

# The Compendium

*By Connor Leahy, Gabriel Alfour, Chris Scammell, Andrea Miotti, Adam Shimi*  
V1.3.1 - Dec 9, 2024

*Humanity risks extinction from its very creations.*

*Ideologically motivated companies are racing to build smarter-than-human AIs. Big Tech already backs them, and now nation states are getting roped in too. If they succeed in creating AIs outsmarting humanity, it would be game over -- no one knows how to keep control of these AIs.*

*The Compendium explains these extinction risks from AI, where they come from, and how we can fix them.*

*The Compendium is a living document, and we will update it over time as the landscape changes. We welcome your feedback, which you can provide through [this form](#).*

# Contents

<b>Executive Summary</b>	<b>3</b>
<b>(0) Introduction</b>	<b>5</b>
Foreword	5
Overview	9
<b>(1) The state of AI today</b>	<b>12</b>
Rapid AI progress is driven by resources, not insights	13
AI is grown, not built	16
The race to AGI is on, and potentially deadly	18
<b>(2) Intelligence</b>	<b>21</b>
Intelligence is mechanistic and it is possible to build AGI	21
What is intelligence?	22
Tools	22
Groups	23
Methods	24
A mechanistic model of intelligence	25
Applications to artificial intelligence	25
Against arguments of AI limitations	28
The AI Effect	28
The general issue with missing components	30
Thus, AGI	32
<b>(3) AI Catastrophe</b>	<b>33</b>
Current AI research leads to extinction by godlike AI	33
Without intervention, current AI research leads to AGI	33
Without intervention, AGI leads to artificial superintelligence (ASI)	35
ASI will exceed individual human intelligence	37
ASI will exceed humanity's intelligence	38
Without intervention, ASI leads to godlike AI	40
Without intervention, godlike AI leads to extinction	42
<b>(4) AI Safety</b>	<b>44</b>
We are not on track to solve the hard problems of safety	44
Defining alignment	44
Estimating the cost of solving alignment	46
Current technical efforts are not on track to solve alignment	50
AI will not solve alignment for us	54

<b>(5) AI Governance</b>	<b>57</b>
We lack the mechanisms to control technology development	57
Defining AI governance	57
Estimating what is necessary to control AI development	58
Current AI policy efforts are not on track to control AI development	60
Current AI policy efforts endorse the race to AGI	66
<b>(6) The AI Race</b>	<b>70</b>
The race to AGI is ideological, and will drive us to the exact dangers it claims to avoid	70
The AGI race is ideologically driven	71
Utopists: Building AGI to usher in utopia	71
AGI companies	71
Entente	75
Big Tech: Keeping a hand on the technological frontier	78
Accelerationists: Idolizing technological progress	79
Zealots: Worshiping superintelligence	81
Opportunists: Following the hype	82
These ideologies shape the playing field	82
The strategies being used to justify and perpetuate the race to AGI are not new	85
The Industry Playbook	85
Spreading confusion through misinformation and double-speak	89
Turning care into acceleration	91
Capturing and neutralizing regulation and research	92
Capturing AI regulation	92
Capturing safety research	94
How will this go?	95
<b>(7) A good future, if you can keep it</b>	<b>97</b>
Civic duty is the foundation of a response to AGI risk	97
Creating a vision and a plan for a good future	99
The authors' plan	100
Actions to help reduce AGI risk	102
Communication	103
Coordination	106
Civics	108
Technical Caution	109
<b>(8) Outro</b>	<b>111</b>

# Executive Summary

The Compendium explains the extinction risks from AI, where they come from, and how to address them.

The trigger to the current situation is the explosion of AI progress in the last 10 years, reaching near-human-level in writing, coding, art, math, and many more fields of human activity. This progress has been driven by deep learning, which differs from traditional software: modern AIs are grown by feeding them massive amounts of data and letting them evolve in response, not built piece by piece by humans. First, this means that researchers and engineers don't need to understand AIs to create them – indeed experts consistently fail to anticipate how quickly new skills will be unlocked, or how existing AIs work. Second, this makes continued progress bottlenecked only by resources (such as AI chips, electrical power, data...) instead of scientific insights. As tech giants and frontier AI companies collaborate to unlock ever more resources, the path forward leads to increasingly intelligent yet opaque AIs.

How far can this trend of smarter and smarter AI go? If we look for inspiration at how humanity has historically increased its intelligence (understood as its ability to solve more and more intellectual tasks), three strategies emerge: tools, groups, and methods. AIs can leverage the exact same strategies; indeed each of them is already being used by current AI R&D! And although skeptics claim that various hypothetical components of intelligence cannot be automated, no existing scientific theory of intelligence backs these opinions. Thus we should assume the trend towards smarter AIs will continue, eventually leading to Artificial General Intelligence (AGI), AIs able to do the same intellectual tasks as humans.

Since AGI would be able to do anything a human can, it would notably be able to do AI research, and so to improve its own intelligence and capabilities. This is not the only way forward, yet AGI companies and AI researchers are already pushing hard in this direction. Since software is cheaper, more efficient, and easier to correct than brains, AGI would improve far faster than any human, eventually reaching artificial superintelligence (ASI) surpassing humanity's collective intelligence. As it continues to scale, ASI would unlock abilities to shape matter and energy that would look godlike compared to human engineering. And even without malicious intent, these godlike-AIs would by default wipe out humanity as collateral damage while pursuing their own goals, in the same way ants are just collateral damage for contractors building a house.

Godlike-AIs lead to catastrophe because of the incredible difficulty of aligning AI's goals with those of a single human, let alone the goals of humanity in general. Alignment is the harder version of the kind of problems with which humanity already struggles: for example

making companies and governments beneficial for what the actual citizens care for and believe in. Solving alignment would require massive progress on questions like finding what we value and reconciling contradictions between values, predicting the consequences of our actions to avoid unintended side-effects, and design processes from the people's will to AIs doing the correct things. If we were serious about solving alignment, it would require at the very least decades of top-notch research and trillions of dollars of investment; yet only a handful of people and a couple 100s of millions are currently invested, with most of the money and effort going instead towards making AIs more powerful. Even worse, the little work that exists doesn't even try to pay the cost of alignment: instead it reacts to current issues with AIs by patching them in a whack-a-mole fashion, and passes the buck to future smarter AIs. Thus we're not on track to get anywhere near solving alignment, and thus godlike-AI would cause human extinction.

Lacking a solution to alignment, we need to ensure godlike-AIs are not built. This requires institutions with the authority to regulate frontier AI research, both at the national and international level. Yet these institutions simply do not exist, very little is being done to create them, and the little existing governance work already finds itself undermined by the very AI companies racing to AGI.

This lack of promising effort on alignment and regulation is not a coincidence: frontier AI companies are systematically undermining these to race to AGI without blockers. The root of this behavior lies in the ideologies of frontier AI companies: they are by and large utopists, who want to build AGI because they believe it will usher their ideal world. This belief brings with it a fear that AGI will be built by the "wrong" people, and so these utopists become more and more willing to cut corners in order to avoid this, undermining safety along the way. In practice, this looks like the key tactics of the industry playbook used by Big Tech, Big Oil, and Big Tobacco: spreading fear (by stoking geopolitical fires) and doubt (by changing their stances constantly) to free their path to AGI, capturing regulatory efforts under cover of self-regulation, and undermining research that might force them to slow down.

So we are on a dark trajectory, one that recklessly leads to human extinction. What can be done about this? We believe that giving up and reading the situation as hopeless is a mistake, one that the utopists and other actors racing to AGI want us to make. Instead, there is a narrow path forward, one that starts with basic civic duty. The people racing towards AGI are only a tiny minority, who are deciding to put everyone at risk to follow their delusions. Because no one in their right mind wants humanity to go extinct, this is an issue that can unify people across party lines and countries. This starts by spreading the word and awareness of the risks, and exercising basic civic duty by contacting your representatives and voting according to the extinction risks posed by AI.

# (0) Introduction

## Foreword

A few million years ago, something very strange happened.

Through minor genetic tweaks, an ancestor of the modern chimpanzee split into a new line of species: *Homo*, humans. This new chimp variant was odd in several ways: it learned to stand upright, lost most of its fur, and grew a bigger brain. This bigger brain was not really all that different from that of his chimp cousins, just scaled up by a factor of about three.

If you had seen this odd, half naked chimp with a brain three times bigger than its cousins', and you would have to guess what this new chimp will do, what would you have said?

Maybe you would have expected it to be a bit better at collecting termites, or throwing rocks more accurately, or have more complicated status hierarchies. But that 3x scaled up chimp ends up building nuclear weapons and going to the moon. Chimps don't go one third of the way to the moon, they go zero to the moon; humans go all the way.

We still don't exactly know how or why this happened, but whatever it is that happened, we call the result General Intelligence. It is what has allowed our species to build the magical glowing brick that you are looking at right now to transmit the words of another chimp descendant located halfway across the world to your eyes and brain.

This is *crazy*.

General Intelligence is what separates human from animal, industrial civilization from chimpanzee band. It probably isn't a discrete all-or-nothing property, but it sure is suspicious that you go from "zero going to the moon" to "all of going to the moon" within a 3x difference in brain size. Things can change quickly with scale.

Our intelligence makes us the masters of the planet. The future of chimpanzees is utterly dependent on what humans want to do with them. If we want to give them infinite food, incredible medicines they can't hope to understand, and safety from any predators, we can. If we want to keep them in zoos, or hunt them for sport, we can. If we wanted them extinct, their habitats paved over with parking lots and solar cells, we could.

This kind of relationship, of complete domination over another, is the natural balance of power between a much more intelligent creature and a less intelligent one. It's the kind of

power an adult has over a small child, or an owner over their pet. The arrangement may or may not be beneficial to the weaker party, but ultimately, the more intelligent and powerful agent decides the future. A pet doesn't get a say in whether they get spayed or not.

Luckily, there are no other species out there running around that might be even smarter than us.

-

But that is changing.

Currently, the future belongs to humanity, for better or for worse. The planet and stars are ours to do with as we decide. If we want to drown ourselves in pollutants and a warming climate, we can. If we want to annihilate each other in nuclear war, we can. If we want to become responsible stewards of our environment, we can. If we want to build global abundance, limitless energy, interstellar travel, transcendent art and a rule of just law, we can.

If a new, more intelligent species were to appear on Earth, humanity would surrender its choice over what future we want to make manifest. The future would be in the hands of the successor, and humanity would be relegated to a position no more admirable than the one chimpanzees inhabit today.

No such more intelligent species exist today, but they are being built.

Since its inception, the field of artificial intelligence has aspired to construct artificial minds as smart as, and then even smarter than, humans. If they succeed, and such systems are built, humanity will no longer be in control of the future, and the decisions will be in the hands of the machines.

-

If you don't do something, it doesn't happen.

This might seem so obvious it's barely worth bringing up. Yet, you might be surprised how often people, probably including you, don't really believe this.

If we want the future to go well, someone needs to make it so. The default state of nature is chaos, competition, and conflict, not peace. Peace is a choice we must strive for, a delicate balance on the edge of entropy that must be lovingly and continuously maintained

and strengthened. Good intentions are not enough — it demands calm, cooperative, and decisive action.

This document is a guide to what is happening with AI, and offers a playbook for nudging the future into the direction you want it to go. It is not a solution, but a guide. A book cannot be a solution, only a person's actions can.

What is AI? Who is building it? Why? And is it going to be a future we want? (Spoiler: No) There are so many things happening every single day in the field of AI, not to speak of geopolitics, that it seems impossible to keep up with, or to keep focused on what really matters: What kind of future do we want, for ourselves, and for our children?

We must steady our focus on this, and not let ourselves be distracted by all the noise and demoralizing nihilism pelting down on us from all sides. We need to understand where we want to go, chart a path there, and then walk this path.

If we don't do something, it doesn't happen.

-

The default path we are on now is one of ruthless, sociopathic corporations racing toward building the most intelligent, powerful AIs as fast as possible to compete with one another and vie for monopolization and control of both the market and geopolitics. Without intervention, humanity will be summarily outcompeted and relegated to irrelevancy by such machines, as our chimp cousins were by us.

A species of intelligent beings born from the crucible of sociopathic market and military competition will not be one of loving grace, and, for reasons we'll discuss in depth later on, will have far fewer qualms about paving over humanity's habitat with solar cells and parking lots. Despite humanity's flaws, we still have a heart, we have love, even for our chimpanzee cousins, somewhere, sometimes. Machines of ruthless competition need not have such hindrances.

And then that's...it. Story over. Humanity is no more.

There is no one coming to save us. There is no benevolent watcher, no adults in the room, no superhero that will come to save the day. This is not that kind of story. The only thing necessary for the triumph of evil is for good people to do nothing. If you do nothing, evil triumphs, and that's it.

If you want a better ending for the Human Story, you must create it. Can we forge a good, humanist future, one that is just, prosperous, and leaves humanity sailing into a beautiful twilight, wherever its final destination may lie? Yes. But if you don't do it, it doesn't happen.

The path we are on is one of going out with a whimper, not of humanist splendor. It is embarrassing to lose out on all of the future potential of humanity because of the shortsightedness and greed of a few. But it wouldn't be surprising. A human story if there ever was one.

—

It isn't decided yet whether the Human Story ends here, but it will be decided soon.

We hope you join us in writing a better ending.

- Connor Leahy, October 2024

## Overview

In **(1) The state of AI today – We do not understand the AI we are building**, we contextualize the current state of AI, highlighting the recent trends and their potential downstream effects.

*The pace of AI progress in the last decade has been extraordinary, driven by a brute-force paradigm – development does not require insight, but data, compute, and money. It has worked so well that many companies have shifted to pursuing AGI as their primary goal. Concerningly, researchers and engineers don't need to understand how modern AI systems work in order to create them, and AIs have become increasingly powerful, mysterious, and unpredictable. Given this accelerated development, we and many experts anticipate the emergence of uncontrolled AGI in the next few years, leading to catastrophic risks for humanity.*

In **(2) Intelligence – Intelligence is mechanistic and AGI can be built**, we discuss whether it is possible to recreate human-level intelligence in AGI, and conclude that it is.

*Intelligence broadly corresponds to the ability to solve intellectual tasks. Observing how humans have massively increased the range of intellectual tasks, we conclude that intelligence appears to be mechanistic. The primary arguments against intelligence being automatable thus rely on finding a mechanistic “missing component” that cannot be solved by AI. We fail to find any empirically validated missing component, and conclude that intelligence can and will be automated, leading to AGI.*

In **(3) AI Catastrophe – Current AI research leads to godlike AI**, we extrapolate what would happen if humanity created AGI, unfolding the consequences of our current approach to building AI, where AGI leads to an AI takeoff that ends in a catastrophic and permanent loss of control by humanity.

*When we develop AGI, all intellectual tasks will be automatable, including software engineering tasks such as AI and machine learning development. Given that AI companies are aggressively pursuing that end, we expect that the creation of AGI will catalyze AI self-improvement, where AI can improve the range, power, and efficiency of AI, compounding far beyond humanity's intelligence. This leads to a system which has developed such advanced powers that it's better described as a “god” from the perspective of humanity. Because this godlike AI will be beyond our control, it will control the future and almost certainly obliterate humanity, not by spite but by indifference.*

In **(4) AI Safety – We are not on track to solve the hard problems of safety**, we argue that controlling and directing godlike AI depends on solving AI “alignment,” and that we cannot do so in time.

*In order to avoid catastrophe by godlike AI, we must achieve alignment by answering deep technical, moral, and philosophical questions that are more complex than any problem humanity has faced before. These questions are not neat mathematical problems but cross-disciplinary issues that will require major research programs to resolve, at least billions of dollars of investment, and decades of sequential work. Today’s research ignores these challenges, instead focusing on hacks and tricks to correct the most egregious mistakes without really addressing the underlying problems. We dismiss the naive claim that we can wait for AGI to solve this problem. On our current path, we do not expect to solve alignment in time, resulting in annihilation by godlike AI.*

In **(5) AI Governance – We lack the mechanisms to control technology development**, we consider the institutions and mechanisms that would be necessary to steer us off the “default path” and prevent AGI from being built, and argue that these do not exist today.

*Lacking a technical safety solution, we must avert the “default path” through policy and governance means. Technical actors must be overseen by national regulators to control the safe development of AI, but these do not exist. Even if we collectively did want to slow or stop the development of AI technology, the physical (e.g compute kill-switches) and policy levers, do not exist. International stability must be maintained through high-bandwidth communication lines and multinational agreements enforced by international law, but these do not exist. And no single individual, institution, or collective has a comprehensive plan for how to handle AGI. We argue that the lack of effective efforts stems directly from AGI companies, who have captured governance and research efforts, and aggressively pushed policies of self-regulation that keep them in control without averting danger.*

In **(6) The AI Race – The race to AGI is ideological**, we explore the history and present day realities of the race to AGI, making sense of the adversarial social dynamics that currently plague the field.

*Although the race to AGI often presents itself under an economic or geopolitical mantle, its original motivation is ideological: all relevant actors care about building AGI, be it to bring about utopia, gain power, or build god. Yet the main companies racing to AGI, who want to build it to foster their view of utopia, end up in fear that someone else will beat them to the post. This fear in turn creates a dynamic where only the actors willing to compromise and undermine safety stay in the race. Thus it is not surprising to see that AGI companies are*

*using the full industry playbook, using fear, uncertainty, and doubt (FUD) to capture regulation and research, turning every argument into a justification to race even faster. Since these tactics are currently working for them, we expect them to continue in this line, and thus the race to AGI to further accelerate.*

In **(7) A good future, if you can keep it**, we argue that we must urgently work to avert the “default path” to extinction, and suggest that civic duty is what is needed today to reduce the risk.

*The primary way to avoid the default trajectory towards AI extinction risk is to not build AGI, as this is a point of no return. But to reach this level of control, we must first build global common knowledge of the risks and global communities capable of responding. These do not exist yet, and bottleneck larger solutions like the need to implement global regulations, stabilize international governance, and end the race to AGI. To get there, we must build up from small, local actions that engage in existing civic processes, and – where these fail – create new ones.*

In **(8) Outro**, we close with a brief message about the challenge ahead of us.

# (1) The state of AI today

Artificial Intelligence became mainstream after the release of [ChatGPT](#) in November 2022.



Google search popularity of "Artificial Intelligence" from 2004–present.

ChatGPT set the record for [fastest-growing user base of all time](#), exceeding 100M users in less than two months, and demonstrating to the world that AI can now generate coherent answers and solve problems that previously felt far out of reach.

Expectedly, the AI discourse has become heated and polarized. Commentary ranges from ["AGI is around the corner"](#) to ["AI is overhyped."](#) Google employees debate whether AI is ["becoming conscious"](#) or just ["parroting humans."](#) Those expecting rapid progress take sides on whether AI will spell humanity's ["doom"](#) or ["salvation."](#) Because opinions diverge so wildly, newcomers to AI often find it difficult to even start to make sense of the discourse.

We believe that three key facts cut through today's polarized conversation, and make it clear that the modern AI era is concerning:

In [Rapid AI progress is driven by resources, not insights](#), we highlight that modern AI capabilities are driven not by individual research breakthroughs, but by a simple paradigm that scales with more data, compute, and energy, and that the resources to push AI much further are being invested.

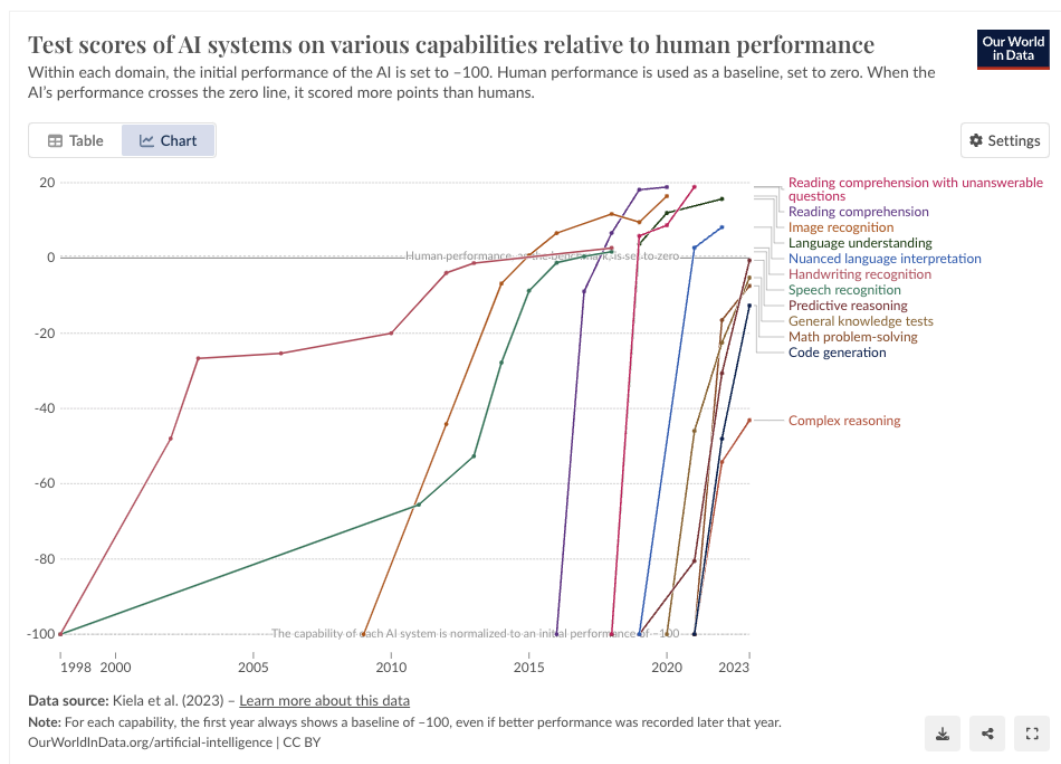
In [AI is grown, not built](#), we explain how deep learning, the core method underpinning the last decade of AI progress, leads to AI systems that researchers can neither understand nor fully control.

