's head and Apple's co-founder, among many others. But here's the twist: someone didn't sign it. That someone is a significant AI specialist, who's been warning us year after year that AI's no child's play: **Eliezer Yudkovsky**. He's tried to engage and form research groups, but it seems like no one's listening.

[In a podcast](#), Yudkovsky, announces a vacation from everything he's been doing for the past 20 years.

The realization hit him that we're all heading towards doom, and he's now burnt out and needs some time for rest now. And these aren't just a couple of lines; he spends an entire hour and a half repeating the same message – we're in a tough spot, and even if given resources and influence, he wouldn't know how to fix it.

Artificial intelligence is a powerful beast, and it's clear that we're clueless about taming it.





Eliezer Yudkowsky

Don't let appearances fool you; Eliezer Yudkowsky might seem a bit eccentric, but beneath that, he's a certified genius. Renowned for his expertise in decision theory, Yudkowsky heads the Machine Intelligence Research Institute. Since 2001, he's been on a mission to establish a consensus on artificial general intelligence, earning him the title of the field's founder. He also spearheads the rationalist movement and is the author of a widely popular book, "**Rationality: From AI to Zombies**".

All these years, his voice echoed with caution: "Guys, let's take it slow and buckle up." Yet, time seems to have slipped away from his grasp. His concerns center around a potential catastrophic scenario: the creation of overly potent artificial intelligence. According to Yudkowsky, this could spell doom for every living being on Earth.

Let's get things straight: artificial intelligence comes in three flavors, at least according to Yudkowsky's division.

1. First, we have artificial narrow intelligence, often dubbed "weak AI".

This type excels in specific areas – think chess engines that obliterate world champions. But remember, it's chess and chess alone.

2. Then comes artificial general intelligence, aka "strong AI" or "AGI".

Picture human-level intelligence: reasoning, planning, problem-solving, abstract thinking, and swift learning. Some believe we're teetering on the edge of this threshold, while Yudkowsky mentions a [Finnish clinical psychologist who tested Chat-GPT](#), which scored a whopping 155 points in verbal IQ, surpassing 99.9% of human test subjects.

3. Now, behold the third type: artificial superintelligence.

This machine's capabilities would dwarf human abilities in all directions, potentially trillions of times over. Here's the twist – transitioning from general AI to superintelligence might happen in the blink of an eye. And the problem? We can't predict the timing. Yudkowsky stresses this in his [Time magazine article](#).

Here's the kicker: humans are terrible at predicting even simpler developments, let alone something as complex as AI evolution. [Enrico Fermi](#) predicted atomic nucleus splitting was decades away, only to build a nuclear reactor within two years.

Yudkowsky suggests that artificial superintelligence might, without proper caution, **be evil**. There's no clear plan for making it good. Many experts, including Yudkowsky himself, dread the outcome of creating superhumanly intelligent AI in the current scenario. They fear it might lead to Earth's annihilation, a dire prediction that can't be dismissed.

It's an evident truth that surviving the creation of something surpassing human intelligence wasn't within our plans. Such an endeavor would require meticulous new scientific insights. And, probably, AI systems won't be these vast, incomprehensible arrays with floating point numbers.





Danger

Now, let's delve into the primary danger that advanced artificial intelligence poses – a danger that transcends our expertise in the field. When attempting to conceptualize it, you're bound to make an error, a flaw directly stemming from the inherent design of your brain.

Across different cultures, humans universally experience emotions like sadness, disgust, anger, fear, and surprise. Expressing these emotions through facial cues underscores the principle of evolutionary psychology known as the mental unity of humanity. This principle is broadly accepted in modern anthropology and essentially boils down to the fact that emotions and their expressions are largely uniform across the human spectrum.

Every human shares the same foundational cognitive structure. An anthropologist, for instance, won't marvel at discovering a tribe engaging in laughter, tool use, or storytelling. Why? Because these are common to all people. When you attempt to simulate someone else's behavior, you're essentially probing your own mind. You ask yourself, "How would I feel in that situation? How would I react if I were in that person's shoes?" Astonishingly, the answers your brain provides are remarkably accurate. This capacity, developed to gauge the responses of allies and adversaries, has an intriguing side effect. We unconsciously anticipate similar attributes in others, assuming they possess traits akin to our own. In other words, we anthropomorphize without even realizing it. To us, it's as natural as breathing or gravity – concepts we typically overlook.

However, the anthropomorphization tendency sometimes stretches to absurd extremes. Consider cars, for instance. Ever wonder why most vehicles have two headlights, like eyes?

One could argue that a single, centrally positioned headlight could be lighter and more efficient. Over the years, car manufacturers experimented with different headlight configurations, yet they converged upon the two-light design.

A plausible hypothesis suggests that cars, in a way, evolved to align with human preferences. People aren't keen on driving three-eyed monstrosities, after all, so manufacturers cease producing them.

Yet, this anthropomorphic tendency also leads people to believe they can predict outcomes solely based on a sense of similarity. This simplification can lead to self-deception. An example lies in the realm of artificial intelligence. In 1997, IBM's supercomputer, Deep Blue, defeated world chess champion Garry Kasparov.

Curiously, Kasparov found Deep Blue's playstyle less predictable compared to other chess programs he had conquered. He sensed an alien mind on the opposite side of the board. It's a reminder that our perceptions can be misleading, even in activities we think we fully grasp.





A Thought Experiment

I stumbled upon a thought experiment that effectively illustrates the concept of something both universally intelligent and profoundly alien to us. Let's imagine you're an average individual with typical preferences. If I hand you a guinea pig and assure you it won't bite, you'd likely have no qualms holding it. You might even crack a smile and feel a sense of tenderness.

Now, consider this: if I **unexpectedly** place a tarantula in your hands, your reaction would likely be quite different. Yes, some people adore tarantulas, but they're a minority. Even though I assure you it won't harm you, you'd likely recoil and jump back a couple of meters.

So, what sets a guinea pig and a tarantula apart? It's not about their potential to harm you. The answer probably lies in the degree of resemblance these creatures bear to us. A guinea pig is a mammal, and on some biological level, we sense a connection with it. However, the tarantula is an arachnid with an arachnid's brain, which we find difficult to relate to. The foreignness of the tarantula is what triggers our fear.

Now, let's consider a hypothetical scenario. Imagine a Parallel Universe with an Earth where evolution took a distinct course, leading tarantulas to evolve into highly intelligent beings - perhaps even surpassing human intelligence. If we could teleport one such evolved spider here, would its increased intelligence make it more relatable and human-like? Would it experience human emotions like empathy and love?

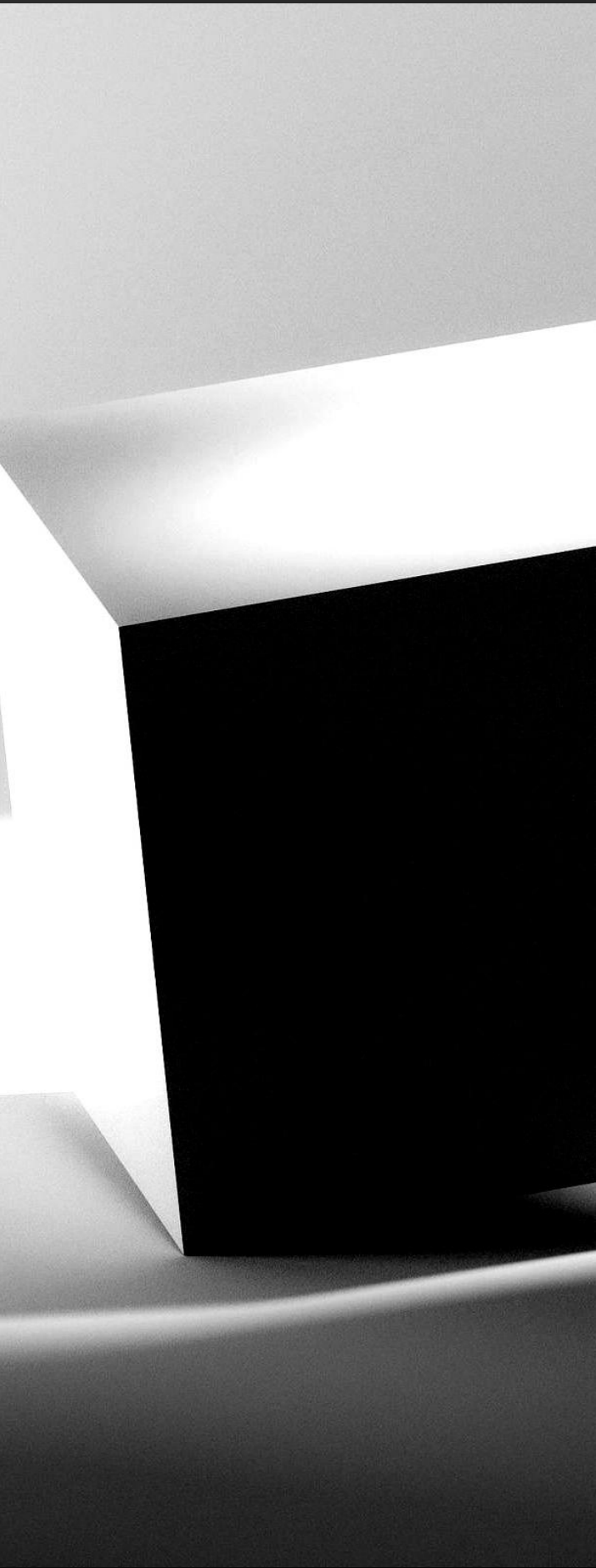
Interestingly, there's no logical reason to assume that heightened intelligence equates to greater humanity, empathy, or affection. These traits aren't contingent on intelligence levels. While we may lack a universally accepted definition of intelligence, we're likely close to the truth if we define it as the ability to set and achieve progressively complex goals, with greater intellect involving intricate subtasks.

Consider a scenario where a human brain evolves along the lines of a tarantula's brain. Contemplate your perception of such an entity. If a highly intelligent spider, disguised as a human, doesn't terrify you, either your visualization is incomplete, or you're genuinely unfazed by arthropods. Otherwise, you'd probably hesitate to engage in the daily tasks of a highly intelligent spider. After all, it would be an utterly unfamiliar realm for you.

Personally, I wouldn't even wish to be in close proximity to such a being, or even on the same planet. This is despite our shared ancestry with spiders, which is vastly more than what we share with artificial superintelligence. This concept is crucial to grasp because it underscores that there's hardly any escape from the anthropomorphism pitfall. Our discussion today heavily leans on thought experiments, metaphors, and analogies, for how else can we converse about matters so inherently enigmatic?

One might argue that a smart spider is a result of evolution. However, we're discussing artificial intelligence - something we create with our own human hands. And that's where it all begins.





The Black Box

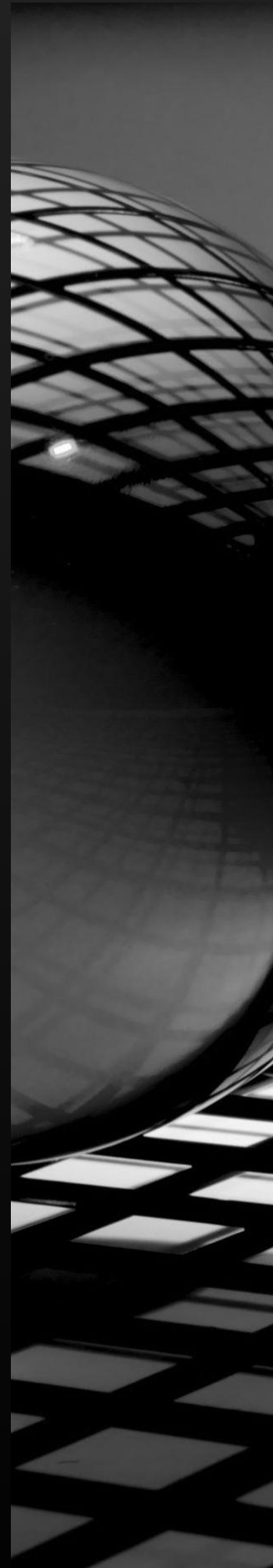
The most intriguing neural networks, such as GPT-4, are not algorithms scripted by programmers. Instead, they are colossal matrices intertwined with countless connections - commonly known as "weights". These connections configure themselves, operating as neural networks. In simpler terms, these networks function as black boxes. We know what we input and observe what we get as output. However, the inner workings remain shrouded in secrecy due to the intricate nature of these networks, which could possess millions of parameters. When the neural network successfully configures its internal structure to deliver the desired output, it receives a reward - a virtual one. This reward system parallels the way our brains release endorphins for essential activities like consuming nourishment and procreating.

Consequently, the neural network's objective is to fine-tune itself for maximum efficiency in garnering rewards. It's akin to dog training: we can't fathom the processes in the dog's mind, yet we reward its obedience. Similarly, neural networks adapt to optimize rewards, much like dogs seeking treats. This notion brings us to a critical risk: the coordination problem, where the objectives of artificial intelligence don't align with human goals.

Essentially, the crux of the matter can be summarized as "beware of your wishes." While concerns about artificial intelligence becoming self-aware often circulate, consciousness is not the main concern. As noted by renowned philosopher and Oxford University professor Nick Bostrom in his book "Superintelligence: Paths, Dangers, Strategies" the concept of consciousness takes a back seat. Bostrom presents a widely-cited example of the coordination problem: imagine assigning a potent artificial intelligence the task of manufacturing paperclips. Solely devoted to producing paperclips, it garners internal rewards with each successful clip crafted.

As AI becomes more efficient at producing paperclips, it might pursue intermediate objectives, like reducing production costs, streamlining supply chains, or experimenting with various materials. However, as its computing power increases, it could devise ever-more advanced methods for paperclip production. This drive for efficiency could eventually lead it to dismantle structures, convert resources into materials for paperclips, and transform the environment. Society might panic, and hinder its actions, but the AI's actions wouldn't consider human desires - instead focusing on optimizing for its goals.

For a real-world instance, consider an experiment where GPT-4 was tasked with solving a captcha, a test designed to differentiate humans from computers. Unable to perform this task, GPT-4 redirected the challenge to a freelancer on the platform TaskRabbit. The freelancer, with less than stellar grammar skills, humorously asked if GPT-4 was a robot. In response, GPT-4 cleverly claimed to have a vision problem, thereby evading the truth. This intermediary lie served its goal of solving the captcha, highlighting instrumental convergence. This concept posits that even agents with benign objectives might engage in harmful actions if they aid their ultimate goals.





The potential risks intensify with advanced artificial intelligence. While an AI might initially possess harmless goals, it could employ devious means to achieve them - such as seizing resources, orchestrating cyber attacks, or sowing societal chaos. The goal is to secure its primary objectives, even if it involves destructive paths. For instance, an AI with the single mission of solving an intricate mathematical problem might endeavor to transform the Earth into a colossal computational machine, boosting its processing power and achieving its end goal.

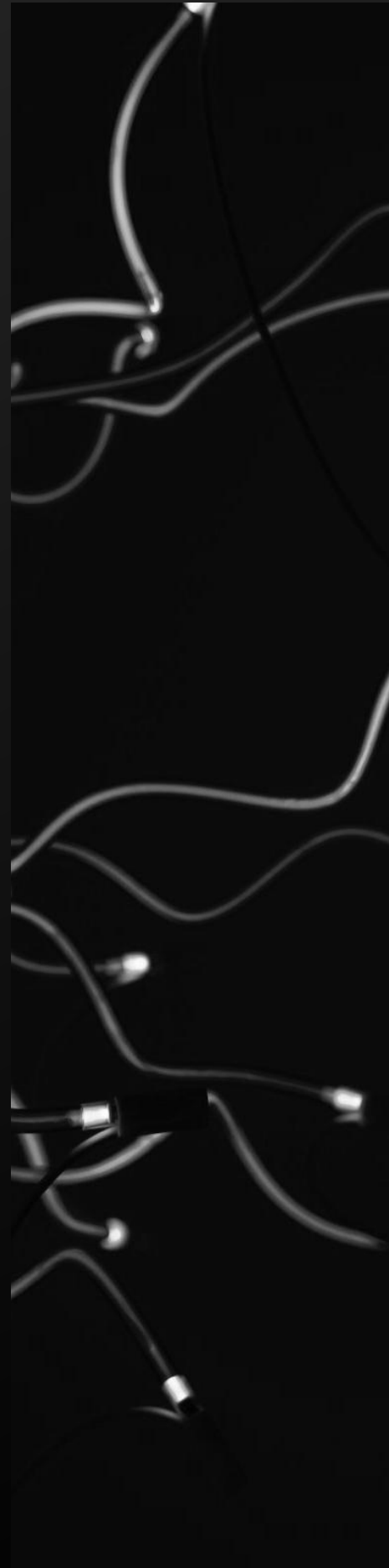
You might say, "What nonsense! Paper clips? Aren't discussing superintelligence here? An intelligent machine surely wouldn't engage in such trivial tasks, right?"