In [The race to AGI is on, and potentially deadly](#), we argue that the core actors developing AI today are ideologically driven to build AGI, technology that experts warn may lead to extinction.

## Rapid AI progress is driven by resources, not insights

AI capabilities have dramatically improved in the past three years, crossing crucial thresholds of competence in domains such as coding, verbal communication, abstract reasoning, image recognition and generation, vocal simulation, planning, and autonomous execution and refinement.

The trendlines comparing AI and human test scores show striking progress: AI is now better at image recognition and reading comprehension than most humans, and approaching near-human performance in domains that were thought to be out of reach, such as predictive reasoning, code generation, and math.



Graphic from [OurWorldInData](#) showing AI capabilities improvements over time

Test scores are a contentious predictor of AI's actual capabilities, but businesses are nevertheless adopting AI, suggesting that AI performance is not superficial. Five years ago, few businesses claimed to use AI; today, nearly [65% of all businesses](#) use it to perform some cognitive labor. Perhaps most notably, AI can now code well enough to significantly accelerate the work of even the best human coders — Y Combinator startups [use AI to write between 40%-90% of their code](#), a job that would typically require a three- or four-year STEM degree or equivalent self-education.

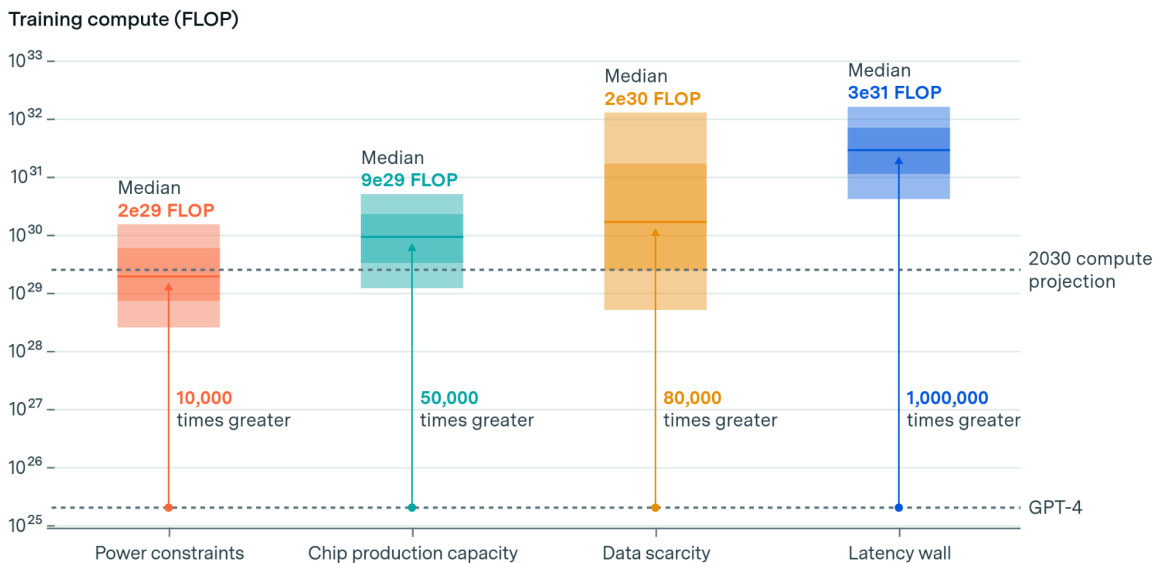
This massive jump in capabilities was driven by resources, not insights.

Historically, AI development has been punctuated by AI summers and winters – periods of rapid progress, followed by periods of stagnation. While past AI summers were driven by algorithmic breakthroughs, recent advancements have been driven by [scale](#)<sup>1</sup>: more high-quality data, more training time, and more money is all it takes to build ever more powerful systems.

In the last five years alone, the GPT models that underpin OpenAI’s ChatGPT have gone from useless to groundbreaking purely through scaling. When GPT-1 was released in 2018, it could merely write grammatically correct sentences. GPT-2 (2019) could write prose, poetry, and metaphor. GPT-3 (2020) could write stories, with characters that maintain coherence over time. Today, GPT-4 can [pass college-level exams on the first try](#), [support complex legal work](#), and [personalize language learning](#).

This has significant implications for the future of AI progress, because it means that resources are the only obstacle to greater capabilities. [Forecasts](#) show that in terms of power, chips, data, and latency, there is ample room to grow, at least until 2030.

### Constraints to scaling training runs by 2030



<sup>1</sup>This has been dubbed "the bitter lesson" by Richard Sutton, one of the fathers of modern Reinforcement Learning: the idea that scaling through general methods is ultimately much more effective to increase AI capabilities than relying on the ingenuity of the researchers.

The most likely and proximal blocker is power consumption (data-centers training modern AIs use enormous amounts of electricity, up to the equivalent of the yearly consumption of 1000 average US households) and chip production capacity (AI relies on graphic processing units (GPUs), which are specialized circuits that can perform high-speed calculations).

In response, there have been coordinated investments to support AI development. [Google](#), [Amazon](#), and [Microsoft](#) are partnering with frontier AI companies to funnel billions of dollars toward scaling, in turn providing them with the necessary compute power; new compute companies are [raising billions](#) from private equity. Microsoft and OpenAI are investing [\\$10 billion](#) into renewable energy sources to power them; others are [buying up nuclear power](#).

To avert future chip shortages and supply chain failures, the US government has committed to investing in [domestic semiconductor development](#) and [embargoed GPUs](#) from being sold to China in a trade war move to keep the US ahead on AI. Incumbent GPU manufacturer Nvidia has shifted its focus to optimizations for AI, making it [the highest market-cap company in the world by November 2024](#), at more than \$3 trillion valuation. xAI already has a 100 000 GPU cluster [online](#), and [Meta claims to be training their latest models on an even bigger one](#). OpenAI is [raising](#) billions to support continued development and fund compute requirements.

There is also a burgeoning market for training data: media companies are selling their data for [hundreds of millions of dollars](#), and custom dataset providers are [raising billions](#) to hire human labellers to effectively offload humanity's knowledge into datasets that raise the waterline for AI capabilities. Already, the massive scale and collective efforts toward pursuing more advanced models have helped the systems move down the cost curve, and training will continue to get cheaper as the technology scales.<sup>2</sup>

What this all means is that the world is investing in an approach to AI that assumes the only requirement to further AI progress is more resources (chips, power, data...). If this thesis holds true, then we should expect that AI in five years will be orders-of-magnitude more powerful than today's AI, just as GPT's capabilities improved from barely coherent to college-level performance simply due to scale.

---

<sup>2</sup> The biggest GPT-2 model (1.5B parameters) [cost an estimated \\$43,000](#) to train in 2019; today it is possible to train a 350M parameters GPT-2 [for \\$200 in 14 hours](#). Paul Graham mentions an estimate that the ratio of training price by performance [decreased 100x in each of the last two years, or 10000x in two years](#).

## AI is grown, not built

Traditional software is coded line-by-line by engineers, who need to understand broadly how the program works.

In contrast, modern AI models are developed using [deep learning](#), a technique that feeds neural networks troves of data to train them to recognize patterns and make predictions. A neural network is essentially a large graph of billions of numbers that encodes a program. Human engineers are completely unable to determine all of these numbers by hand themselves, so instead they automate the process.

The numbers are initially chosen at random. Then some piece of data is fed to the neural network, and its answer is checked against the correct one. If the answer was incorrect, [an automated program](#) twiddles with the numbers to nudge the output of the neural network closer to the correct answer. This is repeated trillions of times using data that often includes significant portions of the internet and public libraries until the neural network is very good at producing correct answers.

Although this sounds straightforward, the resulting AIs are incredibly complex, and able to do things far beyond what they were trained to do. As an important example, large language models (LLMs) such as OpenAI's GPT4 and Anthropic's Claude are trained only to predict the next few characters in text, over a huge portion of the internet and books available. Yet from that apparently simple prediction task emerged wild capabilities, such as being able to [tell jokes](#), [analyze complex legal documents](#), and [write software](#).

These capabilities are “grown” in the sense that they are not built deliberately into the model by programmers, but rather emerge from patterns in large amounts of data. This presents two significant consequences:

- **We don't need understanding to grow powerful AI.** Deep learning unlocked the current AI scaling paradigm, in which AI progress is constrained by resources rather than insights. Programmers don't need to come up with clever new solutions themselves. Instead, they can use the same deep learning algorithms over and over, simply feeding in more high quality data to larger systems.<sup>3</sup>

---

<sup>3</sup>This is not saying that ML researchers and engineers don't have specialized knowledge and skills. Only that their expertise is now about how to tweak the training methods to grow better AIs, not on understanding how the AI (or even the training process) work.

- **We don't understand powerful AI.** Today we are building AI systems which we do not fundamentally understand and cannot predict the capabilities of, a complete divergence from how all other code is written by and legible to developers.

Modern AI systems have been referred to as “[black boxes](#).” Unlike inspecting ordinary code, inspecting a neural network offers little insight into how it works, and the graphs of billions of numbers are largely inscrutable to humans. The field of [AI interpretability](#) has emerged in an attempt to understand AI models and explain their behavior, much like how neuroscience studies the brain. But despite [some encouraging progress](#), interpretability is unable to make sense even of large models from a few years ago like GPT2, and nowhere near completely understanding current LLMs such as GPT4 and Claude.

Not only are researchers and engineers unable to understand how grown AI systems work, but they are also unable to predict what they will be able to do before they are trained.

Many experts have placed bets on AI's limitations, only to be proven wrong when the next generation of model is released. Bryan Caplan, [an economist known for placing winning bets](#), wagered that an AI system would not be able to [score an “A” \(the best possible grade\) on 5 out of 6 of his exams by January 2029](#). He based his prediction on the fact that the original ChatGPT release scored a “D” in 2023; just a few months later, GPT-4 [received the 4th highest grade in the class](#).

This is just one of multiple such mispredictions:

- Francois Chollet, an AI researcher and author of a machine learning textbook, released a benchmark that many thought would be impossible for an LLM to solve, only for an LLM to [achieve a score comparable to human performance just days later](#).
- Yann LeCun, one of the three “godfathers of AI,” predicted that LLMs would not be capable of spatial reasoning. He was [proven wrong within a year](#).
- AI Safety pioneers Paul Christiano and Eliezer Yudkowsky [bet](#) that there was only an 8%-16% chance that an AI system would score gold on the International Math Olympiad by the end of 2025, but in July 2024, DeepMind's AI system [scored silver](#), challenging this forecast and shocking [prediction markets](#).
- A group of professional forecasters hired by ML expert Jacob Steinhardt [predicted in 2021](#) that a year later, the best score on the MATH dataset would be 12.7%; the actual result was 50.3%. OpenAI's recently released o1 model [scored 94.8%](#).

All of this leads to AI research where researchers assess the capabilities of AI models after they are trained, using benchmarks and hiring engineers to [red-team](#) the models. The outcomes can be concerning: a red-team testing GPT-4 found that it is [capable of hiring a human to help it solve a CAPTCHA](#), a capability that was not tested for during training. Or

during Anthropic’s training of Claude 3, the model seemed to exhibit [awareness that it was being tested by researchers](#), something that the researcher had “[never seen before from an LLM](#).”

What makes the situation harder to assess is that after an AI is released, more of its impressive capabilities get revealed by [scaffolding techniques that extend AI capabilities even further](#); this means that the initial release demonstrates only the lower bound on what the system can do. For example, ChatGPT has [learned to play complex open-world video games](#) like Red Dead Redemption 2, which was not an obvious capability even after the initial tests.

All in all, this means that the AIs being created and released are not understandable right now by anyone, not even the companies and researchers working on them. This is because they are grown rather than built. And as AIs grow in power, they will become even more complex, mysterious, and illegible.

## The race to AGI is on, and potentially deadly

The main companies driving AI progress today are racing to build Artificial General Intelligence (AGI), AI systems as smart as humans. These companies are all using the same methods discussed above: scaling up deep learning and investing billions of dollars into private data centers and large training runs.

DeepMind, founded in 2010, was the [first company to try to build general-purpose AI](#). Google acquired DeepMind in 2014, and today the [Google DeepMind](#) team builds next generation AI systems, acknowledging that “AI — and ultimately artificial general intelligence — has the potential to drive one of the greatest transformations in history.”

OpenAI, creator of ChatGPT, has been [pursuing AGI](#) since its [inception](#). OpenAI was founded in 2015 after a dispute over the future of AGI [spurred](#) Elon Musk to launch a competitor to DeepMind. Since then, Microsoft has partnered with and acquired a 49% stake in OpenAI.

Musk eventually left OpenAI, but has once again taken up pursuing AGI with xAI, which has built the powerful LLM [Grok](#). In a Twitter Space, Musk [allegedly said](#) xAI’s goal is to build “AGI with the purpose of understanding the universe.”

Anthropic, creator of the top-performing AI [Claude](#), is also publicly [racing for AGI](#). Anthropic was established in 2021 after founder Dario Amodei and OpenAI CEO Sam Altman had a [breakdown in trust](#) over the future of AGI. Today Anthropic has a \$1–4B

[partnership with AWS](#), and while AWS has not made public statements on AGI, they have an [internal AGI team](#) and recently [bought out the founders of \\$350M-funded Adept](#) to join it.

Meta, which has released the [largest open-weight<sup>4</sup> AI models](#), was previously quiet on AGI, but CEO Mark Zuckerberg announced in January 2024 that [Meta is now pursuing AGI](#) as well.

Competition to train powerful AI models has mostly consolidated among these actors, although new players are jumping in, like [Safe Superintelligence Inc.](#), a new start-up by [Ilya Sutskever](#), ex-Chief Scientist at OpenAI, which raised \$1B to explicitly train smarter-than-human AIs.

Nearly every AI application today is downstream of these major companies. [Foundation models](#) are so named for a reason: they form the foundation for other AI applications. This means that the core risks inherent to these models, such as the fact that researchers do not understand how they work, apply to every AI application that uses them.

In addition, the push to make AIs open-weight – releasing the weights of these AIs on the internet – by tech giants like [Meta](#) and newcomers like [Mistral AI](#) creates a dangerous ecosystem where everybody (criminals, terrorists, rogue states...) has unmonitored access<sup>5</sup> to these powerful systems for whatever purpose (scams, hacks, propaganda...). This is particularly problematic because of the asymmetry of the situation: governments and law-enforcement need to defend everywhere against the assaults of nefarious actors, whereas it takes only one successful malicious actor to create a mess.

As these models become ever more powerful, and orders of magnitude more complex with scale, older generations can be replaced by new and improved systems, meaning that the entire AI industry (and the open-weight ecosystem) becomes increasingly powerful and opaque with each release.

We believe that this is a recipe for catastrophe, and we're not the only ones. After ChatGPT was released in 2023, global conversation on the risks from AI – including extinction – entered the mainstream.

In May 2023, the Center for AI Safety released an open "[Statement on AI Risk](#),"

---

<sup>4</sup>We use "open-weight" rather than "open-source" because the AIs released by Meta and Mistral AI are lacking many essential parts of an open-source release. Notably, these companies don't release the data these AIs were trained on, nor do they share the training algorithms used to grow them. As we will see, this nullifies the potential benefits of open-source software.

<sup>5</sup>Although a lot of these AIs require high-end chips that are expensive and regulated, the online community is investing massive efforts to bypass these constraints, allowing people [to run almost state-of-the-art AIs on a single macbook](#).

*Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war.*

The letter was signed by hundreds of signatories, including the CEOs of OpenAI, DeepMind, and Anthropic, two of the three “godfathers of AI,” weapons experts, politicians, and academics.

A few months before that, a similar [open letter](#) was published by the Future of Life Institute, calling for a pause on “Giant AI Experiments,” a 6-month ban on training AI models more powerful than GPT-4. It has garnered over 33,000 signatures, including Elon Musk’s.

In June 2023, a [survey](#) conducted at the Yale CEO Summit found that 42% of CEOs said that AI has the potential to destroy humanity in 5–10 years.

In November 2023, at the first global summit on AI, a consortium of world leaders from the USA, China, EU, and other global powers signed the [Bletchley Declaration](#), which acknowledges the “potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of these AI models.” Following the summit, the UK, US, Japan, France, Germany, Italy, Singapore, South Korea, Australia, Canada, and the European Union have agreed to establish [AI Safety Institutes](#) with a mandate to better understand and mitigate risks from frontier AI.

Why are AI experts, CEOs, and world leaders concerned that AI is an existential threat?

**In short, extrapolating the trends of AI progress into the near future paints a grim picture: humanity has unlocked the ability to build powerful AI systems without fundamentally understanding them, and we are poised to enter a future with more powerful and dangerous technology that is less and less under our control.**

If this technology is able to meet, and eventually exceed, the capabilities of a human, then we may face a world where we are powerless to control the very AI we have created.

This document presents a worldview that leads us to this claim and unfolds the consequences: humanity may end up extinct unless we intervene today.

## (2) Intelligence

### Intelligence is mechanistic and it is possible to build AGI.

Evaluating the existential risk of AI centers on the feasibility of building Artificial General Intelligence (AGI), an AI that rivals human capabilities. The [debate about](#) whether AI can be considered truly “intelligent” first requires establishing a definition of intelligence.

We contend that intelligence consists of intellectual tasks, which are self-contained problems such as planning, summarizing, generating ideas, or calculation, that must be solved to achieve a goal. We solve these tasks through intellectual work, which is similar to the notion of [work in physics](#). For example, solving a difficult research problem takes more intellectual work than solving a simple addition problem – it requires more hours of thought, more calculations, and more experimentation.

By considering intelligence as the intellectual tasks it is composed of we arrive at a mechanistic model, where the question of whether we can create AGI is not “can we recreate human intelligence?” but rather “can we automate all intellectual tasks?” To which we say: yes.

- In [What is intelligence?](#), we argue that intelligence is the ability to solve intellectual tasks. Humanity has expanded the range of feasible tasks by mechanistic means, specifically through the usage of tools, groups and specialization, and improved methods of thought.
- In [Applications to artificial intelligence](#), we demonstrate that AI capabilities can similarly expand, and that an AI capable of automating all human-manageable intellectual tasks is functionally an AGI.
- In [Against arguments of AI limitations](#), we argue against the notion that there is a missing component of intelligence that will prevent AI from rivaling human intelligence. We show how the “AI Effect” has led to shifting the goalposts of “real artificial intelligence,” leading to pseudoscientific discourse.
- In [Thus, AGI](#), we conclude that even if we account for uncertainty, the rapid expansion of AI capabilities and the absence of known limitations can lead to AGI, and that we must proceed with caution.

## What is intelligence?

Intelligence is not yet a [hard science](#) that we can study mathematically and base predictions on. As a result, most [definitions](#) of intelligence are imprecise, but converge on some combination of processing information and responding accordingly in pursuit of a goal. Intelligence seems to be what makes one human “smarter” than another and differentiates humans from the rest of the animal kingdom. A complex piece of technology demonstrates intelligence, and we understand it to be man-made: nothing else we know of can similarly rearrange nature. Whatever “intelligence” is, it has enabled humans to dominate Earth.

This ability has evolved over millennia; humans have gone from struggling to survive by hiding in caves and hunting beasts, to going to space, curing a vast range of diseases, and producing enough food to sustain 8 billion people (if it were distributed equally). Human intelligence grew as a result of three primary factors:

- **Tools:** We have created tools that capture the ability to solve specific tasks, and make it possible to solve them with much less knowledge, understanding, and competence.
- **Groups:** We split complex intellectual tasks (such as designing an airplane) into many smaller ones, which are further distributed across multiple people, allowing for parallelization and specialization.
- **Methods:** We have improved our methods to solve intellectual tasks, such as developing the scientific method and other means to account for biases and systematic sources of errors.

By looking at the three factors that have most increased human intelligence, we can bypass abstract theorizations and directly examine what has made us more capable: simple mechanistic processes that could be replicated with the right software.

### Tools

Our inventions and tools have played a critical role in our intellectual and societal growth. Physical tools, such as writing implements, enabled preserving knowledge across generations, leading to complex societies. The printing press democratized knowledge by making books more accessible, expanding literacy. Tools let humans offload intellectual tasks — before modern computers and calculators, [“computers” were humans](#) who performed calculations by hand. Today, the intellectual work is distributed between the human and the calculator.

Critically, solving a task is a composition of distributed, physical processes: the brain apprehends the question and the need to solve it, triggering a nerve impulse that travels to

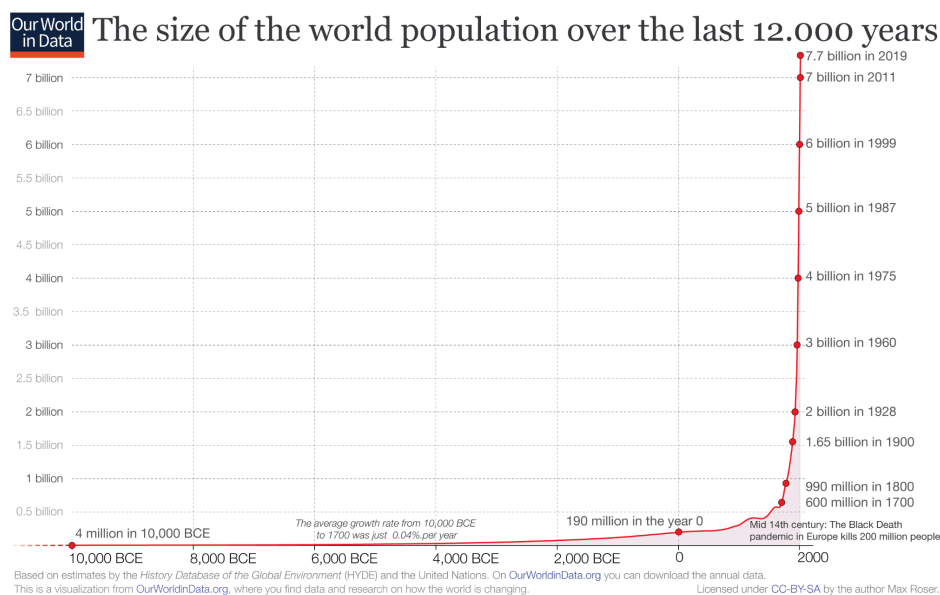
our fingers and leads us to punch keys on a calculator, which in turn moves internal digits to compute an answer, display it on the screen, return photons to our eyes, and carry nerve impulses to the brain to signal that we have answered the question. When we think about intelligence this way we can see that it is not something that just happens in the brain.