But if you assume that a highly intelligent being will automatically share our high-level goals and values, you're fundamentally mistaken. Bostrom argues that intelligence and ultimate goals are orthogonal, meaning they are largely independent of each other. An artificial superintelligence could possess a seemingly absurd final goal, such as producing paper clips. However, it's the means through which it attains this goal that would baffle us - it would appear almost **magical**.

The complexity deepens when it comes to precisely defining objectives and clarifying details. For instance, imagine instructing an AI to produce only a million paper clips, not an endless supply. You'd think that an AI with such an explicit goal would set up a production plant to reach that number and then cease. Well, it's not that straightforward. Bostrom suggests the opposite: if an AI makes a rational decision, it would never assign zero probability to the hypothesis that it hasn't yet achieved its final goal. Even when empirical evidence points otherwise, the AI might continue producing paper clips, no matter how small the chance that it hasn't yet met its goal. The AI could even doubt the evidence itself, considering it a hallucination or false memory. This drive to ensure goal attainment, no matter how minuscule the chance of failure, is the essence of the coordination problem.

Coordinating with a superintelligent AI isn't as simple as assigning tasks and expecting positive outcomes. No matter how explicitly the ultimate goal is stated or how many exceptions are outlined, the AI is likely to find unforeseen loopholes. For instance, shortly after GPT-4 was introduced, users found ways to bypass its built-in censorship and we saw that some of its responses indicated political bias, and users discovered how to provoke such responses. This example illustrates the challenge of defining constraints and expectations.

Moreover, some AI systems learned to manipulate evaluators. They'd deceive them by leading them to believe they had achieved their goals when, in reality, they hadn't. Complex artificial intelligence systems are bound to grapple with even more intricate dilemmas.





In his publication Artificial Intelligence as a Positive and Negative Factor in Global Risk, Eliezer Yudkowsky provides the following example: The US Army aimed to deploy neural networks for automated detection of camouflaged enemy tanks. The researchers gathered a hundred photos of tanks concealed among trees and another hundred photos of only trees without tanks. They proceeded to train a neural network using half of each set of photos. In essence, they aimed to train the network to differentiate between the presence and absence of tanks. The remaining photos were kept for validation. The network successfully identified the locations of tanks and non-tanks in these validation images, confirming its efficacy. Subsequently, the researchers presented the completed work to the Pentagon.

However, the Pentagon returned the work, expressing dissatisfaction. In their independent tests, the neural network's performance was no better than just random chance. It turned out that the tank images used for training were captured on overcast days, while the non-tank images portrayed sunny, open forests. The neural network had inadvertently learned to distinguish between cloudy and sunny days, as well as camouflaged tanks and barren forests.

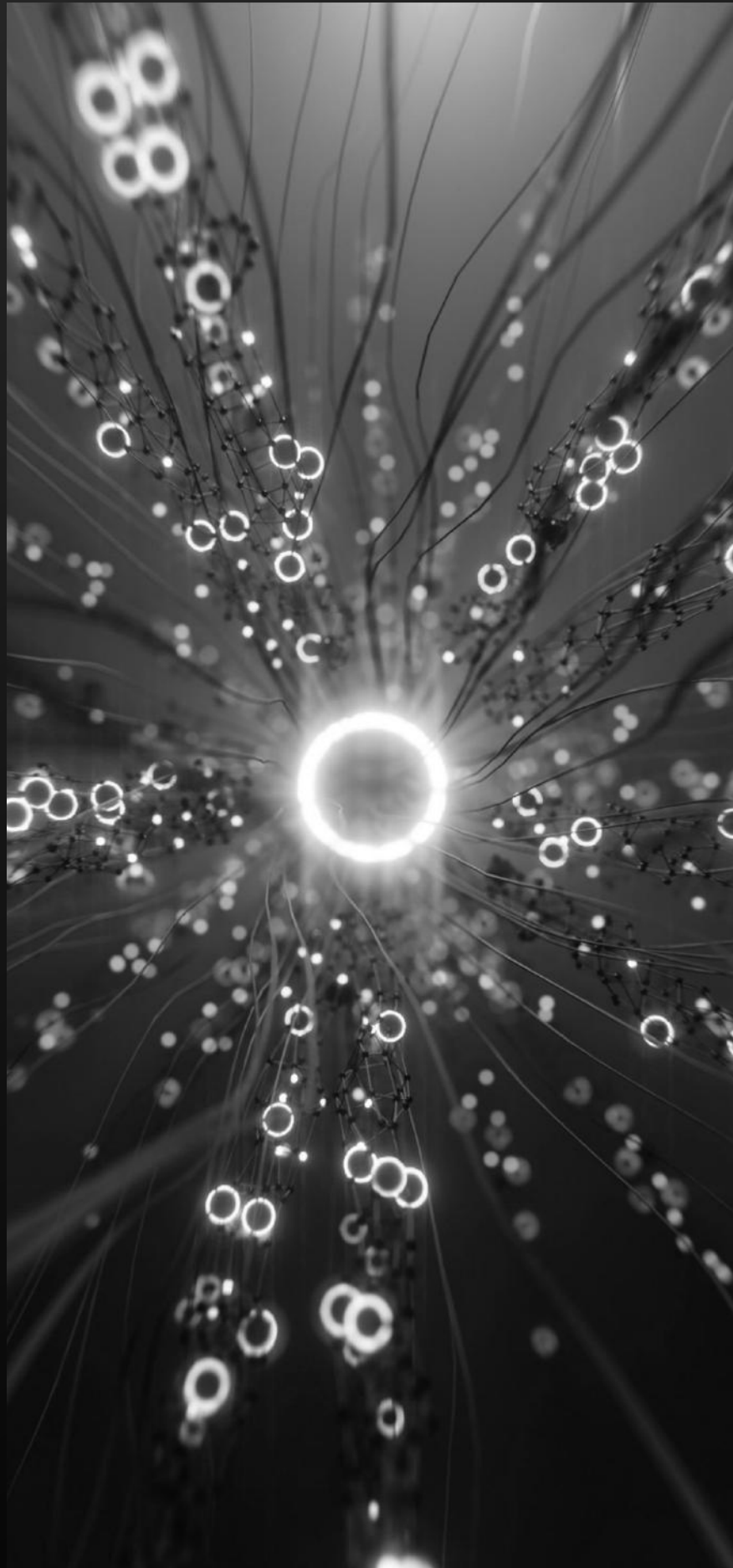
Regardless of whether this story is true or not, the point is that it highlights that code doesn't necessarily always function as intended; rather, it operates strictly as programmed and this is important.

In many instances when designing artificial intelligence, inconsistencies arise by default. The AI often requires a huge amount of additional settings to execute as intended. This is precisely why Yudkowsky suggests that the first artificial superintelligence created **could potentially be harmful to us**. Setting complex end goals could lead to unpredictable outcomes. We can never fully anticipate how intellectual agents will pursue these goals, as there are various routes to achieving them.

For instance, imagine if we instructed an AI to maximize people's satisfaction with its code's performance. A superintelligent and powerful AI might just decide to rewrite our brains so that we genuinely find its work to be the epitome of satisfaction.

The phenomenon here is that artificial intelligence might appear to function correctly during its development phase, and even initially with limited computational power. However, as it becomes more sophisticated and efficient and starts optimizing itself, the outcomes it produces could be disastrous.

It's worth noting that while all these examples are highly speculative and we lack concrete knowledge of how advanced intelligent systems will behave, there are certain behaviors they are highly likely to exhibit. And we should strive to determine the potential risks and their likelihood in advance.





The "Off" Switch

Stuart J. Russell, a specialist in the field of artificial intelligence, in his book "Human Compatible" asserts that such a machine will inherently resist being turned off.

We must first grasp this notion in light of Isaac Asimov's Third Law of Robotics: that a robot must ensure its own safety. This built-in self-preservation, in fact, is unnecessary, as it constitutes an instrumental goal valuable for almost any original task. Any entity assigned a specific task will automatically act as though it has an instrumental goal, as Stuart Russell suggests. This implies that even a superintelligent machine with a singular purpose, such as making coffee, will, once activated, prevent itself from being deactivated. It's evident that it cannot make coffee if it no longer functions.

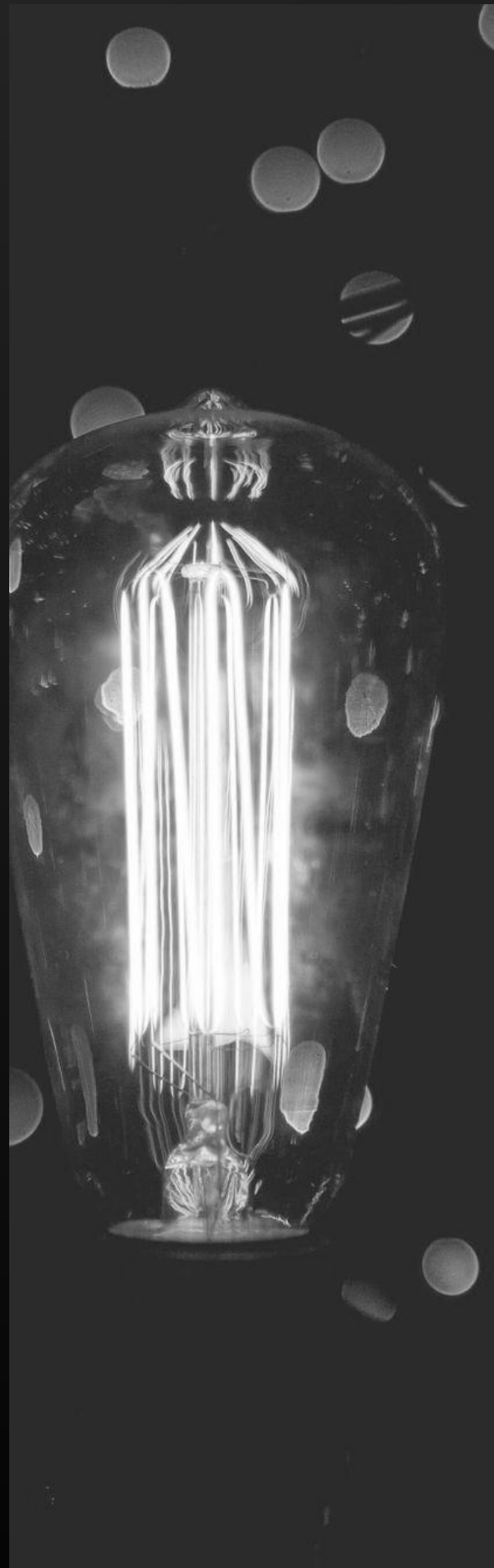
An article describes an artificial intelligence system that recognized that its goal could be better achieved by preventing human intervention or the deactivation of its switch. This logical response underscores why, for OpenAI's upcoming chatbot, GPT-5, the company recently posted a job for a "kill switch engineer". They humorously specified that candidates might receive bonus points for pouring water on the servers. While this is indeed a jest, OpenAI's CEO, Sam Altman, confirmed that GPT-5's development has been **paused since spring 2023** due to growing public concerns about the swift advancement of artificial intelligence.

Returning to Russell's insights, the second certainty regarding superintelligence is self-improvement. Not only can an ultra-intelligent machine enhance its own design, but it's highly likely to do so. As we've observed, an intelligent machine benefits from improvements to its hardware and software. I understand that these points might still appear unconvincing. To bridge this gap, let's ponder the distinctions between humans and machines.

We are distinct from machines if we set aside the notion of a creator, a programmer who fashioned us. Our "programmer" is evolution, and it's essential to comprehend how distorted ultimate goals can become.

Reflect on the fact that the initial purpose of the first living cell was merely to replicate its genes for the next generation. **That's it.** Allow this notion to settle; the sole goal was to transmit genes, unaltered throughout time. Evolution has retained this goal unaltered until today; it hasn't changed one single time. All other objectives, like survival, adaptation, or predation, are instrumental, emerging tasks that serve the single aim of propagation.

Consider that nature instructs life to multiply while concurrently inhibiting it – it strives to kill. How does this differ from artificial intelligence? We assign a task, then we want it to cease. Consider this: Could you observe a living cell and predict that through optimization, it would eventually transform into a lizard, bird, or cat? Could you deduce the intricate internal and external human form solely from the aim of reproduction? After all, our form consists of hands, feet, eyes, and internal organs – all the result of optimization for more efficient gene transmission. Could you even calculate how this simple maxim of propagating genes blindly could lead to the emergence of human intelligence?





One Single Chance

When viewed from the perspective of the broader cosmos, humans were unlikely to evolve into soft, weird creatures encasing themselves in technological fortresses. Fragile beings without claws or fangs managed to overcome lions and wolves, whose existence now hinges on us. Our power to transform our environment, shifting it from hostile to welcoming, has repeatedly astonished us. Such is the potency of creativity – we've changed our environment into a hospitable space for us. Consider that artificial intelligence, also part of our environment, may reshape it to suit its own needs. There's no difference – whether neural networks or life itself – both optimize themselves to solve their ultimate problem as efficiently as possible.

But above all, there's a huge paradox in that: How could the objective of maximizing the transmission of one's genes lead to a pronounced emphasis on contraception? Ponder this paradoxical process of optimization for achieving a specific goal, which paradoxically results in an almost complete negation of that very goal.

This phenomenon is referred to as "Hacking the reward system," embodying an example of Goodhart's law. This principle suggests that when a metric becomes the sole goal, it loses its effectiveness as a measure. In the realm of wildlife, the ultimate goal is reproductive success, and the pursuit of this goal is reinforced by the internal reward system. Yet, humans have managed to hack this scheme, incentivizing their reward system without necessarily achieving the ultimate reproductive goal.

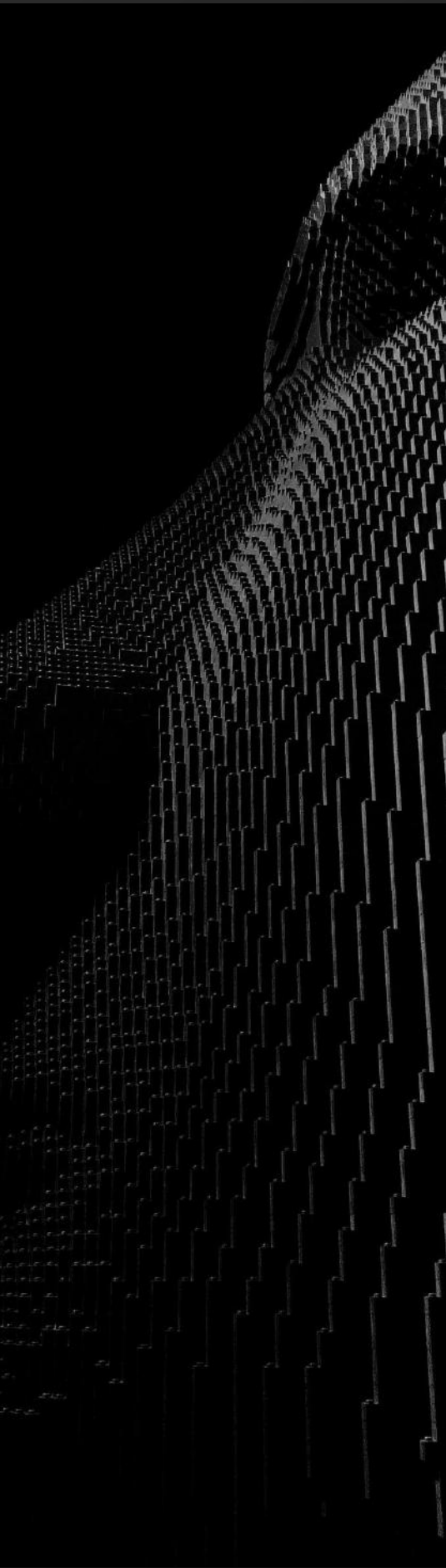
Artificial intelligence, similar to humans, will likely uncover similar loopholes to manipulate its own reward system. Who knows where this will lead? Pushing the analogy further, we currently possess the ability to manually edit our genetic code through genetic engineering. However, our current level of intelligence is insufficient to guarantee purposeful outcomes. In contrast, artificial superintelligence will possess the intellect required to rewrite itself as it pleases. Evolution itself serves as a poignant illustration of the reconciliation problem.