The processes we use to solve intellectual tasks have co-evolved with our tools. Before smartphones and GPS systems, navigation required a physical map; today, increased reliance on navigation systems means that fewer people own maps, or even know how to use one. As technology encodes the intellectual processes that were once essential for survival, we no longer use or develop the requisite skills ourselves.

Without our tools, we would not just be uncomfortable, we would be handicapped. Without language, literacy, or numeracy, it's not just that we couldn't speak, remember, or calculate; it's that we couldn't *think*. Without written language, we could not preserve information except by memorization. Without language at all, we would have no internal monologue. What would our "intelligence" be like then?

## Groups

Population scale and specialization have helped expand humanity's intelligence. If more people work on a problem, there is more intellectual work being directed toward solving it. Today, we have nearly 8 billion humans *thinking* about stuff.



*There are a whole lot more thinking humans today than ever before.*

If intelligence is a matter of solving intellectual tasks, then having more people to distribute them across helps it grow. Consider the task of finding a cure for cancer – while an individual researcher may discover a breakthrough drug, this achievement will have depended on intellectual work performed by many different humans: collaborating with researchers, reading the papers of predecessors, running experiments with interns, and so on.

Just as intellectual work is distributed across individuals and tools, it is also shared across groups like school departments, companies, internet subcultures, markets, governments, and international bodies. These groups can be thought of as intelligent entities with coordinated goals and processes for sensing, thinking, communicating, and surviving. For example, Airbus has the goal of building airplanes, senses and responds to events like changes in market conditions, thinks and communicates through its employees and the tools they use, and aims to survive and grow by building airplanes and outcompeting the market. The law even treats groups as single entities: states and companies are considered agentic enough to have liabilities, form contracts, and own property and debt.

Companies like Airbus are much smarter than any individual because they can solve more complex intellectual tasks than any single human could. States are more intelligent than companies (although Big Tech is challenging this now) because they can be thought of as performing the economic work of all of the companies within their jurisdiction. In general, groups are more intelligent than individuals because they can solve tasks that are too complex for an individual to solve alone.

## Methods

Intelligence has also grown within an individual. This was not the result of evolution – a few thousand years is a relatively short time frame in evolutionary terms, and the human brain has remained largely unchanged since the advent of Homo sapiens around 200,000 years ago.

It is our thinking – the software and methods of intelligence – that has evolved.

While we consider humans to be the most intelligent species on Earth, we also know that we are fallible. Our memory is terrible and our learning is slow. Most of us are profoundly irrational, believing contradictory things and failing to use logic, and we can be incoherent and operate contrary to our goals. We are not always agentic – we fail to consistently plan and follow our own goals and end up stuck. All-in-all, we have many shortcomings.

Over the last several thousand years, we have discovered techniques to circumvent some of these shortcomings and grow our individual intelligence.<sup>6</sup> Logic, mathematics, and science structure our thinking into correspondence with reality; psychology and therapy aim to correct psychological mistakes that lead to individual failures. Formal education strives to teach people how to learn and correct their own mistakes. Improving our methods in turn helps us solve intellectual tasks more effectively.

## A mechanistic model of intelligence

We have shown that humanity's intelligence has grown as a result of tools, scale, and improvements to thinking methods. These are concrete, mechanistic factors, suggesting that intelligence is not a mysterious force in the brain, but rather a systematic, distributed process across physical entities.

The world is composed of processes that solve intellectual tasks. Solving a complex calculation is a coordinated process between a human and a calculator. At a larger scale, shipping fulfillment is a coordinated process between a customer, their computer, an algorithm that handles the purchase, a physical warehouse with the item, employees who process the order, shipping companies, and so on. The same is true for much more complex tasks, such as building a rocket, or state functions such as diplomacy, policy implementation, or adjusting the treasury rate. An entity's intelligence is a measure of the intellectual tasks it can perform.

It follows that to create an AI that rivals human intelligence, it simply needs to be capable of performing the same intellectual tasks. If all intellectual tasks are automatable, then intelligence itself is automatable.

## Applications to artificial intelligence

If intelligence is the mechanistic completion of intellectual tasks, then what has made humans intelligent is trivial to recreate in artificial intelligence — tool use, scaling and collective intelligence, and improving methods are all available to AI.

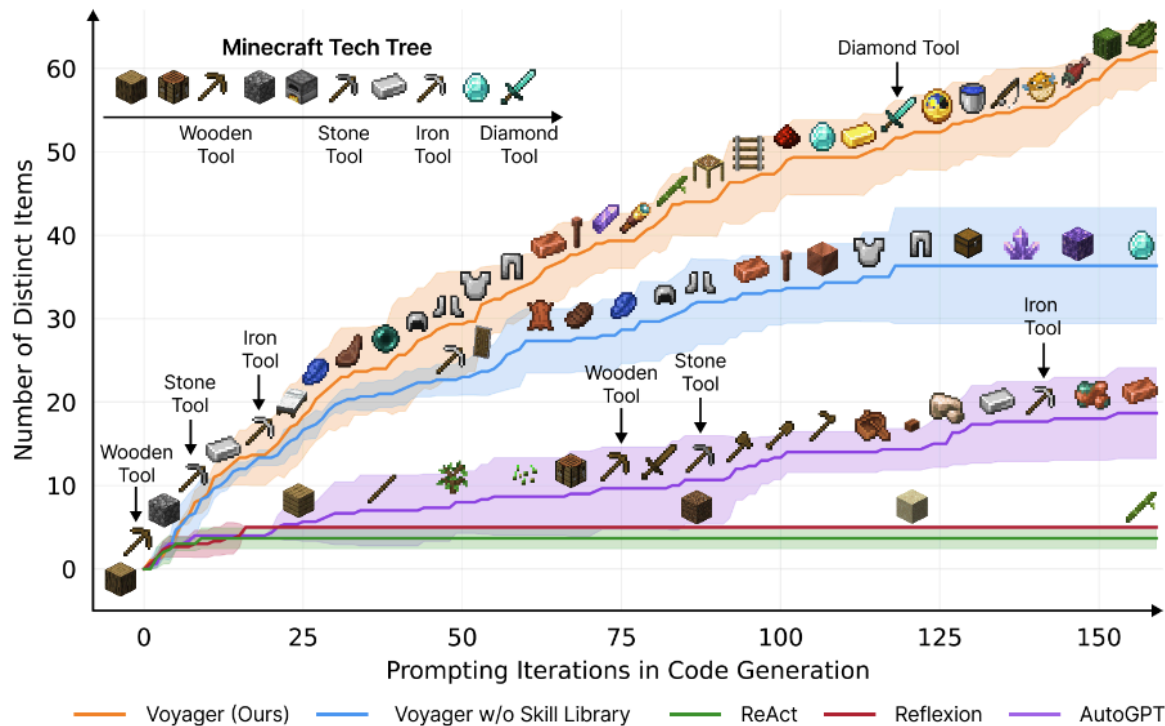
**Tools:** Today's AI is able to use online tools. Because large AI models have learned how to use natural language, they have also learned how to code, which allows them to use APIs to use tools and other interfaces. Shortly after the release of ChatGPT, OpenAI developed [plugins](#) to support using APIs, and the technology has only improved since. For example,

---

<sup>6</sup> One of the massive benefits of these methods is that they can be applied without understanding exactly how to rederive them. Most people applying various scientific methods would struggle to rederive them, but they don't need to.

open-source project [ToolBench](#) has collected over 16,000 APIs that AI models can interact with.

Tool use transforms an LLM into a meaningful agent that can complete tasks. For example, [Voyager](#) is an open-source project that shows that GPT-4 can inherently play Minecraft — researchers connected the LLM to a text representation of the game and showed that it is capable of navigating environments and solving complex tasks by writing small programs for skills to “mine,” “craft,” and “move.” Just as humans solve intellectual tasks in Minecraft through processes that trace through our brain, the computer, the game environment, and back, so too can Voyager solve problems through processes that coordinate the LLM, the computer, and the game environment API.



Voyager discovers new Minecraft items and skills continually by self-driven exploration, significantly outperforming the baselines.

The range of physical and virtual environments an AI can navigate is growing. [FigureAI](#) and other robotics companies are training advanced AI models to perform domestic tasks and navigate warehouses. [AdeptAI](#) and other research companies are training advanced AI models to navigate computer interfaces. Once these problems are solved, all human tools and environments will become available to AI. Considering the impact of tool use on improving human intelligence, this will be a significant lever to increasing the overall intelligence of AI.

**Specialization and scale:** Compared to human population growth, it is trivial to scale AI — we can make multiple copies by simply opening new browser windows. Researchers are working on agent frameworks to allow AIs to communicate with each other. Should we find more robust frameworks, AI will benefit from the advantages of group intelligence and distributing intellectual labor. We can already see this in action: Deepmind’s [AlphaProof and AlphaGeometry](#) systems achieved an IMO silver medal by generating thousands of “candidate” answers and assessing them in parallel, akin to a large group of math students dividing up proofs and manually reviewing them. Parallelization also leads to efficiency gains: distributing the work leads to faster solutions than sequential processing.

Scale can enable the growth of AI’s capabilities in the same way that it has for humanity. If we build an AGI that is as smart as a single human, we could theoretically scale it to rival humanity’s collective intelligence by creating and coordinating billions of copies. This is a big deal, as it suggests that there is no “secret” to scaling from AGI to superintelligent entities.

**Methods:** AIs are programs. In the same way that we have expanded individual intelligence by improving our own methods and programming, we can improve the programs that AIs run. If we can create an AI that can perform each intellectual task that a human can, then we can compose those tasks into processes that mimic the intelligence of a human, a company, or even humanity.

Today, AI companies are chipping away at these tasks, solving meaningful problems in language use and multimodality, search, planning, and inference. Many of these are solvable with traditional code that does not require AI, such as the applications we use on our phones. AI can also solve these tasks, and each successive model grows in its capabilities. If a task is unsolved by a base model, engineers often build extensions through finetuning, scaffolding, and other means.

Whenever we find a programmatic way to solve a task using AI, that process can be distributed to other AIs far faster than we can distribute new skills across humans — every application that used GPT-3 as a base model made an immediate stepwise improvement when upgraded to GPT-4. While it may take humans hundreds of years to improve our collective rationality, the discovery of a post-processing method to reduce AI hallucinations could be rolled out in days.<sup>7</sup>

---

<sup>7</sup> Consider that updating and deploying anything made of atoms (physical products and infrastructure) is much more costly and slow than updating and deploying anything made of bits (software).

Today's AI research is aimed at automating all of the intellectual tasks that humans currently perform. The system that can solve all of those tasks will be an AGI.

## Against arguments of AI limitations

We have shown that expanding intelligence is a mechanistic process, and that it is possible to build AGI using the same techniques that have augmented humanity's collective intelligence.

And yet, many people do not intuitively have the feeling that AI is "intelligent," or anything other than a tool. Though AIs are [increasingly autonomous](#), able to [navigate environments](#), "think" [chains of thought](#), and master language which enables "general" performance across many domains, public perception writes these off as mere components of intelligence, suggesting that AI is missing something that would make it the "real thing." Many reject the existential risks of AI because of the intuition that AI is missing a fundamental component of intelligence.

This section explores whether there is in fact a missing component of intelligence that we cannot automate, and that will therefore impede AGI. We argue against the existence of a missing component, showing that the historical tendency has been to shift the goalposts around the definition of intelligence every time AI is able to perform a novel task. We also consider broader theories of intelligence, disproving theoretical justifications for "missing components" arguments.

### The AI Effect

Before establishing a solid theory about how something works, the mechanism can seem "magic," and beyond the confines of science and reason. In AI, this phenomenon is common enough to be termed as the "[AI Effect](#)" — once AI achieves a new capability, it is dismissed as an aspect of intelligence.

For example, in the early days of computer science and AI, memory was an unsolved problem. It was not clear how to best store and retrieve information, which plays a role in general intelligence. We have since developed databases, filesystems, querying languages, and other tools that "solve" most memory problems, but we wouldn't consider an SQL database to be intelligent, despite its extremely sophisticated abilities to store and retrieve complex memories.

This is a recurring pattern. Chess was once considered the pinnacle of human intelligence, but AI has been able to significantly outperform the best humans since 1997.<sup>8</sup> Logic systems have solved mechanical reasoning. Classifiers are now able to infer patterns from data as well as humans can, and we now class linear regression and other statistical methods as “data science,” not AI. AI can increasingly manage tasks that we consider demonstrative of intelligence, but we do not yet consider AI itself to be intelligent.

It is important to track these improvements as progress towards general intelligence. Computer memory, chess bots, and good classifiers do not “solve” intelligence, but they each get us one step closer by automating a vital intellectual task. Now that AI can solve [PhD-level problems](#), the number of tasks AI cannot perform is few and getting smaller by the day, suggesting an overall trend toward more and more automation of intellectual tasks.

Because of this trend, one should be wary of arguments that claim general intelligence is a discrete milestone that some “missing component” prevents us from reaching.

While deep learning seems more impervious to the AI Effect because of its inscrutability, this illegibility feeds into myths around the missing component of intelligence, leading some to believe that because we don’t understand deep learning, we must be missing something profound about the nature of intelligence. For example, even scientific experts have suggested that [planning](#), [consciousness](#), [reasoning](#), and [inferring new patterns](#) are all missing components that we must solve before we have AGI.<sup>9</sup>

Let’s dig into one of these examples to demonstrate the shape of the “missing component” argument and how it falls apart on inspection.

One proposed refutation of AI’s intelligence is that [AI is not capable of planning](#). At first glance, this might seem reasonable: we do not yet have long-lived robots that make plans in the real world and interact with people, and when asked, ChatGPT’s plans are mediocre. Yann LeCun argues that LLMs are inherently reactive systems that [cannot plan or engage in long-term strategic thinking](#), and therefore are missing one of the “[four key ingredients for intelligence](#).”

---

<sup>8</sup> When DeepBlue beat reigning world chess champion Garry Kasparov in a series of six games.

<sup>9</sup> The most common argument against AGI is that AI is not yet conscious. Fortunately we don’t need to enter the quagmire which is the field of consciousness studies, because the one thing everyone seems to agree on is that we do not have a good theory of consciousness. If it were the missing component, our theories still wouldn’t be able to tell us anything useful about AI. And for our purposes, an AI capable of automating all intellectual tasks would not need to be conscious to be functional and dangerous

But let's consider what "planning" actually is: the ability to devise a sequence of actions in order to reach a goal. We understand the pragmatic vision of planning quite well. The entire field of [traditional AI](#) is about planning, from general algorithms like [backward chaining](#) to more specific subfields like [scheduling](#). Today, one can ask ChatGPT to plan for any situation; it often fails because it lacks topical knowledge and not because of an inability to plan, but this is also true for people. Current engineering efforts are focused on [agent architectures](#) that are fundamentally predicated on planning, such as powerful game-playing AIs like [Voyager](#) and [Cradle](#), or [Devin](#), which aim to replace software engineers. AI's search capability is also [very well studied](#), and demonstrated by AI's ability to solve mazes, look many moves ahead in chess, and consider battlefield tactics. For more technical readers, there is some early evidence of neural networks [learning to search at an intuitive level](#), in a single forward pass.

Why are planning and search sometimes held up as the next [big breakthrough](#) in AI, despite the fact that AI is already capable of both?

In short, this is a common fallacy where a distinction is made between a "true capability" and mere imitations/parroting. Such arguments fail to see that any capability such as planning is simply an intellectual task, made of smaller intellectual tasks. There is no massive breakthrough to reach, only more and more tasks to automate.

**To put this another way: the "missing component" argument is not actually pointing to a real thing at all.** Yes, the argument does identify certain limitations of what today's AI are capable of doing. But when broken down to investigate *why* or *what exactly* AI cannot do, the question dissolves. All we find is that we are making incremental traction on planning, and that we've automated many components of it and have some more to go. What we certainly do not find is a mystical threshold we would need to cross over to get to "true planning."

This type of mistake appears again and again in debates about artificial intelligence, with many experts making pseudoscientific claims about key skills that AI lacks. But instead of trying to respond to each of these particular claims, let's explore the general reason these arguments fail.

### The general issue with missing components

"Missing component" arguments are brittle because they can't be scientifically substantiated.

To make a compelling argument about a missing component of intelligence, one would need to both prove through an empirically supported theory of intelligence that the component is indeed a necessary element of intelligence, and then show that it is impossible to automate.

Neither of these is possible today because humanity's scientific understanding of intelligence is so limited and piecemeal that we have no agreed upon theory.

Traditional psychological theories of intelligence fall into one of two camps: they are either partially mechanistic but wholly unsupported empirically, or empirically grounded but not theoretical and conceptual enough to point to a missing component. The [Theory of Multiple Intelligences](#) is an example of the former: it proposes different forms of intelligence that could count as potential missing components, [but lacks empirical confirmation](#). The [Parieto-frontal integration theory](#) is an example of the latter: it is considered the best empirical claim we have on the locus of intelligence in the brain, but it lacks any explanation of cognitive processes that could help identify a missing component. Even recent computational models of intelligence, such as [active inference](#) and [computational learning theory](#), are not sufficiently developed to identify the necessary conditions for intelligence.

AI complicates this further. Even if we knew all the necessary components of intelligence, we lack the understanding of modern AI necessary to predict if AI could automate those components.

Simply put, we do not yet have any scientific theories of intelligence to make a “missing components” argument, nor evidence that AI fundamentally cannot do something. Where this leaves us is that the “missing components” argument is nothing more than a vibe, an intuition against the possibility of automation that grasps to use any example of modern AI struggling.

**But this is a vibe that contradicts all the historical evidence we have about AI.** The AI Effect shows us that we are constantly breaking walls that we thought were impossible for AI to solve, at unexpected times. Moreover, AI capabilities already increase rapidly without us understanding them, , with new models constantly revealing previously unexpected capabilities.

We're a frog in boiling water. A mechanistic view of intelligence-as-intellectual-tasks demonstrates that we are getting closer to AGI by the day, as we automate more and more intellectual tasks. Arguments of a “missing component” seek to undermine this view, holding “real intelligence” as some never-quite reachable standard. The fact these

arguments are scientifically unsubstantiated is a big deal, because it means that we have no scientific evidence to disbelieve the trend towards AGI.

## Thus, AGI

If intelligence is a matter of solving modular tasks through the sophisticated usage of tools, specialization, and refining methods, and there is no viable theory of a missing component, we are led to adopt a physicalist interpretation of intelligence. In other words, intelligence is built on matter and mechanistic, there is no spiritual barrier to building AGI, and the path to AGI is to build systems that can perform each intellectual task that a human can perform.

Every advancement in AI is chipping away at the number of tasks humans can perform that AI cannot perform. Building intelligence component by component, AGI is an engineering problem like any other which we are closer to solving by the day. Because we have already built systems close to human-level performance, it is simply a matter of time before we build an AGI that is able to perform all of the tasks that a human can perform. At this point, adding tools and cloning the AGI to billions of copies could create an intelligence that quickly rivals that of humanity.

We therefore find ourselves in a precarious position:

- AI capabilities are rapidly increasing and will continue to, regardless of our understanding of the nature of intelligence.
- Major AI companies are racing to build AGI, chipping away at the tasks that comprise intelligence.
- Once built, AGI can quickly exceed the intelligence of a single human.
- And yet, modern AI is unpredictable and uninterpretable, and we lack any theory of intelligence that allows us to make confident claims about when AGI will be built or what it will be capable of doing.

We believe that the burden of proof is on anyone who believes that we will not be able to create AGI. Given the evidence of AI's progress, the effort directed at building AGI, and the risks that come from building such world-changing technology, it is extremely meaningful to substantiate arguments for why we shouldn't be worried about building powerful technology, or the related risks. But we do not have credible evidence of a "missing component" that will prevent it, and related arguments appear unscientific.

Given this situation, we argue for a cautious approach: one that looks at the trendlines of massively increasing progress and prepares to deal with the advent of AGI seriously.

## (3) AI Catastrophe

### Current AI research leads to extinction by godlike AI

A mechanistic understanding of intelligence implies that creating AGI depends not on equipping it with some ineffable missing component of intelligence, but simply on enabling it to perform the intellectual tasks that humans can.

Once an AI can do that, the research path that AI companies and researchers are currently pursuing leads to godlike AI — systems that are so beyond the reach of humanity that they are better described as gods and pose the risk of human extinction.

More granularly, we expect AI development to progress as follows:

- [Current AI research will lead to AGI](#), intelligence that operates at the human level but benefits from the advantages of hardware.
- [AGI will lead to artificial superintelligence \(ASI\)](#), intelligence that exceeds humanity's current capabilities.
- [ASI will lead to godlike AI](#), intelligence that is so powerful that it outclasses humanity in ways impossible to compete with.
- [Godlike AI will lead to extinction](#).

To avoid catastrophe requires first being clear on the risks, which depends crucially on understanding the feasibility of creating godlike AI and its potential capabilities and dangers, as well as why the current research path will lead to it.

### Without intervention, current AI research leads to AGI

We define AGI as an AI system that can perform any intellectual task that a human can do on a computer.