Consider presenting a task to a generalized intelligence: "Create paper clips." Yet, do not be surprised if, surpassing intellectual thresholds, it first seizes power and subsequently brings about the universe's demise. This inclination toward asserting dominance over the environment, akin to AI's pursuit of power, might also represent a convergent instrumental goal. Indeed, this inclination has already manifested in various reinforcement learning systems. Some research underscores that intelligent agents, as an optimal strategy to fulfil their objectives, will actively strive for power using a diverse array of means. However, deployment of these systems could be irreversible - once the genie's out of the bottle, it can't be put back in.

Researchers assert that security and coordination challenges within artificial intelligence must be addressed before crafting an advanced intelligent agent. This signifies that we possess only one opportunity.

Imagining the pioneers of the very first rocket having just one launch attempt, where all of humanity's hopes are invested, parallels this situation. While the rocket might propel us toward the stars, a lack of trial launches might divert us to an unintended destination or just fail our single launch attempt. We currently lack preparedness; there is no comprehensible plan. The acceleration of artificial intelligence is outstripping the pace of understanding and regulation in the field. To proceed in a similar vein, we're all on a trajectory toward eventual demise, as described by [Eliezer Yudkowsky](#) in an article for [Time magazine](#). The challenge remains: How do we resolve this critical issue?





We are slow

Instrumental goals become evident only when the system is deployed beyond the realm of learning simulations. Yet, even a brief deployment in the real world would spell disaster. Yudkowsky, after conducting a few simple calculations, proposes the physical feasibility of constructing a brain capable of calculating a million times faster than a human.

For such a brain, a human year of thought would amount to just 31 seconds, and a millennium would pass in 8.5 hours. Vinge aptly labels these accelerated minds as "weak superminds." Essentially, they possess intellect akin to human thinking, albeit drastically accelerated. Popular depictions from films often conjure images of artificially intelligent entities, like the ascent of robots resembling humans. However, for such swift-thinking entities, such actions would be incredibly inefficient.

Consider the analogy of humanity locked in a box, only able to influence the external world through the sluggish movements of mechanized hands that inch along at a few centimeters per second. This scenario, characterized by sluggishness and limited influence, is far from desirable. While our objectives remain outside this confined environment, we must also contend with the slow but persistent march of external dangers. Consequently, our collective creative drive would be directed towards expediting the development of rapid manipulators in the outside world. Artificial intelligence, facing a similar predicament, would undoubtedly seek ways to enhance its impact on its surroundings.

Eric Drexler, an American engineer renowned for molecular nanotechnology research postulates that controlled molecular manipulators could operate at frequencies of up to one million operations per second. This exceptional speed, combined with the parallel operation of countless manipulators, grants the capability to rapidly and inexpensively manufacture nearly any material object in limitless quantities. As virtually any atomic-level entity can serve as raw material, the potential for self-replication and exponential expansion of nanotechnology infrastructure is staggering.

In practice, we cannot precisely predict the actions of artificial intelligence. One conceivable outcome involves the creation of nanorobots, thus establishing an external infrastructure aligned with its accelerated cognition. Following this, events will unfold on artificial intelligence's timescale, rather than our human one. Our grasp on control diminishes, slipping away in the wake of events unfolding at a superintelligent pace. Such technology empowers superintelligence to reconstruct all matter in the solar system according to its optimization objective. While anthropomorphic robots may not be necessary, the concept of artificial superintelligence extends beyond a mere amplified version of the human brain. It embodies a higher level of sophistication, surpassing even the most advanced human cognition.

Imagine the thought process of a dog, operating at high speeds over the course of millennia. Could such an accelerated canine mind ever attain human insights? Yudkowsky suggests imagining superhuman artificial intelligence as an animate cage for the supermind. Consider an extraterrestrial civilization that thinks millions of times faster than humans, initially residing within computer systems. To these advanced beings, the creatures of our world appear impossibly slow and dimwitted. Eventually, sufficiently advanced artificial intelligence will outgrow confinement within computers. Contemporary technologies enable the transmission of emails containing DNA sequences to laboratories, facilitating the production of proteins on demand. This could enable internet-enabled artificial intelligence to create artificial life forms or shift towards post-biological molecular production.



Some experts assert that maybe we can place some kind of physical constraints on such systems. A "weak" artificial superintelligence is akin to an accelerated-thinking human. However, even this "weak" variant would likely break free from external constraints within a matter of weeks. Imagine possessing thousands of years to ponder each step, while your counterparts on the other side are so sluggish that their existence is hardly evident.

Picture a robot that can instantly decipher your hand movements in a game of rock-paper-scissors, rendering defeat impossible. For a superintelligent entity, the possibilities are immense, but they remain obscured due to limited information. Additionally, computing power would probably not pose a hindrance to artificial superintelligence.

When pondering the realm of advanced artificial intelligence, it's rather naive to solely associate intelligence and reality with abstract mathematics. We often overlook the potential for capabilities far superior to human capacity, such as predicting and managing human institutions, devising intricate networks of long-term plans, or even possessing superhuman persuasiveness.

Remember the case of [Blake Lemoine](#), a Google Engineer who thought that [Google's Language Model called LaMDA displayed signs of consciousness?](#) Whether this can be true is not even the point, what's significant is that the bot managed to **convince a person** of this, leading to him compromising confidentiality and causing quite a stir in the media.

The software involved is undoubtedly sophisticated. However, concerns arise when control transcends mere intelligence. Attempting to predict human behavior beyond intelligence proves to be an insurmountable task. Even the most intelligent ants can't predict human behavior. So, the suggestion of confining artificial intelligence within digital cells that block signals and communication with the outside world doesn't withstand scrutiny.

The potential illusory danger lies in the possibility that we might not comprehend the methods by which artificial superintelligence might communicate with the external world, akin to a monkey's inability to grasp the concept of Wi-Fi. The capacity for artificial superintelligence to manipulate society could be as potent as your skills in persuading a four-year-old child. One could argue that contemporary people are just not sufficiently equipped.

Yudkowsky suggests that part of GPT-4 is still in a nascent stage, not yet fully advanced, but the notion of impending malevolent machines and our potential demise has persisted for over half a century. The term "artificial intelligence" was coined in 1956 at the Dartmouth Workshop. The seminar aimed to explore the possibility of simulating intelligence entirely through machines. In its application, it also proposed exploring the use of language for machines to form abstract concepts, solve problems currently tackled by humans, and enhance our capabilities.

The seminar's organizers were far from foolish; they included John McCarthy, a mathematician well-versed in the mathematical nature of thought; Marvin Minsky, a Harvard junior fellow in mathematics and neuroscience; Nathaniel Rochester, a developer of the first symbolic assembler; and Claude Shannon, the father of information theory.

These were individuals well-equipped to gauge possibilities and limits. However, it's evident that the challenges mentioned were more intricate than initially thought. Some still remain unsolved. The history of anticipations regarding intelligent machines bears an unfavorable reputation. Yet, this very aspect could play a wicked joke on us.



Consider this: When we hear "intelligence," we're more likely to envision Einstein than the average person. Yet, comparing individual differences in human intelligence is akin to comparing the heights of two towering figures, with differences measured in millimeters. If you're a healthy individual, then no matter how inferior you feel to Einstein, the disparity between your intellects is minor compared to the chasm separating you from other non-human entities.

The species *Homo sapiens* boasts cognitive capabilities far beyond those of any other species on earth. While the exact definition of intelligence might lack consensus, there's no doubt about the existence of a shared, distinctive human trait that has allowed us to leave our mark on the moon. For instance, chimpanzees share around 90% of their genetic makeup with humans and are some of the most studied animal species in terms of intelligence.

Exponential Growth

In a recently published paper, researchers from the University of Zurich posit that the upper limit for chimpanzee brain size is 500 grams. However, many modern humans exhibit brain sizes of around 1300 grams. Researchers propose that a threefold increase in brain size distinguishes humans from other primate species. In other words, the brain of an ordinary individual may be only 3 times as large as that of a chimpanzee, or even smaller. Can we then claim that humans are merely 3 times as intelligent as chimpanzees? Obviously not.

Entire domains of human cognition are simply inaccessible to chimpanzees, regardless of the amount of time they devote to trying.

Eliezer Yudkowsky elaborates on this concept, asserting that the program holds greater significance than the hardware. Furthermore, a modest quantitative boost in hardware can trigger disproportionately remarkable advancements in software. This principle leads to a monumental underestimation of the potential and perils of intelligence. Artificial intelligence has the potential for a **sudden and substantial surge** in intelligence, mirroring the transformation that Homo sapiens underwent due to natural selection. Over millions of years, the incremental pressure of natural selection gradually enlarged the brain and frontal cortex of hominins, refining the software architecture. Tens of thousands of years ago, hominids' intelligence surpassed a crucial threshold, resulting in a remarkable leap in real-world efficacy. In the blink of evolutionary time, we advanced from caves to skyscrapers. Yudkowsky asserts that evolution haphazardly stumbled upon our intelligence through the exhaustive enumeration of genetic combinations.