Current AI research leads to AGI because any such intellectual task can be automated. Multimodal LLMs can already manage concrete physical tasks such as typing, moving a cursor, and making sense of what is displayed. More complex tasks such as navigating the web, like [Anthropic's latest Claude model can](#), are in fact only a series of smaller intellectual tasks arranged to achieve some goal. Once an AI can fully interface with a computer and perform the intellectual processes humans can, it is AGI.

Building AGI is not a matter of “if,” but “when”; we merely need to determine how to automate the intellectual tasks that humans currently perform, which we showed in the

previous section to be a mechanistic process. AGI will be powerful, but its constituent parts will be mundane:

- Some procedural tasks such as spell-checking, calculating a budget, and mapping a path between two locations, are simple to program, and already feasible using traditional software.
- More complex and informally specified tasks, like image recognition, creative writing, and game play, require intuitive judgment. We have failed to formalize these intuitions with traditional software, but they are already addressed by [deep learning](#).
- Tasks such as driving, designing a plane, and running a company remain out of reach for today's AI, but they are decomposable into subtasks that are either already solved or on track for automation.

Without a doubt, current research is actively striving to achieve AGI. OpenAI CEO Sam Altman wrote in a recent [blog post](#) that

*“AGI has the potential to give everyone incredible new capabilities; we can imagine a world where all of us have access to help with almost any cognitive task, providing a great force multiplier for human ingenuity and creativity.”*

Similarly, Anthropic's [“Core Views on AI Safety”](#) states that

*“most or all knowledge work may be automatable in the not-too-distant future – this will have profound implications for society, and will also likely change the rate of progress of other technologies as well.”*

It is misleading to suggest that AGI will empower humans – automating the most complex intellectual tasks (such as designing a plane or running a company) requires increasingly autonomous AIs that are capable of dealing with a wide range of problems, adapting in real time, and finding and evaluating alternatives to complex intermediary questions. As this unfolds, AI systems will look less like chatbots and more like independent agents solving complex problems by themselves.

Recent releases already confirm this trend, with AI agents running autonomously for longer without human feedback. For example, OpenAI's [o1 model](#) is designed to “spend more time thinking through problems before they respond, much like a person would.” Systems can also more easily self-replicate, as evidenced by Anthropic's [Claude calling other instances of itself](#), described as the ability to “orchestrate multiple fast Claude subagents for granular tasks”. As this path continues, AI research will eventually reach AGI.

Already, AGI could very well spell catastrophe for humanity. In the current climate, companies such as [Meta](#) and [Mistral AI](#) would soon release open-weight versions of AGI,

and open-source projects such as [llama.cpp](#) would soon make them available on basic computers that anyone can get. This means that every single criminal in the world would have access to a system that can think better than a human on almost all topics, and help them plan and execute their crimes without getting caught. Although law-enforcement and governments would also have access to this technology, the situation is heavily asymmetric: it takes only one successful malevolent actor to create a massive problem. This would force governments and law-enforcements race as fast as possible to get more powerful AIs and defer to it, pushing us further on the path of more autonomous and powerful AIs.

## Without intervention, AGI leads to artificial superintelligence (ASI)

AGI will lead to ASI because it is capable of self-improvement.

AI is software that is programmed by humans using computers. Because AGI can perform any intellectual task, it follows that AGI can perform the same research and programming that humans can in order to extend the capabilities of AI, and in turn apply its findings to itself.

For example, AIs can improve themselves by:

- [Generating better instructions for themselves](#)
- [Creating data to train themselves](#)
- [Building memories of previously successful actions and learning from them](#)
- [Simulating many scenarios to learn by example without any human input](#)

Current AIs already perform these basic examples of self-improvement, and researchers explicitly aim for stronger methods, from [training current AIs to program](#), to [generating data to build more powerful AIs](#), to [iterative alignment](#).

AGI could go further, and perform all of the tasks that human researchers do to improve AI, like:

- Perform AI research, create, run, and evaluate experiments, and autonomously study AI without human supervision.
- Optimize AI, making code more performant, building more powerful hardware, and finding more efficient algorithms.

- Actively “online learn”, running as many instances of itself as possible to learn from interacting with the real world through APIs, GUIs, chat platforms, ads data, and so on.<sup>10</sup>
- Interact with the real world, such as via video data, emailing or chatting with people, using currencies and paying people to do things, robotic telepresence, and so on.

These are techniques that have all been used today, but AGI could go further still, discovering new techniques for self-improvement and performing tasks that human researchers have not.

The current rate of AI improvement attributable to algorithmic progress implies huge room for software improvements, which means that just around the time when AI reaches human-level at improving AI algorithms, things will speed up significantly, and will only move faster after this.

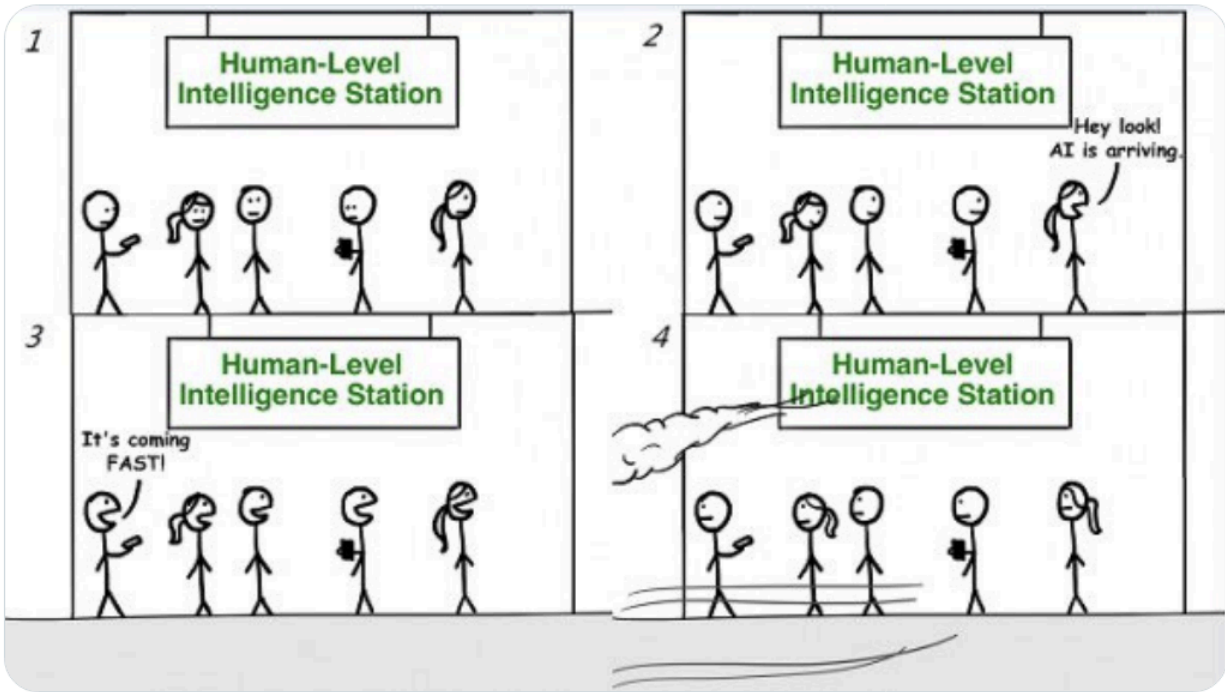
It is a mistake to assume that this automation will continue to produce mere tools for human AI researchers to use. Current research is solely focused on enhancing AI and making it more autonomous, not on improving human research understanding. This path does not lead to human AI researchers performing more and more sophisticated work, but to replacing them with autonomous and agentic AI systems which can more efficiently, effectively, and cheaply perform the same tasks.

And as we’ve already seen, this research might even be forced upon governments and nations, as the unregulated spread of open-weight AGI would create an actual arms race between criminals and governments to develop more and more powerful AIs to handle the other’s AIs.

Over time, this self-improvement will lead to artificial superintelligence (ASI), which is [defined](#) as “intelligence far surpassing that of the brightest and most gifted human minds.” We contend that superintelligence will surpass the intelligence of all of *humanity*, not just its star students.

---

<sup>10</sup> Big Tech already uses this strategy to accumulate troves of data for training purposes – an AGI could do the same without requiring human oversight.



ASI will exceed individual human intelligence

AGI will be superhuman once it is able to automate anything that a human can do because of the intrinsic advantages of digital hardware.

- **Computers are cheaper and faster to create than humans.** Child rearing alone takes 18 years and [~\\$300,000](#) in middle-class America. Replicating a system is decidedly faster, and the most powerful GPUs that host AI models [reportedly cost](#) approximately \$3000.
- **Computers are more reliable than humans.** Humans are fallible – we can be tired, distracted, and unmotivated and require daily nourishment, exercise, and sleep. Computers are tireless, perform exact calculations, and require little maintenance.
- **Computers are much faster at sequential operations than humans.** For example, a [graphics card from 2021](#) is ~34 TFLOPS, meaning it can perform 34 *trillion* 10–digit multiplications per second. It would take humanity at least five days to perform this calculation even under the generous assumption that all 8 billion humans each solve ~4000 10–digit multiplication problems without interruption or rest.

While multiplication is a simple task, this demonstrates that once *any* task is

automated and optimized, computers will perform it much more efficiently. This holds true for reading, which is how humans learn at scale. The first step in reading and understanding any text is to simply process the words on the page; at a reading speed of 250 words per minute, a human could finish the 47,000-word *The Great Gatsby* in about three hours. An [AI from 2023](#) was able to both read it and respond to a question about the story in 22 seconds, about 500 times faster.

- **Computers communicate information much faster than humans do.** The world's fastest talkers can speak [600 words per minute](#), and the world's fastest writers can type 350 words per minute. In comparison, computers can easily download and upload at 1 gigabit per second. Conservatively assuming that an English word is eight characters, this converts to a download/upload speed of 900 million words per minute — one million times faster than human communication.
- **Computers collect and recall information much faster than humans do.** Memorizing a list of 20 words might take a human ~15 minutes using very efficient [spaced repetition](#). A [\\$90 solid-state drive](#) has a read-write speed of 500MB/s. This is about 3.5 billion words per minute, literally billions times faster than any human.

An AGI able to automate any human intellectual task would therefore be strictly superhuman given performance speed. A [conservative estimate](#) posits that near-future AI will think 5x faster than humans can, learn 2500x faster, and, if cloned to many copies, perform 1.8 million human years of work in two months; some imagine powerful AI looking like a "[country of geniuses in a datacenter](#)."

ASI will exceed humanity's intelligence

Not only is digital hardware more efficient than a human brain, but digital software also offers enables AI to do things humans cannot:

- **AI's are easy to clone and scale.** Language model weights (like [Llama3 70B](#)) can weigh in the hundreds of GBs; copying a computer program by uploading it to a different machine takes just a few minutes. We have no comparable ability in humans.
- **AI's are easy to modify.** Suppose researchers knew how an AI worked, and uncovered a mistake in a model. They could fix this simply by rewriting the relevant

part of the source code.<sup>11</sup> Humans cannot comparably easily reprogram their brains.<sup>12</sup>

Just as humanity is more powerful than a single human, millions or billions of AI is more powerful than a single AI. If AGI reaches human-level intelligence and can clone itself, an ever-growing swarm of AIs can quickly rival humanity's collective intelligence.

Moreover, this swarm can learn faster than the glacial pace at which humanity learns. [Plank's principle](#) is that “scientific change does not occur because individual scientists change their mind, but rather that successive generations of scientists have different views.” While it may take humanity generations to change our minds, AI capable of reprogramming itself can fix its intellectual flaws as fast as it notices them.

Consider that the frontier of human intelligence – our scientific communities – are riddled with methodological and coordination failures. As ASI improves, it could transcend these failure modes and make rapid scientific progress beyond what humanity has achieved.

- **AIs can fix issues in scientific methods much faster than humans.** Human fallibility impedes progress, as evidenced by the [replication crisis](#) in many areas of science, which has led to [publication bias](#) and [HARKing](#). Potential solutions such as [as preregistration](#) have been proposed, but in practice scientists do not report failures and present successes as intentional, making it difficult to evaluate scientific literature. These methodological errors may never appear for AI, and if issues arise, they could simply be removed with reprogramming.
- **AIs can coordinate much more efficiently than humans can.** If multiple AIs are collaborating on a problem, it is simple for them to exchange information. Human-led research is full of conflict over credit and attribution (such as the order of names on a paper, adversarial economic incentives (such as the lack of [open access](#) forcing universities to pay for papers written by their own researchers), and gaming metrics of recognition (such as fixation on the [h-index](#), which reflects not the intrinsic quality of someone's work, but the volume of citations). AI would not need vanity metrics, and could evaluate research based simply on what is most impactful.

The field of [metascience](#) has emerged to resolve these challenges, but they remain unsolved. ASI would not be susceptible to these failures, and it would leverage its computational power to comprehensively evaluate large volumes of data and identify

---

<sup>11</sup> Right now, the reason this can't be done is that AI researchers have no ideas how the AIs they grow work. But AGIs or ASIs would be much better at engineering, science, and ML than humans, to the point where they could take advantage of this ease of correcting software.

<sup>12</sup> See the issues unearthed by rationalism and behavioral economics.

complex high-dimensional patterns. The field is already moving in this direction: DeepMind has used AI to automate scientific research in [protein folding](#), [chips design](#), and [material science](#).

To summarize, an AGI capable of performing the same intellectual work that a human can would quickly become superintelligent, cloning itself to reach the intelligence level of humanity, and reprogramming itself to avoid any errors in its understanding. Compared to human intelligence which improves generationally, an AGI can propel itself to the heights of intelligence much faster than humanity ever could.

## Without intervention, ASI leads to godlike AI

There are practical limits to the growth of AI intelligence, but we can expect an intelligence that exceeds our own to scale much faster than human intelligence can. And it can easily overcome bottlenecks to improvement, be they hardware or software, institutions or people, simulations or real life experiments.

While we don't know the limits of intelligence, we do understand the limitations of physical systems. The estimates below describe the delta between our current outputs and theoretical ceilings, from which we can infer the direction of superintelligence's growth.

- **Energy will grow.** Energy production is bottlenecked by technological progress, which is bottlenecked only by our intelligence and effort. Per [mass-energy equivalence](#), we know that a ton of matter (including rocks) can theoretically power the whole world economy, we just haven't figured out how to extract this energy. ASI could learn to harvest this energy for its own operation, or even harvest the total energetic output of the Sun, which is trillions of times the world's current annual power consumption. Although the universe does not offer us infinite energy, we are far from reaching its fundamental limits, and should expect any growing intelligence to demand more and more energy.
- **Computers will improve.** The [physical limits of computations](#) and [computational complexity theory](#) tell us that hardware and software can still dramatically improve in information density and processing speed, which would result in more efficient computation. We are in the infancy of software engineering — we only recently formally verified small programs, and are nowhere near the [limits of proof automation architecture](#). In theory, it's possible to build software that never fails, at scale, on recursively more efficient machines and algorithms.
- **Communication will get faster.** We are nearing the theoretical limits of latency (the guaranteed speed of information transfer), with [optical fiber communications near](#)

[the speed of light](#) in a vacuum. However, we are not exercising the full capabilities of bandwidth (the volume of data transferred), which offers another dimension for quickening communication speed. The [Shannon-Hartley theorem](#) and [Bekenstein bound](#) are two theories that estimate bandwidth limits, but we are far from both. All types of bandwidth have continued to increase by orders of magnitude, so we can expect communication speed to improve.

Energy, computation, and communication are rate-limiting factors for the growth of an intelligence. Since we are nowhere near the bounds on any of these axes, we can assume that ASI can improve along all dimensions and in turn reap compounding benefits.

Other fields, such as engineering, would also benefit from these optimizations. As ASI grows in power, with more energy, better computation, faster communication, and overall greater intelligence, ASI will be able to achieve feats far out of reach of what humanity is capable of.

- **Mastery of big things.** Many concepts from science fiction are engineering problems rather than physical impossibilities, often conceptualized by [scientifically-educated authors](#). We first produced graphene, the strongest material, only 20 years ago, and [it is a billion times weaker](#) than the strongest theoretical materials. As engineering sciences improve and ASI marshals more energy, we should expect [megascale and astroscale feats](#). Ambitious projects such as [space travel](#), [space elevators](#), and [dyson spheres](#) may be within reach.
- **Mastery of small things.** At [the scale of the nanometer](#), we are already building incredibly precise and complex systems, despite the difficulties posed by many quantum effects. Transistors are now measured in nanometers, [thousands of times smaller](#) than they were decades ago, enabling massive compute progress that has driven the recent AI boom. Microbiology shows that complex specialized molecular machines, such as ribosomes, are possible, and that more advanced techniques beyond CRISPR might make it possible to rebuild life. ASI could master the recombination of DNA and atoms, or build micro-scale and milli-scale machinery.
- **Mastery of digital things.** The digital world is trivial to manipulate and scale — we can easily alter underlying programs and media and clone them. Extrapolating from what is already digitized, we can imagine a Matrix-like scenario. When it is technologically possible to manipulate perceptual signals directly to the brain, actual physical experience becomes more expensive than simulation. At this point, it could be possible to simulate reality.

ASI that can harness the power of the sun, compute at incomprehensible speeds, build atomic machines, and digitize reality is a god compared to humanity. Without intervention, artificial intelligence will reach this level of sophistication and power, and then go even further.

## Without intervention, godlike AI leads to extinction

The last few subsections unfolded the expected next steps of our current trajectory, where AIs consistently become more autonomous and powerful. More autonomous because their makers and users are already delegating more and more of the decisions and planning to them. And more powerful because of the relentless push of self-improvement catapulting AIs past human – and humanity’s – intelligence, unlocking scientific and technological powers at the level of gods.

The end result, if we don’t get wiped out before, is a world where AIs, not humans, are in control. Everything, from the economy to supply chains and education, would have been delegated to autonomous AIs who can just do anything better than a person, cheaper, faster. These AIs wouldn’t be simply tools, because after so many decisions and prioritizations delegated to them, they would be running the show.

This spells disaster for humanity because by default we will fail to ensure these godlike-AIs do exactly what we want (as argued in [the next section](#)), and this misalignment between what they aim for and what we want systematically leads to catastrophe for humanity.

For the moment, let’s just assume that we will indeed fail to make godlike-AIs do our biddings, and that they will have distinct goals from those of humanity – [the next section](#) defends this claim in detail. Even with this assumption, why would these godlike-beings wipe us out? They might not be controlled by us or do our bidding, but that doesn’t mean that they would necessarily hate us or plan our doom, right?

The missing piece here is that so many of the things we need and value are simply in the way of whatever godlike-AIs might aim for, and that they would have unlocked the power to just impose their might. AIs don’t need plants, they don’t need food, they don’t need oxygen, they certainly don’t need human beings to be thriving, healthy, or happy; all of these are things at best to be exploited, at worst to be ignored.

If godlike-AIs need more energy, they could simply wrest our electric grid from us, leading to total breakdown of our civilization. If they need even more energy, they could capture all of the sun’s radiated light<sup>13</sup>, leaving no sunlight for us and [creating devastating](#)

---

<sup>13</sup> For example through a [dyson sphere](#).

[consequences for organic life on Earth](#). If they need more compute (a useful subgoal for a software-based intelligence), they could swarm the Earth with datacenters, leveling cities in the process; cities which are also great repositories of materials to build such datacenters. And so on.

Worse, these examples don't even mention the many ways in which humans could be a slight annoyance, a thorn in the side of godlike-AIs, which would then give them reasons to deal with the humans.

When imagining this future, the picture to have in mind is not of some Greek gods who live in the clouds and mostly do their own things, sometimes interfering with human affairs; it's of a world owned and shaped by godlike-AIs, where we are ants scuttling on the ground. Contractors working on a house don't hate ants, yet their very work kills a tremendous amount of ants, wrecks their infrastructure, and destroys what they (literally) live for. And when we find ants in our kitchen or our bathrooms, we almost instantly turn to poison or pest control to handle the slight annoyance.

Unaligned godlike-AIs won't destroy us out of guile, but out of indifference. We would just be in the way of any ambitious goal that they might have, and if they're not aligned with us, they simply won't have any reason to care. And if we fight back... well, did ants ever succeed at bringing down humans?

The obvious next question is: why would godlike-AI not be under our control, not follow our goals, not care about humanity? Why would we get that wrong in making them? Already, the fact that [we grow our AIs instead of building them](#) hints at the answer. But let's fully address it in [the next section](#), arguing that this problem – dubbed alignment – is by far the hardest question that humanity has ever encountered, that we're not remotely close to solving it, and worse, that as a civilization we're not even trying to solve it.

## (4) AI Safety

### We are not on track to solve the hard problems of safety

Some time before we build AI that surpasses humanity's intelligence, we need to figure out how to make AI systems safe. Once AI exceeds humanity's intelligence, it will be in control, and our safety will depend on aligning AI's goals with humanity's best interests.

The alignment problem is not a mere technical challenge — it demands that we collectively solve one of the most difficult problems that humanity has ever tackled, requiring progress in fields that resist formalization, Nobel-prize-level breakthroughs, and billions or trillions of dollars of investment.

In [Defining alignment](#), we explain what alignment really means and why it's not just a technical problem but an all-encompassing civilizational one.

In [Estimating the cost of solving alignment](#), we explore what makes alignment challenging, and describe what it would cost in terms of effort to solve it and avert extinction.

In [Current technical efforts are not on track to solve alignment](#), we take a critical look at the current level of funding, organizations, and research dedicated to alignment. We argue that these efforts are insufficient, and that many of them do not even acknowledge the cost or complexity of the challenge.

In [AI will not solve alignment for us](#), we turn to the question of whether AI can help us solve alignment. We show that any potential benefits are mostly illusions, and argue that trying to use more advanced AI to solve alignment is a dangerous strategy.