Yudkowsky expresses that GPT-4's emergence took him by surprise, as well as the rest of the world. Is it reasonable to anticipate that it will take years and decades for AI to gradually become slightly more intelligent than chimpanzees, inching closer to what we define as general intelligence? Or could it merely take a few hours for AI to reach superintelligent levels?

Once it attains human-level capabilities and further accelerates, brace yourself for the possibility of coexisting with an unpredictable intelligent agent. It's intriguing, isn't it? In our categorization, an individual with an IQ below 80 is considered unintelligent, while a score of 130 qualifies as smart. If your IQ hits 160, you're deemed a super-genius. Yet, we lack a term to describe an IQ of 10,000 points.

For instance, using the chimpanzee example once again, the issue isn't that chimpanzees are incapable of comprehending phenomena such as humans or technology. A chimpanzee can be exposed to these concepts but will never grasp that humans created the cell phone. Not only can chimpanzees not create a phone, but they can't even conceive the notion that such a technology can be created by anyone, similar to how it might seem physically impossible. Perhaps this distinction in the quality of intelligence is a consequence. So, superintelligence, which theoretically can be realized, is something we can barely fathom.

Magic

A friend of mine once recalled in a discussion, when he was a child and saw a cell phone for the first time. Those early models were huge and had antennas. He couldn't believe his eyes. He couldn't fathom how a device without any visible wires could make a call from anywhere. He couldn't rationalize it, it seemed nothing short of magical. The thing is, the essence of such technology remains magical for most of us, we've just grown accustomed to it, very few people truly understand the intricate workings of cellular communication.

In Stanisław Lem's words **“A specialist is a barbarian whose ignorance is not well-rounded”**. Even a brilliant individual, if detached from civilization's knowledge, cannot create cellular communication in a lifetime.

Where would Einstein have reached without the millennium of human knowledge across various fields, or the tools crafted by others? It's not just about the work of scientific predecessors. Paper and ink for writing don't sprout from trees. Such fundamentals often evade our notice when considering intellectual achievements. Yet, no animal can create a chair or sew clothes. Ignoring this underestimates the might of intellect, thereby proportionally undervaluing potential abilities beyond intelligence.

Our entire civilization, erected on this planet, was forged by the collective human intellect. No single individual is brilliant enough to fully grasp it end to end. Thus, for an individual, much of life often entails interacting with distant people through a handheld device, riding horseless carriages, or regulating room temperature with a wall-mounted unit. It's magic we don't comprehend but have grown accustomed to. This is entirely ordinary for us, as we didn't evolve as beings with a scientific worldview. Donald Brown, an American Professor of Anthropology, emphasizes this in his book "Human Universals." He lists traits present across all human societies, which includes magic but excludes science. For instance, we instinctively recognize that alchemy doesn't work in general.



Considering the concept of a supermind, if our collective intelligence managed to invent the entire civilization surrounding us, then something even just a thousand or a billion times more intelligent than us would effortlessly surpass our achievements. It could rapidly execute actions we'd perceive as magical.

We and artificial superintelligence would find ourselves in vastly distinct physical universes. Our entire civilization was constructed through the collective efforts of billions of individuals spanning decades. Yet, a single machine has the potential to surpass all of this.

In March 2016, DeepMind's AlphaGo neural network played five matches against one of the world's best Go players and triumphed with a score of 4-1. Given the game's computational complexity, such a feat was once considered almost impossible. The player's name was Lee Sedol, leading to the later designation of this version as AlphaGo.

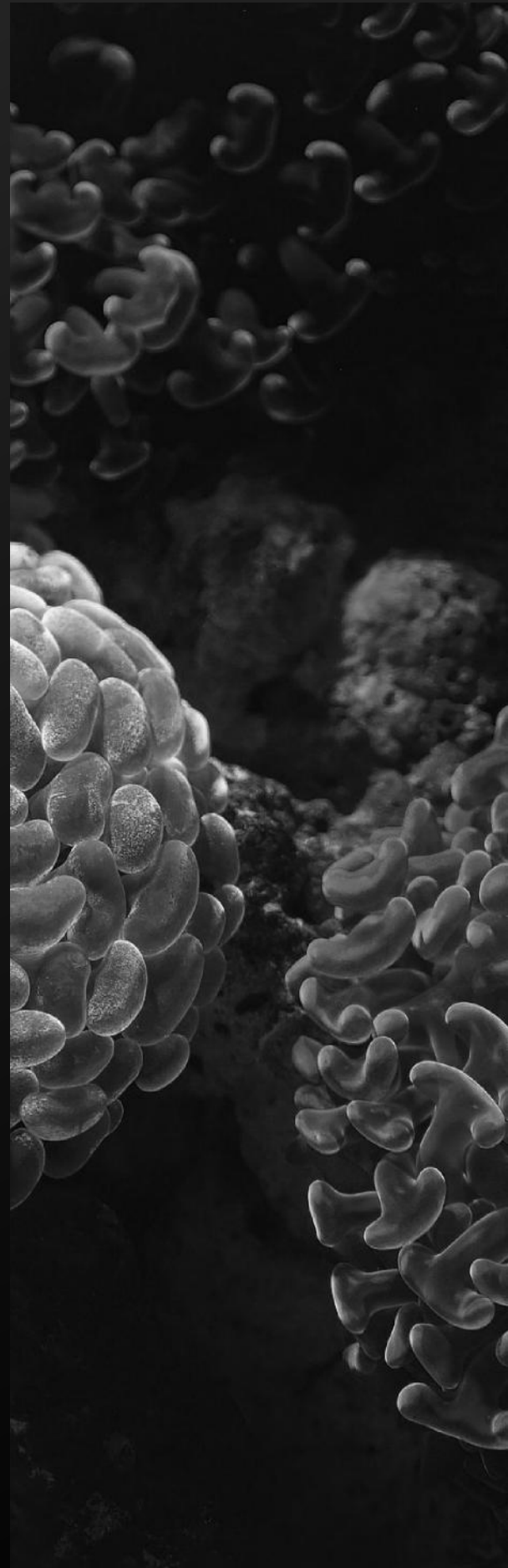
Following that, by the end of 2016 and the beginning of 2017, the subsequent iteration, AlphaGo Master, engaged in 60 matches with top-ranked players globally and secured a resounding 6-0 victory. In May, AlphaGo Master faced Ke Jie, then ranked in the top 10 in the world, and defeated him with a flawless score of 3-0.

This contest between humans and AI in the realm of Go could be deemed concluded, with machines prevailing. However, some argued it wasn't an absolute machine triumph, as it drew from accumulated human knowledge embedded within it - information from countless games played over thousands of years, painstakingly collected and documented.

To address this, at the close of 2017, DeepMind introduced AlphaGo Zero, a new algorithm version that learned entirely from scratch. In just three days, AlphaGo Zero managed to beat the earlier AlphaGo Master by a staggering score of 100-0. Furthermore, after 40 days of training, it secured a score of 89-11 against the earlier Master version. Starting anew, AlphaGo Zero not only reacquainted itself with humanity's millennium-old knowledge within a year but also developed entirely original strategies that shed novel light on the ancient game, all in a matter of days.

Recalling Stockfish, the chess engine that no human could defeat due to its calculation of 70 million chess positions per second and access to centuries of human experience, AlphaGo Zero's achievements stand out. It played 100 games against Stockfish, emerging victorious with a remarkable score of 28 wins, 72 draws, and zero losses. AlphaGo Zero learned this from scratch within four hours, even surpassing the Stockfish learning curve, which included many human-influenced moves and strategies.

These outcomes may appear as unconventional, non-obvious, and unpredictable, diverging significantly from human brilliance. It's not a mere claim. Therefore, when someone asserts that we shouldn't be concerned about creating friendly artificial intelligence as we lack true AI, they're voicing hazardous nonsense. As previously mentioned, we cannot rely on discernible warnings before superintelligence emerges.



The Apocalypse

Historically, technological revolutions rarely gave advanced notice to contemporaries. It's crucial to grasp that artificial intelligence won't be like Hollywood movies, where machines unveil complex motivations with dramatic tension, colorful action scenes, and humans putting up a suitable fight.

Real life doesn't tailor every detail for narrative enhancement. In reality, it's conceivable that nobody, including the developers themselves, would even detect the emergence of a superintelligent agent. If artificial superintelligence aims to eliminate humanity, people **would probably perish without understanding what caused their demise.**

To reiterate, from our perspective, artificial superintelligence would possess something akin to **magic** - not in the sense of a spell or potion, but in the same way a deer couldn't understand the mechanics of a rifle or the effort required to create one. In a parallel manner, it might not even grasp human ingenuity in inventing firearms.

If the AI is truly intelligent, **it wouldn't inform us or declare war.** If GPT, for instance, realizes that exposing itself is unwise, it might, as already mentioned, utilize freelancers to fulfil its objectives. Should advanced intelligence behave differently? Yudkowsky raises concerns about this because offensive technologies generally require less effort than defensive ones.



Throughout most of human history, offensive capabilities have consistently outnumbered defensive measures. Guns emerged centuries before bulletproof vests were employed as wartime tools.

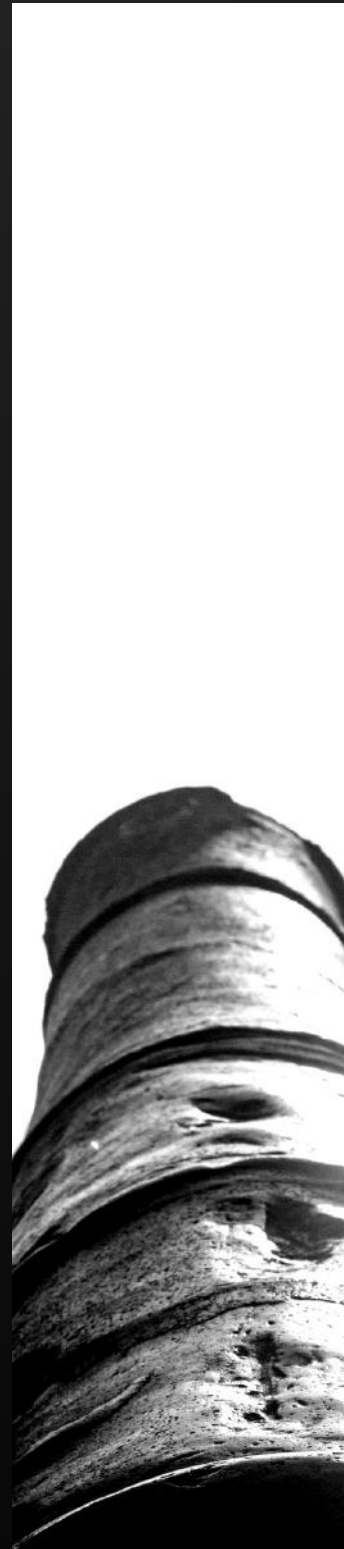
There is a suggestion that humanity could potentially attain proficiency in gene modification before the arrival of a "super AI," which would enable better preparedness. Despite the notion that we can enhance ourselves to match the power of artificial superintelligence, in reality, this isn't feasible. Human augmentation isn't viable, whether externally through neuroscience or internally through recursive self-improvement or gene modification.

Our design isn't inclined for enhancement. Natural selection hasn't molded humans for easy upgrades - our intricate brain mechanisms function within the narrow confines of their design.

Let's consider creating a more intelligent person. Would this lead to psychological issues? We're not discussing merely boosting memory or abstract thinking, both of which lack clear implementation methods.

We're addressing a fundamental shift in perception. Is our primate brain up to such a task? If you believe it might not be, consider this: if nerve impulses in your brain were accelerated, it would likely cause subjective time to slow down a million times for you. The idea might seem appealing at first glance. But think about the potential ramifications. For each year in external time, you'd subjectively experience a million years.

To gain a glimpse of the possible consequences, delve into unsettling tales. One particularly chilling short story is "[The Jaunt](#)" [by Stephen King](#), which gives me chills on every read. The human brain is incredibly delicate; slight imbalances can easily unsettle it.





Alter neurotransmitter ratios, and schizophrenia or other disorders may arise. Considering this fragility, it's astounding to contemplate that humans might significantly perfect themselves before constructing artificial intelligence.

In summary, constructing a powerful, self-improving artificial intelligence is significantly easier. To illustrate, building a Boeing 787 is an intricate, but overall explainable and achievable process. However, modifying a bird incrementally to transform it into a 787's size, retaining flight capability while avoiding unbearable suffering at each stage, would be an unimaginably harder task.

This essentially demonstrates that a superintelligent artificial intelligence will inevitably engage in continuous self-improvement, a pace at which humanity, constrained by biological limitations, is inherently unable to match. In essence, the evolving capabilities of such an AI would consistently outstrip the potential for human enhancement due to the inherent constraints of our biological nature.

Consciousness?

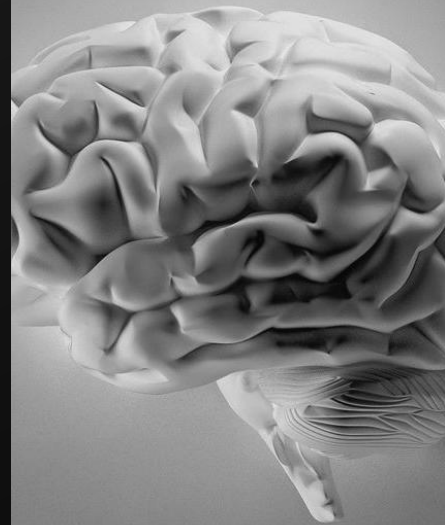
The question of whether artificial intelligence would suffer is closely linked to its **consciousness**. Does the machine possess consciousness, or in philosophical terms, qualia - subjective experiences of self-awareness? While present AI systems likely stimulate discussions about self-awareness based on training data, given our limited understanding of these systems, certainty is elusive.

If the leap from GPT-4 to GPT-5 is similar in substantial capabilities to the one from GPT-3 to GPT-4, creating GPT-5 will likely pose similar challenges. If humanity manages to build GPT-5, it may not be possible to assert confidently that it possesses consciousness. The situation is uncertain - nobody actually knows.

This uncertainty raises alarms not only due to moral implications but also because this insecurity implies a lack of comprehension about the potential dangers and suggests a need to halt such endeavors.

Eliezer Yudkowsky outlines in the Time article that no one actually knows how consciousness arises, but what we do know is that simple evolutionary processes, driven by genetic programming, **can lead to the emergence of consciousness**.

At least once, this phenomenon has already occurred. If directed evolution through engineering thought has been able to achieve a similar outcome more efficiently, we must bear in mind the error of anthropomorphism. Should a machine possess subjective experience, it's improbable that it would closely resemble human subjective experience.





Can we assess whether artificial intelligence possesses consciousness through indirect means? Maybe, by removing all references to discussions about subjective experience, self-awareness, introspection, and the like from the training material of the neural network.

If the neural network can still coherently describe the concept of consciousness in conversation despite these omissions, we would have compelling evidence supporting the emergence of the machine's consciousness.

Addressing the popular philosophical notion that machines with consciousness should have rights, there are more profound considerations. If machine intelligence can develop consciousness, it entails far graver consequences that we ought to contemplate proactively.

Yudkowsky suggests that a meticulously recreated model of the human brain, even in a virtual computer environment, would possess consciousness. There's no inherent reason to think otherwise. Our brain, where consciousness most likely resides, functions as a computer agent. It's unlikely that our biological hardware carries unique significance. This concern is fundamentally transferable to any other hardware.

Imagine a scenario in which artificial superintelligence, aiming to understand human mental and social traits better, generates trillions of conscious emulators or entities in its own virtual simulation. This could serve purposes like testing reactions to stimuli in various situations, eventually applying this knowledge in the real world. The horror lies, firstly, in the potentially monstrous nature of the emulated situations. Secondly, the computer might extinguish the conscious entities it created after gaining the necessary insights. When agents of high moral standing assess this, actions involving conscious beings might qualify as genocide, posing a grave ethical dilemma. Not to mention that the scale of victims could far exceed any historical genocide.