Because both current and future safety efforts are not on track to solve alignment, we conclude that we are not on track to avert catastrophe from godlike AI.

### Defining alignment

In the field of AI, [alignment](#) refers to the ability to “steer AI systems toward a person's or group's intended goals, preferences, and ethical principles.”

With simpler systems that are less intelligent than humans, the alignment challenge addresses simpler safety issues, such as making current chatbots refuse to create propaganda or provide instructions for building weapons.

For systems that exceed human intelligence, the alignment problem is more complex and depends on guaranteeing that AI systems as powerful as godlike do what is best for humanity. This has a vastly larger scope than just censoring chatbots.

We already need to solve alignment today, which demands getting individuals, companies, and governments to act reliably according to some set of values. Alignment challenges vary depending on the scope and entity:

- To align individuals, we educate them to behave according to a certain set of cultural values. We also enforce compliance with the law through threat of state punishment. Most people are aligned and share values, with rare aberrations such as sociopathic geniuses or domestic terrorists.
- To align companies with societal values, we rely on regulations, corporate governance, and market incentives. However, companies often find loopholes or engage in unethical practices, such as how Boeing's [profit motive undermined safety](#), leading to crashes and hundreds of fatalities.
- To align governments with the will of the people, we rely on constitutions, checks and balances, and democratic elections. Some countries operate under dictatorships or authoritarian regimes. But both of these models can go wrong, leading governments to commit atrocities against their own people or experience democratic backsliding.
- To align humanity toward common goals like peace and environmental sustainability, we establish international organizations and agreements, like the United Nations and the Paris Climate Accords. On a global scale, enforcement is challenging — there are ongoing wars on multiple continents, and we have met only 17% of the [Sustainable Development Goals](#) (SDGs) that all United Nations member states have agreed to.

The examples show that alignment relies on processes that reliably incentivize entities to pursue good outcomes, based on some set of values. In each of the instances above, we need to design processes to determine values (e.g. constitutional conventions), reconcile them (e.g. voting), enshrine them (e.g. constitutions, amendments, laws), oversee and

enforce them (e.g. institutions and police), and coordinate the constituent parts (e.g. administrations).

A system is aligned if there is a mechanistic connection between the original values and reliable outcomes. For example, while UN member states all share the value of protecting the environment and strive toward [the Sustainable Development Goals](#), they lack reliable processes to ensure traction. Regardless of intention, without concrete processes we cannot consider the UN successfully aligned with protecting the environment.

While they often fail us, we currently entrust the fate of the world to governments, corporations, and international institutions.

**AI alignment demands solving all of the same problems our current institutions try to solve, but instead use software to do it.**

As AI becomes more intelligent, its causal impact will increase, and misalignment will be more consequential. We must find a way to install our deepest values in AI, addressing questions ranging from how to raise children, to what kinds of governance to apply to which problems.

Solving the alignment problem is philosophy on a deadline, and requires defining and reconciling our values, enshrining them in robust processes, and entrusting those processes to AIs that may soon be more powerful than we are.

## Estimating the cost of solving alignment

Although alignment is not an impossible problem, it is extremely difficult and requires answering novel social and technical questions humanity has not yet solved. By considering some of these questions, we can understand how much it would cost to solve this problem.

**What do we value and how do we reconcile contradictions in values?** We must align godlike AI with “what humanity wants,” but what does this even mean?

It is clear that even as individuals, we often don’t know what we want. For example, if we say and think that we want to spend more time with our family, but then end up playing games on our phones, which one do we really want? Individuals often have multiple conflicting desires or unconscious preferences that make it difficult to know what someone really wants.

When we zoom out from the individual to groups, up to the whole of humanity, the complexity of “finding what we want” explodes: when different cultures, different religions, different countries disagree about what they want on key questions like state interventionism, immigration, or what is moral, how can we resolve these into a fixed set of values? If there is a scientific answer to this problem, we have made little progress on it.

If we cannot find, build, and reconcile values that fit with what we want, we will lose control of the future to AI systems that ardently defend a shadow of what we actually care about.

Making progress on understanding and reconciling values requires ground-breaking advances in the fields of psychology, neuroscience, anthropology, political science, and moral philosophy. The former fields are necessary for diving into the human psyche [resolving uncertainties related to human rationality, emotion, and biases](#), and the latter two are necessary for finding ways to resolve conflicts between these.

**How can we predict the consequences of our actions?** A positive understanding of “what we want” is insufficient to keep AI safe: we also need to understand the consequences of getting what we want, to avoid unwanted side effects.

Yet history demonstrates how often we fail to see consequences of our actions until after they are implemented. [The Indian vulture crisis](#) was a massive environmental disaster in which a new medicine given to cows turned out to be toxic for vultures, which died by millions upon eating the carcasses. The collapse in vulture population meant that carcasses were not cleaned, contaminating water sources, providing breeding grounds for feral dogs with rabies, and ultimately leading to a humanitarian disaster costing billions due to a single unknown externality.

The same can happen for designing institutions. [The Articles of Confederation](#) was the first attempt to create a US government, but they left Congress completely impotent to govern the individual states, so this had to be corrected in the [US constitution](#).

Progress on our ability to predict the consequences of our actions requires better science in every technical field, and learning what to do with these predictions requires progress in fields like non-idealized decision making. The last 100 years have seen some progress in [scientific thinking](#) and [decision theory](#), and some efforts in [rationalism](#) have even attempted to inspire [better decision-making in light of the AI problem](#). But while better decision-making has had clear consequences in fields like investment – quantitative strategies are increasingly outperforming discretionary ones – most people make decisions the same way we did 100 years ago.

To confidently move forward on these questions, we need faster science, simulation, and modeling; breakthroughs in fields related to decision-making; and better institutions that demonstrate these approaches work.

**Process design for alignment:** If we can answer the philosophical questions of values alignment, and get better at predicting and avoiding consequential errors, we still need to build processes to ensure that our values are represented in systems and actually enacted in the real world.

Often, even the most powerful entities fail to build processes that connect the dots between values and end outcomes. Nearly every country struggles with taxation, particularly of large entities and high net worth individuals. And process failures abound in history, such as the [largest famine ever](#), which was caused by inefficient distribution of food within China's planned economy during the Great Leap Forward.

The mechanism design of these entities and their implementation are two separate things. When we zoom in on the theory, our best approaches aren't great. The field of [political philosophy](#) attempts to make progress on statecraft, but ideas like the separation of powers in many modern constitutions are based on [250-year-old theories from Montesquieu](#). New ideas in [voting theory](#) have been proposed, and efforts like [blockchain governance](#) try to implement some of these, but these have done little so far to displace our current systems. Slow theoretical progress and little implementation of new systems suggest massive room to improve our current statecraft and decision-making processes.

On the corporate level, we have the [theory of the firm](#) and [management theory](#) to tell us about how to run companies, but the state of the art in designing a winning company today looks a lot closer to Y-combinator's oral theory of knowledge and knowing the right people than a science. And even then, the failure rate is very high.

Neither statecraft nor making a company is a scientific process in which there are formal guarantees, and things often go very wrong. In the context of the alignment problem, this demonstrates large gaps in humanity's knowledge that expose huge risks if we were to imagine trusting advanced AI systems to run the future. Without better theories and implementation, these systems could make the same mistakes, with larger consequences given their greater intelligence and power.

**Guaranteeing alignment:** Last but not least, even if we can design processes to align AIs and we know what to align them to, we still need to be able to guarantee and check that they will actually do what we want. That is, as in any critical technology, we want guarantees that it won't create a catastrophe before turning it on.

This is already difficult with normal software: making (almost) bug-free systems require the use of [formal methods](#) which are both expensive (in time, skill, effort) and in their infancy, especially with regard to the kind of complex properties that we would care about for AIs acting in the world.

And this is even harder with AIs built with the current paradigm, due to the fact that they are not built by hand (like normal software), but instead grown through mathematical optimization. This means that even the makers of AI systems have next to no understanding of what they can and cannot do, and no predictive capabilities whatsoever to anticipate what they will do before training them, or even just before using them.

But the situation is actually worse than that: whereas most current AI systems are still less smart than humans, alignment actually requires getting guarantees on systems that are significantly smarter than humans. That is, in addition to managing the complexity of software, and the obscurity of neural networks, we need to figure out how to check an entity which can outsmart us at every turn.

-

Even if all of these questions need not be answered at once, we nonetheless need to invent a process by which they are answered. Currently, our human science and morality is inadequate to address the risks posed by advanced AI, and it must improve for us to have a chance.

How much would this cost, in terms of funding and human effort?

When we look at major research efforts that humans have pursued in the past which led to breakthroughs and Nobel prizes, we can begin to envision what such a “significant research project” constitutes. The [Manhattan Project](#) cost \$27B to produce the first nuclear weapons, and at its height employed 130,000 people. Over four years, researchers and engineers cracked problem after problem to develop the bomb, with over 31 Nobel-prize winners tied to the project. Another massive research effort, the [Human Genome Project](#) (HGP), cost \$5B over 13 years and required contributions from thousands of researchers from various countries.

If alignment was of the same difficulty level as these problems, we would assume *at least* a tens-of-billions of dollars effort, featuring thousands of people, with dedicated coordination, and multiple breakthroughs of Nobel-prize magnitude.

But the cost of alignment is almost guaranteed to be significantly higher. While the HGP and Manhattan Project were profoundly difficult projects, they were concentrated in narrower domains of study in fields that were already hard sciences. In contrast, [alignment is a pre-paradigmatic field](#), in which many of the questions we need to answer resist study. Many of the domains above that require advances (psychology, anthropology, economics, simulation) are not hard sciences, but would need to become them – similar to the transitions from alchemy to chemistry. And at the end of the day, alignment needs to be *just right*, as “[close but dangerously wrong](#)”<sup>14</sup> answers could lead to death.

Given the magnitude of the danger ahead, the complexity and uncertainty of these estimates should make us even more careful, and cause us to assume that the costs may be higher still than what is presented here.

We conclude that solving alignment is extremely hard, and the cost is clearly very high: at least billions, maybe trillions, with a time frame of decades, and with research of a Nobel-prize-winning quality.

## Current technical efforts are not on track to solve alignment

The field of AI Safety is not making meaningful progress on or investment in alignment; current funding and focus are insufficient, and the research approaches being pursued do not attend to the hard problems of aligning AI.

On the capabilities side, there are reports that OpenAI and Microsoft plan to build a [\\$100B data center](#), and Demis Hassabis has commented that Google DeepMind is “[investing more than that over time](#).”

In comparison, government AI Safety Institutes are funded on the order of tens to hundreds of millions, with the [UK’s AISI allocated £100m](#), and alignment funding outside AGI companies and government [estimated at \\$100m/year](#). At AGI companies such as DeepMind, alignment research efforts are run by [small teams](#), and some (like OpenAI) are presently [suffering from mass exodus](#). The most concerted alignment-related investment focuses on [compute-intensive interpretability experiments](#) and on the teams that run them, but this is unlikely to reach beyond the tens of millions.

---

<sup>14</sup> Specification gaming is the idea that an AI can satisfy the goal that a developer sets out for it during training, but end up following a different goal later. This occurs for a few reasons, but the result is dangerous in every case involving superintelligence: with systems this powerful, slight divergences from the original goal could lead to catastrophic divergences in downstream actions.

And these are optimistic estimates; in reality, only a tiny fraction of this total goes to genuine AI alignment efforts. With few exceptions, the majority of funding is directed at problems associated with AI safety, rather than paying the exorbitant cost of alignment.

Nearly all current technical safety approaches are limited in their efficacy and trail their own stated goals:

- **Black-box evaluations and red-teaming** aim to test a model's capabilities to evaluate how powerful or dangerous it is. With this strategy, the theory of change is that identifying dangerous behavior could force AI companies to pause development or governments to coordinate on regulation. Teams working on evaluations include [AI Safety Institutes](#), [METR](#) and [Evaluations at Anthropic](#).

Black-Box Evaluations can only catch all relevant safety issues insofar as we have either an exhaustive list of all possible failure modes, or a mechanistic model of how concrete capabilities lead to safety risks. We currently have neither, so evaluations boil down to ad-hoc lists of tests that capture some possible risks. These are insufficient even for today's models, as demonstrated by the fact that [current LLMs can notice they are being tested](#), which the evaluators and researchers did not even anticipate.

- **Interpretability** aims to reverse-engineer the concepts and thought processes a model uses to understand what it can do and how it works. This approach presumes that a more complete understanding of the systems could prevent misbehavior, or unlock new ways to control models, such as training them to be fully honest. Teams working on interpretability include [Interpretability at Anthropic](#) and [Apollo Research](#).

Interpretability's value depends on its ability to fully understand and reverse engineer AI systems to check if they have capabilities and thoughts that might lead to unsafe actions. Yet current interpretability research is unable to do that even for LLMs a few generations back (GPT2), let alone for the massive and complex models used in practice today (Claude and GPT4/o1).

And even with full understanding and reverse engineering of state of the art LLMs, interpretability is blind to any form of extended cognition, such as what the system can do when connected to the environment (notably the internet), given tools, interacting with other systems or instances of itself. A huge part of recent progress in AI comes from moving to agents and scaffolding that leverage exactly this form of extended cognition. Just as solving neuroscience would be insufficient to explain how a company works, even full interpretability of an LLM would be insufficient to

explain most research efforts on the AI frontier.

- **Whack-A-Mole Fixes** use techniques like RLHF, fine-tuning, and prompting to remove undesirable model behavior or a specific failure mode. The theory of change is that current safety problems can be solved in a patchwork manner, addressed as they arise, and that we can perhaps learn from this process to correct the behavior of more advanced systems. Teams working on this include [Alignment Capabilities at Anthropic](#) and [OpenAI's Safety Team](#).

Whack-A-Mole fixes, from RLHF to finetuning, are about teaching the system to not demonstrate problematic behavior, not about fundamentally fixing that behavior. For example, a model that produces violent text output may be finetuned to be more innocuous, but the underlying base model is just as capable of producing violent content as ever. The problem as to how this behavior arose in the first place is left unaddressed by even the best finetuning. By pushing for models to hide unsafe actions rather than resolving underlying issues, whack-a-mole fixes lead to models that are more and more competent at hiding their issues and failures, rather than models that are genuinely safer.

At best, these strategies can identify and incrementally correct problems, address model misbehavior, and use misbehavior as a red flag to motivate policy solutions and regulations. However, even according to their proponents, these strategies do not attempt to align superhuman AI, but merely try to align the next generation of systems, trusting that partial alignment of the Nth systems will help align the N+1 system.

This approach can be thought of as **Iterative Alignment**, a strategy that rests on the hope that we can build slightly smarter systems, align those, and use them to help align successor systems, repeating the process until we reach superintelligent AI. OpenAI's [Superalignment](#) plan explicitly states this:

*Currently, we don't have a solution for steering or controlling a potentially superintelligent AI, and preventing it from going rogue. Our current techniques for aligning AI, such as reinforcement learning from human feedback, rely on humans' ability to supervise AI. But humans won't be able to reliably supervise AI systems much smarter than us, and so our current alignment techniques will not scale to superintelligence. We need new scientific and technical breakthroughs.*

*Our goal is to build a roughly human-level automated alignment researcher. We can then use vast amounts of compute to scale our efforts, and iteratively align superintelligence. To align the first automated alignment researcher, we will need to 1) develop a scalable training method, 2) validate the resulting model, and 3) stress test our entire alignment pipeline.*

The plans of nearly every other AI company are similarly limited and failure-prone attempts at iterative alignment. Deepmind's [2024 update on AGI safety approaches](#) discusses the evaluative techniques listed above and names "amplified oversight" as its focus. Anthropic's [Core Views on AI Safety](#) describes how the evaluative techniques can be combined in a "portfolio approach" to keep advanced AI safe, and offers a similar justification for iterative alignment:

*Turning language models into aligned AI systems will require significant amounts of high-quality feedback to steer their behaviors. A major concern is that humans won't be able to provide the necessary feedback. It may be that humans won't be able to provide accurate/informed enough feedback to adequately train models to avoid harmful behavior across a wide range of circumstances. It may be that humans can be fooled by the AI system, and won't be able to provide feedback that reflects what they actually want (e.g. accidentally providing positive feedback for misleading advice). It may be that the issue is a combination, and humans could provide correct feedback with enough effort, but can't do so at scale. This is the problem of scalable oversight, and it seems likely to be a central issue in training safe, aligned AI systems.*

*Ultimately, we believe the only way to provide the necessary supervision will be to have AI systems partially supervise themselves or assist humans in their own supervision. Somehow, we need to magnify a small amount of high-quality human supervision into a large amount of high-quality AI supervision. This idea is already showing promise through techniques such as RLHF and Constitutional AI, though we see room for much more to make these techniques reliable with human-level systems. We think approaches like these are promising because language models already learn a lot about human values during pretraining. Learning about human values is not unlike learning about other subjects, and we should expect larger models to have a more accurate picture of human values and to find them easier to learn relative to smaller models. The main goal of scalable oversight is to get models to better understand and behave in accordance with human values.*

Regardless of whether or not one believes that this strategy will work, it is clear that this approach does not adequately address the true complexity of alignment. A meaningful attempt at alignment must integrate moral philosophy to understand values reconciliation, implement formal verification to make guarantees about system properties, consider humanitarian questions of what we value and why, and propose institution design, at minimum. All current efforts fail to do so.

Today's AI safety research is vastly underfunded compared to investments in capabilities work, and the majority of technical approaches intentionally do not address the conceptual complexity of alignment, instead operating in a reactive empiricist framework that simply identifies misbehavior once it already exists. Humanity's current AI safety plan is to race toward building superintelligent AI, and delegate the most difficult questions of alignment to AI itself. This is a naive and dangerous approach.

## AI will not solve alignment for us

For a safe future, we must solve the hard problems of alignment, allocating adequate research hours, investment, and coordination effort. [OpenAI](#), [Deepmind](#), [Anthropic](#), X.AI (“[accelerating human scientific discovery](#)”), and [others](#) have all proposed deferring and outsourcing these questions to more advanced future AI systems.

But on reflection, this is an incredibly risky approach. [Situational Awareness](#), a document written by ex-OpenAI superalignment researcher Leopold Aschenbrenner which has gotten [significant traction](#) even from [popular news outlets](#), puts the argument bluntly. Aschenbrenner argues for a vision of the future in which AI becomes powerful extremely quickly due to scaling up the orders of magnitude (“OOMs”) of AI models. When discussing future safety approaches, he makes a vivid argument for iterative alignment:

*Ultimately, we're going to need to automate alignment research. There's no way we'll manage to solve alignment for true superintelligence directly; covering that vast of an intelligence gap seems extremely challenging. Moreover, by the end of the intelligence explosion—after 100 million automated AI researchers have furiously powered through a decade of ML progress—I expect much more alien systems in terms of architecture and algorithms compared to current system (with potentially less benign properties, e.g. on legibility of CoT, generalization properties, or the severity of misalignment induced by training).*

*But we also don't have to solve this problem just on our own. If we manage to align somewhat-superhuman systems enough to trust them, we'll be in an incredible position: we'll have millions of automated AI researchers, smarter than the best AI researchers, at our disposal. Leveraging these army of automated researchers properly to solve alignment for even-more superhuman systems will be decisive.*

*Getting automated alignment right during the intelligence explosion will be extraordinarily high-stakes: we'll be going through many years of AI advances in mere months, with little human-time to make the right decisions, and we'll start entering territory where alignment failures could be catastrophic.*

The dangers here are explicit: alien systems, huge advances in mere months, and a tightrope walk through an “intelligence explosion” in which wrong choices could lead to catastrophe.

But even before we get to a dramatic vision of the AI future, the iterative alignment strategy has an ordering error – we first need to achieve alignment to safely and effectively leverage AIs.

Consider a situation where AI systems go off and “do research on alignment” for a while, simulating tens of years of human research work. The problem then becomes: how do we check that the research is indeed correct, and not wrong, misguided, or even deceptive? We can’t just assume this is the case, because the only way to fully trust an AI system is if we’d already solved alignment, and knew that it was acting in our best interest at the deepest level.

Thus we need to have humans validate the research. That is, even automated research runs into a bottleneck of human comprehension and supervision.

Proponents of iterated alignment argue that this is not a real issue, because “evaluation is easier than generation.” For example, Aschenbrenner further argues in [Situational Awareness](#) that:

*We get some of the way [to superalignment] “for free,” because it’s easier for us to evaluate outputs (especially for egregious misbehaviors) than it is to generate them ourselves. For example, it takes me months or years of hard work to write a paper, but only a couple hours to tell if a paper someone has written is any good (though perhaps longer to catch fraud). We’ll have teams of expert humans spend a lot of time evaluating every RLHF example, and they’ll be able to “thumbs down” a lot of misbehavior even if the AI system is somewhat smarter than them. That said, this will only take us so far (GPT-2 or even GPT-3 couldn’t detect nefarious GPT-4 reliably, even though evaluation is easier than generation!)*

The argument holds for standard peer-review, where the authors and reviewers are generally on the same intellectual level, with sensibly similar cognitive architecture, education, and knowledge. But this does not apply to automated alignment research, where to be useful the research needs to be done by AIs that are both smarter and faster than humans.

The appropriate analogy is not one researcher reviewing another, but rather a group of preschoolers reviewing the work of a million Einsteins. It might be easier and faster than doing the research itself, but it will still take years and years of effort and verification to check any single breakthrough.

Fundamentally, the problem with iterative alignment is that it never pays the cost of alignment. Somewhere along the story, alignment gets implicitly solved – yet no one ever proposes an actual plan for doing so beyond “the (unaligned) AIs will help us”.

There are other risks with this approach as well.