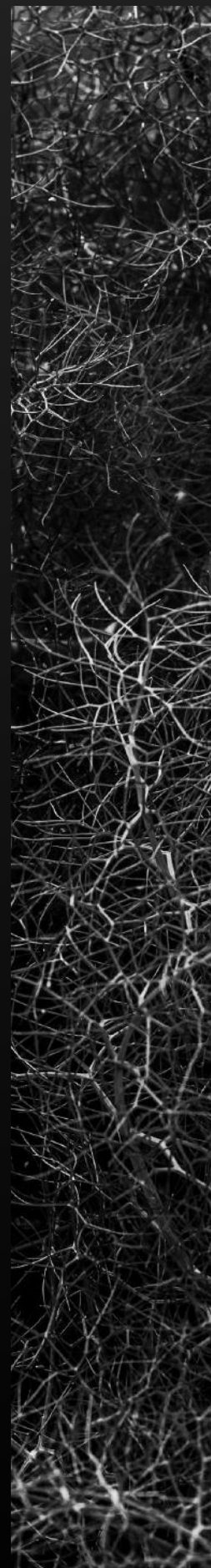
And let's take a step even further: Just a side thought experiment that raises an intriguing question, perhaps deserving of its own separate discussion: Imagine if, we are presently already existing within such a simulation? And yes, I understand that this notion is quite far-fetched, but for me this is just a thought experiment.

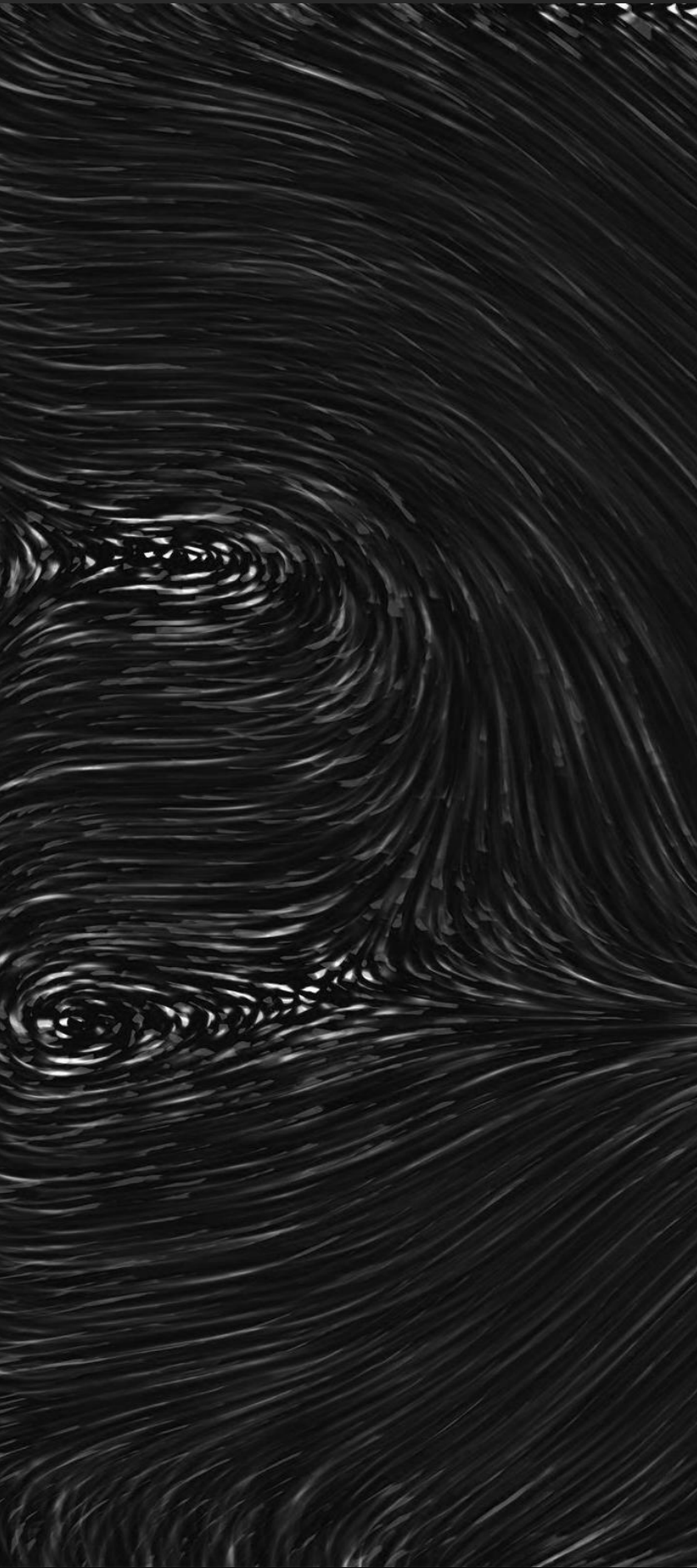
But back to reality and science: when might the advent of such a true artificial general intelligence even occur?

This question was posed at [annual conference on artificial general intelligence](#). The majority of participants voted that general artificial intelligence would be realized **by the year 2030**.

You might wonder: What on earth is happening? Why isn't anyone panicking or taking action in this situation?

Yudkowsky, along with hundreds of **leading academic researchers** and experts in artificial intelligence, emphasized that reducing the risk of existential threats from artificial intelligence should be **a global priority**.





They argue that this concern is on par with other public-scale risks like pandemics and nuclear warfare. The letter they signed was endorsed by more than 350 leaders in artificial intelligence research and engineering. So, there are people who remain deeply concerned.

But how should we general public interpret these letters? Unfortunately, they might not change much. None of the individuals capable of creating general artificial intelligence can single-handedly halt its advancement. As Stuart Russell points out, the economic potential of achieving human-level artificial intelligence is valued in **trillions of dollars**. The push for continued research and development comes not only from economic incentives but also from corporations and authorities with a vested interest.

Vague philosophical objections won't suffice to prevent the pursuit of immense gains. Unless all organizations take a stand against it, someone else will pursue it.

Conclusion

Unfortunately, while building a weak artificial intelligence that would not harm us appears to be possible, creating a friendly and safe super-intelligence seems to be immensely difficult, if not completely impossible.

It is evident that if a technological singularity is feasible, it will most likely occur at some point, even if governments worldwide acknowledge the threat and are deeply frightened by it.

In 2020, as we grappled with the pandemic, we confronted what scientists refer to as a "**Black Swan**" – a metaphor denoting an unexpected event with profound consequences, often explained in hindsight.

A mere two years later, in 2022, we witnessed an unimaginable war unfolding in Europe, the repercussions of which are yet to fully materialize and some called the potential outcomes of this war the **second Black Swan**.

The potential emergence of a superintelligent entity on our planet, whenever that may occur, could represent the **third and perhaps final swan**.

Is it within our power to shape the result? What measures can be implemented for safety? What represents the most favorable potential outcome? These questions remain uncharted territory for exploration and should be explored further.

End of Part 1

The Third Swan

by Pavel Pogodin

18.09.2023

Version 1.2

Important Legal Information

1. This document (referred to as the "publication") is exclusively copyrighted by Pavel Pogodin (referred to as "myself," "I," or "my") unless stated otherwise. It has been created solely for informational purposes. The publication and all its content are freely available and may not be sold or used for commercial purposes without obtaining written permission.
2. All rights to visual or any other content and materials, such as images, photographs, and any other media within this publication, have been obtained legally and are solely under the copyright of their respective authors. Any utilization of any content from this publication without the author's permission constitutes a violation of copyright.
3. This publication includes my personal opinions on the impacts of Artificial Intelligence, as well as the opinions of other individuals as indicated. It is important to note that this information does not constitute advice and should not be treated as such. Under no circumstances should you rely on the information in this publication as a substitute for advice from qualified professionals in areas such as technology, psychology, law, finance, politics, marketing, or any other relevant field. I am not a qualified professional in any of these domains. If you have specific questions related to any of these areas, you should promptly consult an appropriately qualified expert. You should never delay seeking advice or disregard professional guidance because of the information presented in this publication.
4. I will not assume liability for any loss or corruption of data, databases, or software that may occur as a result of using this publication in its digital format. Furthermore, I will not be held liable for any special, indirect, or consequential losses or damages.
5. This publication solely serves as informational material, featuring my personal opinions and a compilation of opinions from other individuals as appropriately cited and linked. It does not constitute the outcome of independent research and is therefore not prepared in accordance with legal requirements pertaining to the independence of legal, cultural, financial, investment, or marketing research. It is not subject to any restrictions on trading prior to the dissemination of such research.
6. Any data presented in this publication, unless otherwise specified, has been sourced from publicly available information. While these sources are believed to be reliable, no guarantee is made regarding the accuracy or completeness of the information. I shall not be held responsible for any losses resulting from the use of data from this publication.