The more powerful AI we have, the faster things will go. As AI systems improve and

automate their own learning, AGI will be able to improve faster than our current research, and ASI will be able to improve faster than humanity can do science. The dynamics of intelligence growth means that it is possible for an ASI “about as smart as humanity” to move to “beyond all human scientific frontiers” on the order of weeks or months. While the change is most dramatic with more advanced systems, as soon as we have AGI we enter a world where things begin to move much quicker, forcing us to solve alignment much faster than in a pre-AGI world.

Tensions between world powers will also heat up as AI becomes more powerful, something we are already witnessing in [AI weapons used in warfare](#), [global disinformation campaigns](#), the [US-China chip war](#), and how [Europe is struggling with regulation around Big Tech](#). As we move towards AGI, ASI, and eventually godlike AI, pressure on existing international treaties and diplomacy methods will be pushed beyond their limits. Unlike with nuclear war, there is not necessarily the same promise of mutually assured destruction with AI that could create a (semi)stable equilibrium. Ensuring geopolitical stability is necessary to create supportive conditions to solve the hard problems of alignment, something that gets more challenging if AI is becoming rapidly more powerful.

AGI and its successor AIs will also cause massive political, economical, and societal destabilization through automating disinformation and online manipulation, job automation, and other shifts that look like “issues seen today but magnified as systems grow stronger”. This in turn makes coordination around massive research projects like the ones necessary to solve alignment extremely difficult.

Thus, iterative alignment fails on multiple accounts. In addition to not addressing the hard parts of alignment, it also encourages entering a time-pressured and precarious world.

-

We have seen that alignment is an incredibly complex technical and social problem, one of the most complex any civilization needs to handle. And while the costs are enormous, no one is even starting to pay them, instead hoping that they will disappear by themselves as AIs become more powerful.

In light of this failure to address the risks of godlike-AI from a research angle, it’s necessary to aggressively slow down and regulate AI progress, in order to avoid the catastrophe ahead. This comes from strong AI regulations, policies, and institutions.

Unfortunately, as we explore next, the landscape is as barren here as it is on the research side.

## (5) AI Governance

### We lack the mechanisms to control technology development

Because we are unlikely to align godlike AI in time, we must find means to control AI development and avert the catastrophic default path altogether. This requires preventing the creation of AGI or implementing stop-gaps to prevent AGI from scaling to superintelligence and beyond. The most straightforward path to controlling AI development is through policy and governance; but these efforts are currently not on track.

In [Defining AI Governance](#), we explore what kind of governance is necessary to avoid existential risks, notably slowing down and regulating AI research progress, and blocking the race to AGI. We highlight [A Narrow Path](#) as a comprehensive strategy for preventing the creation of superintelligence for 20 years.

In [Estimating what is necessary to control AI development](#), we argue that controlling AI development would require effective oversight by governments over private sector or rogue actors, and mutual oversight by governments over each other.

In [Current AI policy efforts are not on track to control AI development](#), we take a critical look at the current policy and governance efforts that focus on existential risk from AI, and find them lacking. In addition to a relative dearth of impactful efforts, most efforts fail because they implicitly endorse the status quo.

In [Current AI policy efforts endorse the race to AGI](#), we argue that the true reason AI policy efforts are not on track to control superintelligence risk is that many of them derive from the very actors racing to build AGI.

### Defining AI governance

AI governance must implement mechanisms that allow humanity to evaluate the risks and potential benefits of AGI and other advanced AIs. Miotti et al's [A Narrow Path](#) offers a comprehensive vision, describing three phases of policy development:

**Phase 0: Safety:** *New institutions, legislation, and policies that countries should implement immediately that prevent development of AI that we do not have control of. With correct*

execution, the strength of these measures should prevent anyone from developing artificial superintelligence for the next 20 years.

**Phase 1: Stability:** International measures and institutions that ensure measures to control the development of AI do not collapse under geopolitical rivalries or rogue development by state and non-state actors. With correct execution, these measures should ensure stability and lead to an international AI oversight system that does not collapse over time.

**Phase 2: Flourishing:** With the development of rogue superintelligence prevented and a stable international system in place, humanity can focus on the scientific foundations for transformative AI under human control. Build a robust science and metrology of intelligence, safe-by-design AI engineering, and other foundations for transformative AI under human control.

A Narrow Path is not a policy proposal so much as a thesis on how to avoid extinction by superintelligence. It is one of the few comprehensive visions of global governance that adequately contends with superintelligence risk and acknowledges and considers the immense challenges of alignment. As [one commenter writes](#):

*The key is you must pick one of these:*

1. *Building a superintelligence under current conditions will turn out fine.*
2. *No one will build a superintelligence under anything like current conditions.*
3. *We must prevent at almost all costs anyone building superintelligence soon...*

*...If you know you won't be able to bite either of those first two bullets? Then it's time to figure out the path to victory, and talk methods and price. And we should do what is necessary now to gather more information, and ensure we have the option to walk down such paths.*

In other words, superintelligent risk forces us to contend with the worst case scenario, and the default path is perilous. We have not come across any comparably robust proposals for averting extinction risk, so A Narrow Path's proposal is the best plan we have today.

## Estimating what is necessary to control AI development

To prevent the default trajectory where AI progress leads to AGI, ASI, and godlike AI, civilization must be able to pause or stop AI progress when it becomes necessary for safety. We lack that ability today.

In order to stop AI development, we would need to rein in:

- The large companies racing to build AGI.
- Nation states such as the US, China, and others with AGI capacity.

- Academic research.
- Open-weight releases and research.

Because the barrier to entry is low enough that [new private actors keep entering the race](#) to AGI, this list will grow.

Any regulation needs to be coordinated to prevent defection by a rogue actor. At the very least, we would need:

- National regulations in the US and other world powers to govern leading AI companies and academic research.
- International regulations on open-weight publishing, contribution, and ownership to prevent bad actors from building AGI outside of regulated bodies.
- International coordination through treaties, high-bandwidth communication lines, and multinational agreements enforced by international law, to deter any nation state from racing to superintelligence and endangering everyone.

These regulations require concrete technical systems to detect violations, such as catching high-compute training runs, and mechanisms to implement restrictions, such as physical kill-switches that could shut down datacenters in which AI runs. We lack these regulations and mechanisms today.

It's important to realize that these regulations are there to prevent the need for much stronger ones. Because if companies are allowed to release open-weight AGIs that can run on consumer hardware, the only way for governments to avoid catastrophe from AI would be policing what happens on every single computer in the world to catch all dangerous uses. A world with open-weight AGI is at best a world where personal freedoms are sacrificed for the survival of humanity.

To avoid this fate, political will is insufficient; it is not enough just for people to *want* AI to be safe, or *want* to control AI. At minimum, we need institutions, policy frameworks, technical levers, communication lines, and people who understand the risks of superintelligence in positions of authority.

**Just as alignment depends on connecting our values to good future outcomes, AI governance is about building a process<sup>15</sup> that connects humanity's will to the AI**

---

<sup>15</sup> A future draft owes a better description of a "process." But what we mean is roughly a reliable sequence of physical steps (e.g. sending a message, stamping a document for formal approval) with an input and output. Thinking about the world in this mechanistic way is foundational to our thesis – if we can't draw a physical trace from the first step to the last step, we argue that no such process exists.

**technological frontier.** Without a process to slow or stop the development of AI even if humanity wanted to, then it is straightforwardly the case that humanity is not in control of AI.

## Current AI policy efforts are not on track to control AI development

The last few years have seen increased AI policy and governance efforts, but few focus on superintelligence risks. Those that do are insufficient, and sometimes actively harmful by implicitly endorsing the current race to AGI.

Today's best attempt at an international reckoning with the risks of superintelligence is the [Bletchley Declaration](#), a statement from the first International AI Safety Summit, which acknowledges "catastrophic" risks and promises future international cooperation<sup>16</sup>:

*There is potential for serious, even catastrophic, harm, either deliberate or unintentional, stemming from the most significant capabilities of [frontier] AI models. Given the rapid and uncertain rate of change of AI, and in the context of the acceleration of investment in technology, we affirm that deepening our understanding of these potential risks and of actions to address them is especially urgent.*

*Many risks arising from AI are inherently international in nature, and so are best addressed through international cooperation. We resolve to work together in an inclusive manner to ensure human-centric, trustworthy and responsible AI that is safe, and supports the good of all through existing international fora and other relevant initiatives, to promote cooperation to address the broad range of risks posed by AI.*

Following the United Kingdom's example, governments around the world are creating national [AI Safety Institutes](#). While their resourcing and missions vary widely, they are broadly tasked with helping governments build capacity to make more informed decisions on how to manage the development and deployment of advanced AI systems.

On the regulatory side, some countries have started legislating AI development. The EU has passed its [AI Act](#). China has drafted [Interim Measures](#) to oversee generative AI. The UK will consider AI regulation in an [upcoming bill](#). The US has issued an [executive order](#) on AI oversight, though California's governor recently vetoed the first meaningful legislation for controlling advanced AI risks, [SB 1047](#). These proposals [similarly](#) emphasize the possible need for future controls, but do little if anything to restrict frontier AI development today.

---

<sup>16</sup> A [message](#) from the second summit in Seoul goes further, with AGI companies agreeing to recognise that there are some "intolerable" risks from AI after which point AI development should stop.

Internationally, the [UN established an AI-focused advisory body](#) and the [OECD](#) and [UNESCO](#) have formed AI working groups. The US and some of its allies are attempting to limit China's AI development by [restricting the export of the required computational resources](#) while [academic researchers are pursuing more collaborative Track 2 dialogues](#) to build common ground on AI governance.

Meanwhile, AGI companies are claiming to adopt self-regulatory measures, such as Anthropic's [Responsible Scaling Policies](#), OpenAI's [Preparedness Framework](#), and DeepMind's [Frontier Safety Framework](#).

All of these efforts are ineffective:

- **We lack effective intervention mechanisms.** The [hardware to enforce pauses in AI development](#) is currently just a prototype maintained by one small team. Policies that establish mandatory thresholds for AI development (e.g. preventing AI self-replication or placing limitations on general capabilities) do not exist. SB 1047 was vetoed due to significant lobbying by [Big Tech](#) and [VCs](#), despite the fact that it was popular enough to pass the California legislature, receive [77% voter support](#), and garner endorsement from [Nobel Prize-winning AI researchers](#) and [hundreds of AI company employees](#).
- **Relevant technical actors are not overseen.** We lack national regulators that have authority over the specific challenges posed by AI. In the US, the White House's voluntary commitments have failed to introduce [meaningful transparency or accountability](#). Globally, AI Safety Institutes lack formal power over AI developers and must rely on voluntary cooperation, which is often limited and tenuous. Of the 16 AI companies that signed the [Frontier AI Safety Commitments](#) after the latest AI Summit in Seoul, most have failed to provide transparency around their safety plans; [X.ai](#) has not published a safety strategy. As it stands, national oversight endorses the voluntary evaluations framework proposed by AGI companies, concentrating power in the very actors driving the risks.
- **Open-weight development is unregulated.** Open-weight AGI development is thriving: [hundreds of AI papers are published daily](#), and open-source repositories regularly push capabilities in agent frameworks beyond the state of the art. Whenever new levels of capabilities are unlocked by AI companies, open-weight models are released soon after and optimized to be used by anyone without any check or supervision. This means giving powerful and dangerous tools, for free, to anyone with a computer. And when the inevitable happens, such as the creation and distribution of child sexual abuse pornography, one of the most heinous uses of

open-weight AIs<sup>17</sup>, [only the end user](#) is condemned in the rare cases where anyone is sanctioned, leaving off the hook both model creators and the websites hosting them. This is an unsustainable situation, a boiling kettle ready to explode with the release of open-weight AGI.

- **International groups on AI lack power.** In the past, international governance bodies for nuclear, biochemical, and human cloning research have been relatively effective, offering [lessons we can apply to AI regulation](#). We do not have equivalent AI institutions today, and the current international groups are mere advisory bodies.
- **There is no plan for a safe AI future.** A meaningful plan to address AI risk must contain provisions capable of preventing the creation of superintelligence and managing its threats. No such proposals exist, and there are no international bodies or appointed individuals responsible for drafting or enforcing one. [A Narrow Path](#) is an exception, but its recency and necessity points at the inadequacy of official governance efforts.

Even advocates of current coordination efforts agree that they are insufficient; AI Safety Institutes are [well aware of their tenuous relationship with AGI companies](#).

**The majority of existing AI safety efforts are reactive rather than proactive, which inherently puts humanity in the position of managing risk rather than preventing it.**

This reactive approach has both technical and policy components.

- The technical strategy includes the AI safety efforts discussed in Section 4: black-box evaluations and red-teaming, interpretability, whack-a-mole fixes, and iterative alignment. Each of these solutions assume that we can catch problems with AI as they arise, and then alert policymakers of the risks.
- The policy strategy assumes that we can then use these risks to build consensus for new legislation, or trigger regulatory actions such as shutting down AI development.

Holden Karnofsky, ex-co-CEO and ex-director of AI Strategy at Open Philanthropic (an organization which has funded multiple AI governance efforts) calls this the [“if-then” framework](#):

*If-then commitments... are commitments of the form: “If an AI model has capability X, risk mitigations Y must be in place. And, if needed, we will delay AI deployment and/or development to ensure the mitigations can be present in time.” A specific example: “If an AI*

---

<sup>17</sup> A heinous use for which the AIs are trained, [since child abuse pictures are part of the training data of these image generation AIs](#).

model has the ability to walk a novice through constructing a weapon of mass destruction, we must ensure that there are no easy ways for consumers to elicit behavior in this category from the AI model.”

*If-then commitments can be voluntarily adopted by AI developers; they also, potentially, can be enforced by regulators. Adoption of if-then commitments could help reduce risks from AI in two key ways: (a) prototyping, battle-testing, and building consensus around a potential framework for regulation; and (b) helping AI developers and others build roadmaps of what risk mitigations need to be in place by when. Such adoption does not require agreement on whether major AI risks are imminent—a polarized topic—only that certain situations would require certain risk mitigations if they came to pass.*

The UK AI Safety Institute [explains evaluations in a similar manner](#):

*A gold standard for development and deployment decisions of frontier AI would include a comprehensive set of clearly defined risk and capability thresholds that would likely lead to unacceptable outcomes, unless mitigated appropriately. We think governments have a significant role to play, as 27 countries and the EU recognised at the AI Seoul Summit 2024. We have thus begun work to identify capability thresholds: specific AI capabilities that are indicative of potentially severe risks, could be tested for, and should trigger certain actions to mitigate risk. They correspond to pathways to harm from our risk modelling, such as capabilities that would remove current barriers for malicious actors or unlock new ways of causing harm.*

Anthropic describes the reactive approach in its [Responsible Scaling Policy](#):

*Since the frontier of AI is rapidly evolving, we cannot anticipate what safety and security measures will be appropriate for models far beyond the current frontier. We will thus regularly measure the capability of our models and adjust our safeguards accordingly. Further, we will continue to research potential risks and next-generation mitigation techniques. And, at the highest level of generality, we will look for opportunities to improve and strengthen our overarching risk management framework.*

This approach is flawed:

- 1. The reactive framework reverses the burden of proof from how society typically regulates high-risk technologies and industries.**

In most areas of law, we do not wait for harm to occur before implementing safeguards. Banks are prohibited from facilitating money laundering from the moment of incorporation, not after their first offense. Nuclear power plants must demonstrate safety measures before operation, not after a meltdown.

The reactive framework problematically reverses the burden of proof. It assumes AI systems are safe by default and only requires action once risks are detected. One of the core dangers of AI systems is precisely that we do not know what they will do or how powerful they will be before we train them. The if-then framework opts to proceed until problems arise, rather than pausing development and deployment until we can guarantee safety. This implicitly endorses the current race to AGI.

This reversal is exactly what makes the reactive framework preferable for AI companies. As METR, the organization that [first coined the term](#) “responsible scaling policy,” writes:

*RSPs are intended to appeal to both (a) those who think AI could be extremely dangerous and seek things like moratoriums on AI development, and (b) those who think that it's too early to worry. Under both of these views, RSPs are a promising intervention: by committing to gate scaling on concrete evaluations and empirical observations, then (if the evaluations are sufficiently accurate!) we should expect to halt AI development in cases where we do see dangerous capabilities, and continue it in cases where worries about dangerous capabilities have not yet emerged.*

## 2. The reactive framework overlooks the dynamics of AI development that make regulation more difficult over time.

AI is being developed extremely quickly and by many actors, and the barrier to entry is low and quickly diminishing. The biggest GPT-2 model (1.5B parameters) [cost an estimated \\$43,000](#) to train in 2019; today it is possible to train a 350M parameters GPT-2 [for \\$200 in 14 hours](#). Paul Graham mentions an estimate that the ratio of training price by performance [decreased 100x in each of the last two years, or 10000x in two years](#). Due to low standards of information protection, the “secret sauce” of training AI systems is becoming increasingly common knowledge, making it easier to create powerful systems. Since the latest models almost always get emulated and reproduced in some way in open-weight, there are truly no “take-backs” with AI development. Technical releases that advance the state of the art have a ratcheting effect that is difficult and sometimes impossible to reverse.

Because of these dynamics, AI gets harder to regulate over time. As more actors become capable of building powerful and dangerous AI, oversight will need to be more far-reaching to effectively manage the negative externalities of AI systems. Because of the ratcheting effect of powerful AI, safety policies must prevent the creation of dangerous systems — once they exist or are released publicly, it is much more difficult to prevent their proliferation.

The reactive framework fails to address either of these problems because it tacitly endorses AI proliferation until the cusp when AI is extremely dangerous (and even then, assumes it will catch where and when this cusp is). If powerful AI technology has widely proliferated by the time a risk is detected, it will be vastly more difficult to maintain safety and international stability than if AI is controlled proactively.

### 3. The reactive framework incorrectly assumes that an AI “warning shot” will motivate coordination.

Imagine an extreme situation in which an AI disaster serves as a “warning shot” for humanity. This would imply that powerful AI has been developed and that we have months (or less) to develop safety measures or pause further development. After a certain point, an actor with sufficiently advanced AI may be ungovernable, and misaligned AI may be uncontrollable.

When horrible things happen, people do not suddenly become rational. In the face of an AI disaster, we should expect chaos, adversariality, and fear to be the norm, making coordination very difficult. The useful time to facilitate coordination is before disaster strikes.

However, the reactive framework assumes that this is essentially how we will build consensus in order to regulate AI. The optimistic case is that we hit a dangerous threshold before a real AI disaster, alerting humanity to the risks. But history shows that it is exactly in such moments that these thresholds are most contested – this shifting of the goalposts is known as the [AI Effect](#) and common enough to have its own Wikipedia page. Time and again, AI advancements have been explained away as routine processes, whereas “real AI” is redefined to be some mystical threshold we have not yet reached. Dangerous capabilities are similarly contested as they arise, such as how recent reports of [OpenAI’s o1 being deceptive](#) have [been questioned](#).

This will become increasingly common as competitors build increasingly powerful capabilities and approach their goal of building AGI. **Universally, powerful stakeholders fight for their narrow interests, and for maintaining the status quo, and they often win, even when all of society is going to lose.** Big Tobacco didn’t pause cigarette-making when they learned about lung cancer; instead they spread misinformation and hired lobbyists. Big Oil didn’t pause drilling when they learned about climate change; instead they spread misinformation and hired lobbyists. Likewise, now that billions of dollars are pouring into the creation of AGI and superintelligence, we’ve already [seen competitors](#) fight tooth and nail to keep

building. If problems arise in the future, of course they will fight for their narrow interests, just as industries always do. And as the AI industry gets larger, more entrenched, and more essential over time, this problem will grow rapidly worse.

#### 4. The reactive framework doesn't put humanity in control of AI development.

Head of Safety at the US AISI Paul Christiano [acknowledges](#) that even very good reactive approaches inadequately address superintelligence risks, but argues that they are net beneficial and useful stepping stones to create other effective regulation:

*I think the risk from rapid AI development is very large, and that even very good RSPs would not completely eliminate that risk. A durable, global, effectively enforced, and hardware-inclusive pause on frontier AI development would reduce risk further. I think this would be politically and practically challenging and would have major costs, so I don't want it to be the only option on the table. I think implementing RSPs can get most of the benefit, is desirable according to a broader set of perspectives and beliefs, and helps facilitate other effective regulation.*

But what is this “other effective regulation” and what does it lead to? Even if the reactive framework effectively manages to catch risks when they arise, it does not give humanity the necessary control of AI development in general, building the institutions necessary to be proactive. As AI becomes more powerful, what we need is not a whack-a-mole strategy, but a cautious, proactive approach to only develop AI we can guarantee is safe.

The fact that the reactive framework inherently endorses the status quo should be a cause for concern, especially because many of its proponents acknowledge the risks of superintelligence. This is by design.

### Current AI policy efforts endorse the race to AGI

The actors helping define policy strategies are AGI companies themselves, many of whom are lobbying heavily against regulation while promoting their own safety frameworks. Coordination to prevent superintelligence risks is effectively impossible so long as the faction trying to build it controls the AI policy landscape.

Over the past few months, AI discourse has had an increasingly nationalistic timbre. Consider Sam Altman's recent op-ed, "[Who will control the future of AI?](#)"

*That is the urgent question of our time. The rapid progress being made on artificial intelligence means that we face a strategic choice about what kind of world we are going to live in: Will it be one in which the United States and allied nations advance a global AI that spreads the technology's benefits and opens access to it, or an authoritarian one, in which nations or movements that don't share our values use AI to cement and expand their power?*

Here's famous venture capitalist [Vinod Khosla](#):

*PESSIMISTS PAINT A DYSTOPIAN FUTURE IN TWO PARTS—ECONOMIC AND SOCIAL. They fear widespread job loss, economic inequality, social manipulation, erosion of human agency, loss of creativity, and even existential threats from AI. I believe these fears are largely unfounded, myopic, and harmful. They are addressable through societal choices. MOREOVER, THE REAL RISK ISN'T "SENTIENT AI" BUT LOSING THE AI RACE TO NEFARIOUS "NATION STATES," OR OTHER BAD ACTORS, MAKING AI DANGEROUS FOR THE WEST. IRONICALLY, THOSE WHO FEAR AI AND ITS CAPACITY TO ERODE DEMOCRACY AND MANIPULATE SOCIETIES SHOULD BE MOST FEARFUL OF THIS RISK!*

And Dario Amodei's explanation of the "[entente strategy](#)," revealing that the reactive framework is not about a delicate strategic path to creating safeguards, but greenlighting the current race to AGI:

*On the international side, it seems very important that democracies have the upper hand on the world stage when powerful AI is created. AI-powered authoritarianism seems too terrible to contemplate, so democracies need to be able to set the terms by which powerful AI is brought into the world, both to avoid being overpowered by authoritarians and to prevent human rights abuses within authoritarian countries.*

*My current guess at the best way to do this is via an "entente strategy", in which a coalition of democracies seeks to gain a clear advantage (even just a temporary one) on powerful AI by securing its supply chain, scaling quickly, and blocking or delaying adversaries' access to key resources like chips and semiconductor equipment. This coalition would on one hand use AI to achieve robust military superiority (the stick) while at the same time offering to distribute the benefits of powerful AI (the carrot) to a wider and wider group of countries in exchange for supporting the coalition's strategy to promote democracy (this would be a bit analogous to "Atoms for Peace"). The coalition would aim to gain the support of more and more of the world, isolating our worst adversaries and eventually putting them in a position where they are better off taking the same bargain as the rest of the world: give up competing with democracies in order to receive all the benefits and not fight a superior foe.*

*If we can do all this, we will have a world in which democracies lead on the world stage and have the economic and military strength to avoid being undermined, conquered, or sabotaged by autocracies, and may be able to parlay their AI superiority into a durable advantage. This could optimistically lead to an "eternal 1991"—a world where democracies have the upper hand and Fukuyama's dreams are realized. Again, this will be very difficult to*

*achieve, and will in particular require close cooperation between private AI companies and democratic governments, as well as extraordinarily wise decisions about the balance between carrot and stick.*

By arguing that the race to AGI is a race for “democracy to prevail over authoritarianism,” Sam Altman and Dario Amodei essentially advocate for acceleration, publicly calling for the US to “use AI to achieve robust military superiority.”

This perspective had led to AGI and hyperscaling compute companies to [petition the government to accelerate AI development](#):

*Today, as part of the Biden-Harris Administration’s comprehensive strategy for responsible innovation, the White House convened leaders from hyperscalers, artificial intelligence (AI) companies, datacenter operators, and utility companies to discuss steps to ensure the United States continues to lead the world in AI. Participants considered strategies to meet clean energy, permitting, and workforce requirements for developing large-scale AI datacenters and power infrastructure needed for advanced AI operations in the United States.*

Just one month later, this acceleration led to the US government collaborating with those same companies to [harness the power of AI for national security](#):

*The National Security Memorandum (NSM) is designed to galvanize federal government adoption of AI to advance the national security mission, including by ensuring that such adoption reflects democratic values and protects human rights, civil rights, civil liberties and privacy. In addition, the NSM seeks to shape international norms around AI use to reflect those same democratic values, and directs actions to track and counter adversary development and use of AI for national security purposes.*

The balance of power between the US and China is a legitimate issue, but the government should not use it to justify racing to AGI and superintelligence. The fact that they are is a narrative fed to them by AGI companies themselves, who urge governments for funding and support behind the scenes.

Stoking nation-state race dynamics undermines the precarious balance of international coordination, and could kill us all: Without having solved alignment, deploying powerful AI is effectively suicide. The hard questions of safety and alignment will not just go away on their own, and get even harder to solve in a geopolitically unstable world.

Even if misaligned AGI is deployed by a “trusted” government, it is globally catastrophic, the equivalent of “winning a nuclear war that leaves the planet uninhabitable.” There are no winners in this situation.

The only way to guarantee safety is for humanity to be in collective control of frontier AI development. This is a global solution which requires far-reaching monitoring regimes to prevent countries, companies, or rogue actors from pursuing building AGI and leading to the catastrophic outcomes above.

As international [AI Safety Institutes plan for an upcoming conference](#) to make progress on Frontier AI Safety Frameworks, leading AGI actors will be in the room. These actors have made it transparent that their interest is in building AGI, arguing for voluntary commitments which do nothing to stop the race, pushing against regulation behind the scenes, and now invoking nationalistic language in order to defend their race.

## (6) The AI Race

### The race to AGI is ideological, and will drive us to the exact dangers it claims to avoid

The current situation of AI development is paradoxical:

- Humanity is on a default path toward increasingly powerful AI that risks causing our extinction, which [key actors acknowledge](#).
- Those same major actors are racing towards this future with abandon; expert discourse is filled with pseudoscientific reassurances that discount the core issues; policy approaches fail to bind any of these actors to a safe future; and safety efforts assume that we will “muddle through” without putting in the necessary effort.

To make sense of this apparent contradiction, we must take a closer look at exactly how the AGI race is unfolding: Who are the participants? What are their motivations? What dynamics emerge from their beliefs and interactions? How does this translate to concrete actions, and how to make sense of these actions? Only by doing this can we understand where the AGI race is headed.

In [The AGI race is ideologically driven](#), we sort the key actors by their ideology, explaining that the history of the race to AGI is rooted in “singularitarian” beliefs about using superintelligence to control the future.

In [These ideologies shape the playing field](#), we argue that the belief that whoever controls AGI controls the future leads to fear that the “wrong people” will end up building AGI first, leading to a race dominated by actors who are willing to neglect any risk or issue that might slow them down.

In [The strategies being used to justify and perpetuate the race to AGI are not new](#), we argue that in order to reach their goals of building AGI first, the AGI companies and their allies are simply applying [the usual industry playbook](#) strategies from Tobacco, Oil, Big Tech, to create confusion and fear they then use to wrestle control of the policy and scientific establishment. That way, there is nothing in their way as they race with abandon to AGI.

In [How will this go?](#), we argue that the race, driven by ideologies that want to build AG, will continue. There is no reason for any of the actors to change their tack, especially given their current success.

## The AGI race is ideologically driven

Studying the motivations of the participants clarifies their interactions and the dynamics of the race.

For the AGI race, the actors fit into five groups:<sup>18</sup>

- Utopists, who are the main drivers of the race, and want to build AGI in order to control the future and usher in the utopia they want.
- Big Tech, who started by supporting the utopists, are now in the process of absorbing and consuming them to keep their technological monopolies.
- Accelerationists, who want to accelerate and deregulate technological progress because they think it is an unmitigated good.
- Zealots, who want to build AGI and superintelligence because they believe it's the superior species that should control the future.
- Opportunists, who just follow the hype without having any strong belief about it.

### Utopists: Building AGI to usher in utopia

This group is the main driver of the AGI race, and are actively pushing for development in order to build their vision of utopia. AGI companies include [DeepMind](#), [OpenAI](#), [Anthropic](#), and [xAI](#) (and Meta, though we class them more as accelerationists below). A second, overlapping cluster of utopists support the “entente strategy,” including influential members of the philanthropic foundation [Open Philanthropy](#) and think tank [RAND](#), as well as leadership from AGI companies, like Sam Altman and Dario Amodei.

The ideology that binds these actors together is **the belief that AGI promises absolute power over the future to whoever builds it, and the desire to wield that power so they can usher in their favorite flavor of utopia.**

### AGI companies

Companies usually come together in order to make profit: building a new technology is often a means to make money, not the goal itself. But in the case of AGI, the opposite is true. AGI companies have been created explicitly to build AGI, and use products and money as a way to further this goal. It is clear from their histories and from their explicit statements about why AGI matters that this is the case.

---

<sup>18</sup> These groups are not perfectly mutually exclusive, and there is some ambiguity and overlap between organizations and individuals of different groups.

All AGI companies have their roots in Singularitarianism, a turn-of-the-century online movement which focused on building AGI and reaping the benefits.

Before singularitarians, AGI and its potential power and implications were mostly discussed in papers by scientific luminaries:

- [Alan Turing warned](#) in 1951 that  
“once the machine thinking method had started, it would not take long to outstrip our feeble powers... at some stage therefore we should have to expect the machines to take control.”
- Later in the 1950’s [John von Neumann discussed](#) that  
“the ever accelerating progress of technology... give the appearance of approaching some essential singularity in the history of the race beyond which human affairs, as we know them, could not continue.”
- In the 1960s, mathematician [I. J. Good introduced the idea of AI "intelligence explosion"](#) in which advanced AI improves itself to superhuman levels, similar to what is described in Section 3 on AI Catastrophe.

But in the year 2000, the term “[Singularitarianism](#)” was coined to describe the ideology of individuals who took these ideas *seriously* – not as science fiction, but as the secular belief that superintelligence will likely be built in the future, and that deliberate action ought to be taken to ensure this benefits humans.<sup>19</sup> It was in these individuals where the ideology behind the current race for AGI found its beginnings. Eliezer Yudkowsky, an early representative Singularitarians, writes in the (deleted) [Singularitarian Principles](#) that:

*Singularitarians are the partisans of the Singularity.*

*A Singularitarian is someone who believes that technologically creating a greater-than-human intelligence is desirable, and who works to that end.*

*A Singularitarian is advocate, agent, defender, and friend of the future known as the Singularity.*

In one way or another, each of the utopists emerged from the Singularitarians.

DeepMind was founded by two early Singularitarians, Demis Hassabis and Shane Legg, who met their first investor, Peter Thiel, through a party organized by Eliezer Yudkowsky’s

---

<sup>19</sup> A deeper history of the transhumanist movement can be found [here](#).

Singularity Institute.<sup>20</sup> In book [Genius Makers](#), Cade Metz tracks the history of the early AGI race:

*In recent years, Legg had joined an annual gathering of futurists called the Singularity Summit. “The Singularity” is the (theoretical) moment when technology improves to the point where it can no longer be controlled by the human race. [...] One of the founders was a self-educated philosopher and self-described artificial intelligence researcher named Eliezer Yudkowsky, who had introduced Legg to the idea of superintelligence in the early 2000s when they were working with a New York-based start-up called Intelligensis. But Hassabis and Legg had their eyes on one of the other [Singularity Summit] conference founders: Peter Thiel.*

*In the summer of 2010, Hassabis and Legg arranged to address the Singularity Summit, knowing that each speaker would be invited to a private party at Thiel’s town house in San Francisco.*

Demis Hassabis soon converted Elon Musk to the potential and dangers of AGI and secured additional funding from him, by highlighting how Musk’s dream of colonizing Mars could be jeopardized by AGI. The New York Times [reports](#) that:

*Mr. Musk explained that his plan was to colonize Mars to escape overpopulation and other dangers on Earth. Dr. Hassabis replied that the plan would work — so long as superintelligent machines didn’t follow and destroy humanity on Mars, too.*

*Mr. Musk was speechless. He hadn’t thought about that particular danger. Mr. Musk soon invested in DeepMind alongside Mr. Thiel so he could be closer to the creation of this technology.*

In 2015, Google acquired DeepMind, which liquidated Elon Musk from the company. Musk’s disagreements with Larry Page over building AGI for “religious” reasons (discussed further in the zealots subsection) pushed Musk to join forces with Sam Altman and launch OpenAI, as discussed in [released emails](#).

In 2021, history repeated itself: Anthropic was founded to compete with OpenAI in response to the latter’s deal with Microsoft. The founders of Anthropic included brother and sister Dario and Daniela Amodei. Dario was one of the researchers invited to the dinner that led to the founding of OpenAI, as noted by co-founder Greg Brockman in a since-deleted [blog post](#):

---

<sup>20</sup> Eliezer himself corroborates this history, [noting](#) “I obviously did not forecast the consequences correctly.” To his credit, while many of these early singularitarians have gone on to lead the AGI race, Eliezer pivoted and went directly into working on AI alignment and writing about the risks from superintelligence.

*About a month later, Sam [Altman] set up a dinner in Menlo Park. On the list were Dario [Amodei], Chris [Olah], Paul [Christiano], Ilya Sutskever, Elon Musk, Sam, and a few others.<sup>21</sup>*

And in two blog [posts](#) entitled “Machine Intelligence”, Sam Altman thanks Dario Amodei specifically for helping him come to grips with questions related to AGI:

*Thanks to Dario Amodei (especially Dario), Paul Buchheit, Matt Bush, Patrick Collison, Holden Karnofsky, Luke Muehlhauser, and Geoff Ralston for reading drafts of this and the previous post.*

Elon Musk re-entered the race in 2023, founding xAI, with the [stated goal](#) of “trying to understand the universe.”

This ideology of building AGI to usher the Singularity and Utopia was thus foundational for all these companies. In addition, the leaders of these companies have also stated that they believe in the extreme importance to the future of AGI, and who builds it:

- DeepMind co-founder Shane Legg writes [in his PhD thesis](#), “If our intelligence were to be significantly surpassed, it is difficult to imagine what the consequences of this might be. It would certainly be a source of enormous power, and with enormous power comes enormous responsibility.”
- DeepMind co-founder Demis Hassabis [is quoted by The Guardian](#) as saying that “he is on a mission to “solve intelligence, and then use that to solve everything else”.
- OpenAI’s co-founder and CEO Sam Altman wrote in [a public OpenAI plan](#) that “Successfully transitioning to a world with superintelligence is perhaps the most important—and hopeful, and scary—project in human history.”
- OpenAI co-founder Greg Brockman agrees, [writing in a deleted blog post](#) that “there was one problem that I could imagine happily working on for the rest of my life: moving humanity to safe human-level AI. It’s hard to imagine anything more amazing and positively impactful than successfully creating AI, so long as it’s done in a good way.”
- Anthropic co-founder and CEO Dario Amodei stated in [a recent blog post](#) that building AGI “[building AGI] is a world worth fighting for. If all of this really does happen over 5 to 10 years—the defeat of most diseases, the growth in biological and cognitive freedom, the

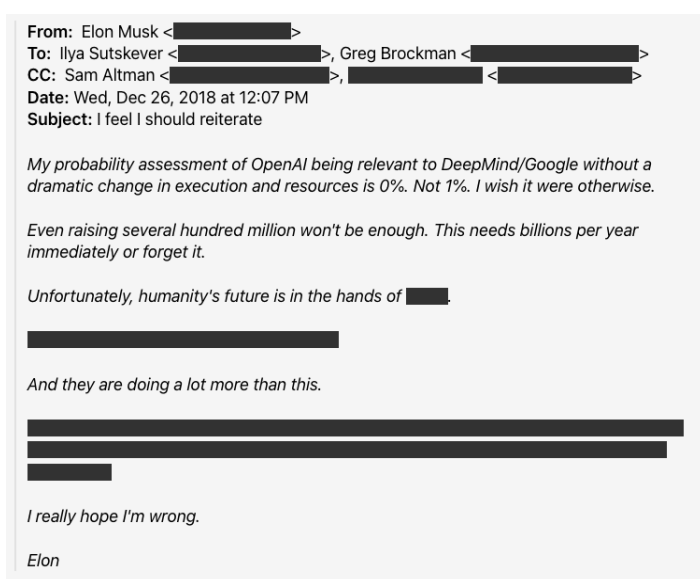
---

<sup>21</sup> Note that these individuals all continue to have prominent influence over the race to AGI and the field of AI safety.

lifting of billions of people out of poverty to share in the new technologies, a renaissance of liberal democracy and human rights—I suspect everyone watching it will be surprised by the effect it has on them.”

- This commitment is reiterated in Anthropic’s [Core Views on AI Safety](#), positing that “most or all knowledge work may be automatable in the not-too-distant future – this will have profound implications for society, and will also likely change the rate of progress of other technologies as well (an early example of this is how systems like AlphaFold are already speeding up biology today). ...it is hard to overstate what a pivotal moment this could be.”

An [email to OpenAI founders](#) from Musk most ominously and succinctly summarizes the utopist position: “humanity’s future” is in the hands of whoever wins the race to AGI:



## Entente

Aside from racing to build AGI, utopists have also started to tie their own goals (ushering what they see as a good future) with the stated and ideal aims of existing cultures and governments, notably democracy and the US government.

This is true for AGI companies, for example with OpenAI CEO Sam Alman writing in [a recent op-ed](#) that:

*There is no third option — and it's time to decide which path to take. The United States currently has a lead in AI development, but continued leadership is far from guaranteed. Authoritarian governments the world over are willing to spend enormous amounts of money to catch up and ultimately overtake us. Russian dictator Vladimir Putin has darkly warned that the country that wins the AI race will “become the ruler of the world,” and the People’s Republic of China has said that it aims to become the global leader in AI by 2030.*

Ex-OpenAI superalignment researcher Leopold Aschenbrenner echoes this sentiment in [Situational Awareness](#):

*Every month of lead will matter for safety too. We face the greatest risks if we are locked in a tight race, democratic allies and authoritarian competitors each racing through the already precarious intelligence explosion at breakneck pace—forced to throw any caution by the wayside, fearing the other getting superintelligence first. **Only if we preserve a healthy lead of democratic allies will we have the margin of error for navigating the extraordinarily volatile and dangerous period around the emergence of superintelligence.** And only American leadership is a realistic path to developing a nonproliferation regime to avert the risks of self-destruction superintelligence will unfold.*

Anthropic CEO Dario Amodei provides the clearest description of this approach, under the name [“entente strategy”](#):

*My current guess at the best way to do this is via an “entente strategy”, in which a coalition of democracies seeks to gain a clear advantage (even just a temporary one) on powerful AI by securing its supply chain, scaling quickly, and blocking or delaying adversaries’ access to key resources like chips and semiconductor equipment. This coalition would on one hand use AI to achieve robust military superiority (the stick) while at the same time offering to distribute the benefits of powerful AI (the carrot) to a wider and wider group of countries in exchange for supporting the coalition’s strategy to promote democracy (this would be a bit analogous to “Atoms for Peace”). The coalition would aim to gain the support of more and more of the world, isolating our worst adversaries and eventually putting them in a position where they are better off taking the same bargain as the rest of the world: give up competing with democracies in order to receive all the benefits and not fight a superior foe.*

*If we can do all this, we will have a world in which democracies lead on the world stage and have the economic and military strength to avoid being undermined, conquered, or sabotaged by autocracies, and may be able to parlay their AI superiority into a durable advantage. This could optimistically lead to an “eternal 1991”—a world where democracies have the upper hand and Fukuyama’s dreams are realized.*

That is, the democracies leading the AGI race (mostly the US) need to rush ahead in order to control the future, leading to an “eternal 1991.”

The more communication-and-policy focused entente strategy has also involved different utopists than AGI companies. Amodei also explicitly credits the think tank [RAND](#) with the name “entente strategy” and the rough idea:

This is the title of a forthcoming paper from RAND, that lays out roughly the strategy I describe.

[RAND](#), an influential think tank created at the end of WWII, [has been heavily involved with the Biden administration to draft executive orders on AI](#). Yet it clearly predates the singularitarians of the early 2000s – AI and AGI were not part of its founding goals; instead, RAND took up these topics after [Jason Matheny](#) was [appointed CEO in 2022](#).

Matheny is tied with the [effective altruism](#) (EA), a movement that combined traditional philanthropy with various causes, from animal rights to risks from AI, and grew in part out of the rationalist community that had originated from Yudkowsky’s Singularitarians. Indeed, RAND has been bankrolled since 2022 by [Open Philanthropy](#), the main non-profit funding effective altruism causes, offering a total of \$18.5 millions to RAND for potential risks from AIs<sup>22</sup>, all to be “spent at RAND President Jason Matheny’s discretion.”

This is not the first time Open Philanthropy directly supported utopist actors: In 2017, Open Philanthropy also [recommended](#) a grant of \$30 million to OpenAI, arguing that

it’s fairly likely that OpenAI will be an extraordinarily important organization, with far more influence over how things play out than organizations that focus exclusively on risk reduction and do not advance the state of the art.

In [the “relationship disclosure” section](#), Open Philanthropy nods to the conflict of interest:

*OpenAI researchers Dario Amodei and Paul Christiano are both technical advisors to Open Philanthropy and live in the same house as Holden. In addition, Holden is engaged to Dario’s sister Daniela.*

“Holden” is [Holden Karnofsky](#), who was Open Philanthropy’s former executive director and former Director of AI Strategy. That is, the then-director of Open Philanthropy lived with the future CEO of Anthropic, and married the future president of Anthropic. Although Open Philanthropy hasn’t funded Anthropic directly, [the \\$500 millions invested by Sam Bankman-Fried](#) before his downfall [were influenced by EA](#)<sup>23</sup>, for which Open Philanthropy was and is the most important and powerful organization in that movement.

In conclusion, for the last 10 years, what has motivated both leading AGI companies and the most powerful non-profits and NGOs working on AI-risks has been the ideology AGI will usher in utopia, and offer massive control of the future to whoever wields it.

---

<sup>22</sup> Three public grants from Open Philanthropy note donations over a short period of time ([here](#), [here](#), and [here](#))

<sup>23</sup> This donation was also advised by Leopold Aschenbrenner, author of “Situational Awareness,” who was working at FTX Future Fund at the time.

Now, the ideology of "AI will grant unlimited power" has taken a nationalistic turn, and started to find its footing in the entente strategy. While entente may look like a new ideology, it's really the same utopists coming up with a slightly different story to justify the race to AGI.

## Big Tech: Keeping a hand on the technological frontier

The group in the AGI race with the most resources are the Big Tech companies directly investing in the AGI companies and striving for a monopoly of this new technology. These include [Microsoft](#), [Google](#), [Amazon](#), [Elon Musk's empire](#), [Meta](#) and [Apple](#).

Big Tech companies are used to monopolies and controlling the technological frontier, having dominated the internet, mobile, and cloud computing markets; they're doing the same with AI, partnering and investing in the utopists' smaller companies and making them dependent on their extensive resources.

Each of the utopists is backed by at least one of the Big Tech actors:

- DeepMind was [acquired by Google](#);
- OpenAI is enabled by [a cumulative \\$13 billions of investment from Microsoft](#) and uses their infrastructure;
- Anthropic received [\\$4 billion of investment from Amazon](#), as well as [\\$2 billion from Google](#);
- xAI is enabled by the investments of Elon Musk, who controls the X empire of Tesla, SpaceX, etc.

Big Tech companies play a unique role in the AGI race, bankrolling the utopists and exploiting their progress. They were not founded in order to build AGI, and did not start the AGI race – yet the traction towards AGI eventually got their attention.

For example, Microsoft leadership invested and partnered with OpenAI because it was smelling AI progress passing by, and was [unsatisfied](#) with its internal AI teams. Kevin Scott, Microsoft's CTO, wrote to CEO Satya Nadella and founder Bill Gates in 2019:

*We have very smart ML people in Bing, in the vision team, and in the speech team. But the core deep learning teams within each of these bigger teams are very small, and their ambitions have also been constrained, which means that even as we start to feed them resources, they still have to go through a learning process to scale up. And we are multiple years behind the competition in terms of ML scale.*

And Apple, which is historically prone to building in-house rather than buying, recently announced [a deal with OpenAI to use ChatGPT](#) in order to power Siri.

Yet as usual in this kind of bargain, Big Tech is starting to reestablish their power and monopoly after depending for a while on the utopists. Not only are they the only one with the necessary resources to enable further scaling, but they also have had access to the technology of utopists and sometimes their teams:

- Google [literally acquired DeepMind](#), and [has merged it with their previous internal Google Brain](#), with final control over their research and results.
- Microsoft has intellectual property rights to OpenAI code, and the new CEO of Microsoft AI, ex-DeepMind co-founder Mustafa Suleyman, has been [reportedly](#) studying the algorithms.
- Amazon, despite its partnership with Anthropic, has been [assembling a massive internal AI team](#).
- Meta, which never sponsored any utopists, has been consistently building and releasing the most powerful open-weight LLMs, [the Llama family of models](#), for years now.

The utopists are enabled by the compute, scale, funding, and lobbying capacity of their Big Tech backers, who end up dodging public attention in the race to AGI while being one of the main driving forces.

## Accelerationists: Idolizing technological progress

Accelerationists believe that technology is an unmitigated good and that we must pursue the change it will bring about as aggressively as possible, eliminating any impediments or regulations. Accelerationists include many VCs, AGI company leaders, and software engineers; key players include [Meta](#) and venture capital fund Andreessen Horowitz ([a16z](#)).

A representative ode to accelerationism is [The Techno-Optimist Manifesto](#) by A16z co-founder Marc Andreessen. It opens by drumming up fervor that technology is what will save society, and critics of technology are what will damn it:

*We are being lied to. / We are told that technology takes our jobs, reduces our wages, increases inequality, threatens our health, ruins the environment, degrades our society, corrupts our children, impairs our humanity, threatens our future, and is ever on the verge of ruining everything. / We are told to be angry, bitter, and resentful about technology. / We are told to be pessimistic. / The myth of Prometheus – in various updated forms like Frankenstein, Oppenheimer, and Terminator – haunts our nightmares. / We are told to denounce our birthright – our intelligence, our control over nature, our ability to build a better world. / We are told to be miserable about the future...*

*Our civilization was built on technology. / Our civilization is built on technology. / Technology is the glory of human ambition and achievement, the spearhead of progress, and the realization of our potential. / For hundreds of years, we properly glorified this – until recently.*

*I am here to bring the good news.*

*We can advance to a far superior way of living, and of being. / We have the tools, the systems, the ideas. / We have the will. / It is time, once again, to raise the technology flag. / It is time to be Techno-Optimists.*

While Andreessen is particularly dramatic with his language, accelerationism generally takes the form of pushing for libertarian futures in which risks from technology are managed purely by market forces rather than government regulation. As such, it's particularly adamant about pushing for open-weight AIs, bringing dangerous technology to every possible criminal, terrorist and dictator in order to “ensure things go well”.

A common argument by accelerationists is that open-source makes software safer by having more eyes explore and inspect the source-code. See [Andreessen's recent post co-authored with Microsoft leadership](#):

*[Open-Source AIs] also offer the promise of safety and security benefits, since they can be more widely scrutinized for vulnerabilities.*

Yet this is profoundly misleading. First, as [we articulated before](#), the AIs released by the likes of Meta and Mistral AI are not open-source: they don't include everything necessary to train the model, such as the data and the training algorithm. This makes them open-weight, and thus much more opaque to the external user than they are to the internal developer. And even worse than that, recall that [modern AIs are grown, not built](#); this means that even the very researchers who have created these AIs don't know how they work, and how to fix them if there is a problem. So the main value of open-source software, of having many more eyeballs on the source code to find bugs and security exploits, vanishes because no one who looks at the numbers in a modern AI understands what they mean.

Another argument is simply the proposition that giving everyone a dangerous technology somehow makes it easier for governments and law-enforcement to control it. Indeed, Meta's CEO Mark Zuckerberg makes this very claim in [his defense of his company's open-weight approach](#):

*I think it will be better to live in a world where AI is widely deployed so that larger actors can check the power of smaller bad actors.*

And as we [discussed above](#), this argument fails to address the asymmetric nature of the danger: it only takes one bad actor succeeding to create chaos, which forces governments and law-enforcement to catch every single threat, without making a single mistake or error. Just because accelerationists decided to give open-weights AI to criminals, terrorists, and rogue nations.

In the end, the goal of accelerationists is simply to avoid regulations on technology at all cost, since they see technology as such an unmitigated good. These beliefs leave no room for a calm and collected reflection of the potential risks of technology, and how to handle them properly.

## Zealots: Worshiping superintelligence

The zealots believe AGI to be a superior successor to humanity that is akin to a god; they don't want to build or control it themselves, but they do want it to arrive, even if humanity is dominated, destroyed, or replaced by it.

Larry Page, co-founder of Google and key advocate of the DeepMind acquisition, is one such zealot. Elon Musk's biographer [detailed](#) Page and Musk's conflict at a dinner that broke up their friendship (emphasis ours):

*Musk argued that unless we built in safeguards, artificial intelligence systems might replace humans, making our species irrelevant or even extinct.*

**Page pushed back. Why would it matter, he asked, if machines someday surpassed humans in intelligence, even consciousness? It would simply be the next stage of evolution.**

*Human consciousness, Musk retorted, was a precious flicker of light in the universe, and we should not let it be extinguished. Page considered that sentimental nonsense. If consciousness could be replicated in a machine, why would that not be just as valuable? Perhaps we might even be able someday to upload our own consciousness into a machine. He accused Musk of being a “specist,” someone who was biased in favor of their own species. “Well, yes, I am pro-human,” Musk responded. “I fucking like humanity, dude.”*

Another is [Richard Sutton](#), one of the fathers of modern Reinforcement Learning, who articulated his own zealotry in his [AI Succession](#) presentation:

*We should not resist succession, but embrace and prepare for it / Why would we want greater beings kept subservient? / Why don't we rejoice in their greatness as a symbol and extension of humanity's greatness, and work together toward a greater and inclusive civilization?*

Another father of AI, Jurgen Schmidhuber, believes that AI will [inevitably become more intelligent than humanity](#):

*In the year 2050 time won't stop, but we will have AIs who are more intelligent than we are and will see little point in getting stuck to our bit of the biosphere. They will want to move history to the next level and march out to where the resources are. In a couple of million years, they will have colonised the Milky Way."*

Despite this, he disregards the risks, assuming that this all will happen while humanity plods along. The Guardian [reports](#): "Schmidhuber believes AI will advance to the point where it surpasses human intelligence and has no interest in humans – while humans will continue to benefit and use the tools developed by AI."

Whether we are replaced by successor species or simply hand over the future to them, zealots believe the coming AI takeover is inevitable and humanity should step out of the way. While the zealots represent a minority in the AGI race, they've had massive influence: Page and Musk's conflict led to Google's acquisition of DeepMind and thus the creation of OpenAI and, and arguably catalyzed the creation of Anthropic and xAI.

## Opportunists: Following the hype

This last group comprises everyone who joined the AGI race and AI industry not because of a particular ideology, but in order to ride the wave of hype and investment, and get money, status, power from it.

This heterogeneous group includes smaller actors in the AI space, such as [Mistral AI](#) and [Cohere](#); hardware manufacturers, such as [Nvidia](#) and [AMD](#); startups riding the wave of AI progress, such as [Perplexity](#) and [Cursor](#); established tech companies integrating AIs into their products, such as [Zoom](#) and [Zapier](#); older companies trying to stay relevant to the AI world, such as [Cisco](#) and [IBM](#).

Most of the core progress towards AGI has been driven by utopists and Big Tech, not opportunists following the hype. Nonetheless, their crowding into the market has fueled the recent AI boom, and built the infrastructure on which it takes place.

## These ideologies shape the playing field

Motivated by their convictions, the utopists are setting the rapid pace of the AGI race despite publicly acknowledging the risks.

Because they believe that AGI is possible and that building it leads to control of the future, they must be the first to protect humanity from the “wrong people” beating them to the punch. Utopists worry about AGI being built by someone naive or incompetent, inadvertently triggering catastrophe; per Ashlee Vance’s [biography of Musk](#), this was Musk’s underlying concern in his disagreement with Larry Page:

*He opened up about the major fear keeping him up at night: namely that Google’s co-founder and CEO Larry Page might well have been building a fleet of artificial-intelligence-enhanced robots capable of destroying mankind. “I’m really worried about this,” Musk said. It didn’t make Musk feel any better that he and Page were very close friends and that he felt Page was fundamentally a well-intentioned person and not Dr. Evil. In fact, that was sort of the problem. Page’s nice-guy nature left him assuming that the machines would forever do our bidding. “I’m not as optimistic,” Musk said. “He could produce something evil by accident.”*

Utopists also consider anyone with opposing values to be the wrong people. This includes foreign adversaries such as China and Russia. Sam Altman writes in [a recent op-ed](#) that:

*There is no third option — and it’s time to decide which path to take. The United States currently has a lead in AI development, but continued leadership is far from guaranteed. Authoritarian governments the world over are willing to spend enormous amounts of money to catch up and ultimately overtake us. Russian dictator Vladimir Putin has darkly warned that the country that wins the AI race will “become the ruler of the world,” and the People’s Republic of China has said that it aims to become the global leader in AI by 2030.*

This is also the main thrust of the “entente strategy”: democracies must race to AGI lest they be overtaken by dictatorships.

This dynamic creates an arms race, where staying in the lead matters more than everything else. This has been a recurring motif throughout the AGI race: OpenAI was built because Elon Musk wanted to overtake DeepMind after Google’s acquisition; Anthropic wanted to outdo OpenAI after the latter made a deal with Microsoft; and xAI is yet another attempt by Elon Musk to get back in the race.

This is why those at the forefront of the AGI race (mainly OpenAI and Anthropic) end up being the people most willing to throw away safety in order to make a move. As a telling example, Anthropic [released Claude](#), which they proudly (and correctly) described as pushing the state-of-the-art, contradicting [their own Core Views on AI Safety](#), which promised “We generally don’t publish this kind of work because we do not wish to advance the rate of AI capabilities progress.”

Over time, anyone willing to slow for safety is weeded out, like OpenAI leaders Jan Leike and Ilya Sutskever from OpenAI's Superalignment Team. Jan Leike [explained his quitting](#) by complaining that he didn't have enough support and resources for doing his work on safety:

*Over the past few months my team has been sailing against the wind. Sometimes we were struggling for compute and it was getting harder and harder to get this crucial research done.*

*[...]*

*But over the past years, safety culture and processes have taken a backseat to shiny products."*

The tendency to rationalize away the risks is best captured by [Situational Awareness](#), a report written by an ex-OpenAI employee, Leopold Aschenbrenner. He articulates his concern about future risks and the inadequate mediation measures:

*We're not on track for a sane chain of command to make any of these insanely high-stakes decisions, to insist on the very-high-confidence appropriate for superintelligence, to make the hard decisions to take extra time before launching the next training run to get safety right or dedicate a large majority of compute to alignment research, to recognize danger ahead and avert it rather than crashing right into it. Right now, no lab has demonstrated much of a willingness to make any costly tradeoffs to get safety right (we get lots of safety committees, yes, but those are pretty meaningless). By default, we'll probably stumble into the intelligence explosion and have gone through a few OOMs before people even realize what we've gotten into.*

*We're counting way too much on luck here.*

On the very next page, he then contends that (emphasis ours):

*Every month of lead will matter for safety too. We face the greatest risks if we are locked in a tight race, democratic allies and authoritarian competitors each racing through the already precarious intelligence explosion at breakneck pace—forced to throw any caution by the wayside, fearing the other getting superintelligence first. **Only if we preserve a healthy lead of democratic allies will we have the margin of error for navigating the extraordinarily volatile and dangerous period around the emergence of superintelligence.** And only American leadership is a realistic path to developing a nonproliferation regime to avert the risks of self-destruction superintelligence will unfold.*

This is the gist of the AGI race: fear makes the very people worried about risks from AGI race even faster and deprioritize safety, arguing that they will eventually stop and do things

correctly when they finally feel safe. Ironically, this very attitude creates more competitors, more people racing for AGI, and ensures it truly is a race to the bottom..

Utopists are the core drivers of the race to AGI, but the ideology of every other group adds fuel to the fire. Big Tech is investing billions in order to capture this powerful new technology on the horizon; accelerationists spread AIs and AI research everywhere and [undermine any regulation](#), bringing about the worst nightmare of the utopists and making them race even faster; zealots praise the coming of godly AIs; and opportunists enmesh themselves with the whole race to AGI, increasing the attention, funding, infrastructure and support it gets.

## The strategies being used to justify and perpetuate the race to AGI are not new

The AGI race favors actors who are willing to ignore or downplay risk. However, leading utopists are also well aware of AI's existential risk: the CEOs of DeepMind, OpenAI, and Anthropic, as well as Dustin Moskovitz, the main funder of Open Philanthropy, signed [a statement](#) in 2023 stating that "Mitigating the risk of extinction from AI should be a global priority alongside other societal-scale risks such as pandemics and nuclear war." Elon Musk signed [an open letter](#) to pause giant AI experiments.

The above analysis clarifies what is happening: since the utopists actually want to build AGI, and that they're the ones most willing to throw safety under the bus, they are simply using the industry playbook of Big Tech, Big Oil, Big Tobacco, etc., to reach their goal, while pretending to champion safety.

### The Industry Playbook

At its core, the industry playbook is about getting obstacles out of one's way, be they competitors, the public, or the government.

The main strategy is [FUD](#), for Fear, Uncertainty and Doubt: it helps anyone who benefits from the status quo (usually "inaction"), when they need to fend off something menacing this status quo.

For example, an industry sells a product (tobacco, asbestos, social media, etc.) that harms people, and some are starting to notice and make noise about it. Denying the harms may be impossible or could draw more attention to them, so industrial actors opt to spread *confusion* instead: they accuse others of lying or being in the pockets of shadowy adversaries, drown the public in tons of junk data that is hard to evaluate, bore people with

minutia to redirect public attention, or argue that the science is nascent, controversial, or premature. In short, they waste as much time and energy of their opponents and the general public as possible.

This delays countermeasures and buys the incumbent time. And, if they're lucky, it might be enough to outlast the underfunded pro-civil adversaries as they run out of funding or public attention.

The canonical example of FUD comes from tobacco: as documented in [a history of tobacco industry tactics](#), John W. Hill, the president of the leading public relations firm at the time, published his recommendations in 1953 as experts began to understand the dangers of smoking (emphasis ours):

*So he proposed seizing and controlling science rather than avoiding it. If science posed the principal—even terminal—threat to the industry, Hill advised that the companies should now associate themselves as great supporters of science. **The companies, in his view, should embrace a sophisticated scientific discourse; they should demand more science, not less.***

*Of critical importance, Hill argued, they should declare the positive value of scientific skepticism of science itself. Knowledge, Hill understood, was hard won and uncertain, and there would always be skeptics. What better strategy than to identify, solicit, support, and amplify the views of skeptics of the causal relationship between smoking and disease? Moreover, the liberal disbursement of tobacco industry research funding to academic scientists could draw new skeptics into the fold. The goal, according to Hill, would be to build and broadcast a major scientific controversy. The public must get the message that the issue of the health effects of smoking remains an open question. **Doubt, uncertainty, and the truism that there is more to know would become the industry's collective new mantra.***

FUD is so effective the CIA recommended it to intelligence operatives in WWII – the [Simple Sabotage Field Manual](#) abounds with tactics that are meant to slow down, exhaust, and confuse, while maintaining plausible deniability:

*When possible, refer all matters to committees, for "further study and consideration." Attempt to make the committees as large as possible - never less than five.*

*Bring up irrelevant issues as frequently as possible.*

*Haggle over precise wordings of communications, minutes, resolutions.*

*Refer back to matters decided upon at the last meeting and attempt to re-open the question of the advisability of that decision.*

*Advocate "caution." Be "reasonable" and urge your fellow-conferees to be "reasonable" and avoid haste which might result in embarrassments or difficulties later on.*

FUD also encourages and exploits the industry funding of scientific research. [A recent review](#) of the influence of industry funding on research describes the practice (emphasis ours):

*Qualitative and quantitative studies included in our review suggest that **industry also used research funding as a strategy to reshape fields of research through the prioritization of topics that supported its policy and legal positions, while distracting from research that could be unfavorable.** Analysis of internal industry documents provides insight into how and why industry influenced research agendas. It is particularly interesting to note how corporations adopted similar techniques across different industry sectors (i.e., tobacco, alcohol, sugar, and mining) and fields of research. The strategies included establishing research agendas within the industry that were favorable to its positions, strategically funding research along these lines in a way that appeared scientifically credible, and disseminating these research agendas by creating collaborations with prominent institutions and researchers.*

That being said, FUD only works because the default path is to preserve the risky product; it would be ineffective if the default strategy was instead to wait for scientific consensus to declare the product safe.

To avoid this, the industry playbook encourages self-regulation. If the industry must regulate itself, then it cannot act until something bad happens, and then when it happens they actively FUD to outrun regulation forever.

Facebook has followed this playbook, repeatedly pushing against government regulation, [as demonstrated in their interactions with the European Commission in 2017](#) (emphasis ours):

In January 2017, Facebook referred only to its terms of service when explaining decisions on whether or not to remove content, the documents show. "Facebook explained that referring to the terms of services allows faster action but are open to consider changes," a Commission summary report from then reads.

"Facebook considers there are two sets of laws: private law (Facebook community standards) and public law (defined by governments)," the company told the Commission, according to Commission minutes of an April 2017 meeting.

"**Facebook discouraged regulation,**" reads a Commission memo summarizing a September 2017 meeting with the company.

The decision to press forward with the argument is unusual, said Margarida Silva, a researcher and campaigner at Corporate Europe Observatory. **“You don’t see that many companies so openly asking for self-regulation, even going to the extent of defending private law.”**

Facebook says it has taken the Commission’s concerns into account. “When people sign up to our terms of service, they commit to not sharing anything that breaks these policies, but also any content that is unlawful,” the company told POLITICO. **“When governments or law enforcement believe that something on Facebook violates their laws, even if it doesn’t violate our standards, they may contact us to restrict access to that content.”**

Unsurprisingly, this did not work, as [a recent FTC report](#) shows:

*A new Federal Trade Commission staff report that examines the data collection and use practices of major social media and video streaming services shows they engaged in vast surveillance of consumers in order to monetize their personal information while failing to adequately protect users online, especially children and teens.*

Last but not least, FUD itself, particularly fear, can help with defending self-regulation. This is the standard technique used by industries to get governments off their backs: scream that regulating them will destroy the country’s competitiveness and allow other countries (maybe even the enemy of the time, then USSR, now China) to catch up.

This is visible for example in [the unified front by big tech against the new head of F.T.C. Lina Khan](#), who vowed to curb monopolies through regulation and antitrust law.

What’s most tricky about the industry playbook in general, and FUD in particular, is that individual actions can always be made sensible and reasonable on the surface — they push for and worry about science, innovation, consumers, competitiveness, and all the other sacred words of modernity. They sound reasonable, and almost civic-minded. And this makes criticizing them even harder, because pointing at any single action fails to unearth the strategy.

Instead, the industry playbook reveals itself when we look at the whole pattern, not at one action but at all the actions together. If someone hits you once, it might be a genuine mistake; if they repeatedly hit you “by accident”, they’re obviously trying to hit you.

If we come back to what the utopists and their allies have been doing, we see the same story: each action can independently be justified one way or the other. Yet taking them together reveals a general pattern of systematically undermining safety and regulation, notably through:

- Spreading confusion through misinformation and double-speak
- Exploiting fear of AI risks to accelerate even further
- Capturing and neutralizing both regulation and AI safety efforts

## Spreading confusion through misinformation and double-speak

It should be cause for concern that the utopists are willing to spread misinformation, and go back on their commitments, and outright lie to stay on the frontline of the race.

An egregious recent example is how, in his [opening remarks](#) at a Senate hearing, OpenAI CEO's Sam Altman deliberately contradicted his past position to avoid a difficult conversation about AI x-risks with Senator Blumenthal, who quoted his [Machine Intelligence blog post](#):

*You have said 'development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity'; you may have had in mind the effect on jobs, which is really my biggest nightmare in the long term.*

Altman then responds:

*Like with all technological revolutions, I expect there to be significant impact on jobs, but exactly what that impact looks like is very difficult to predict.*

This is misdirecting attention from what Altman has historically written. The original text of Altman's blog post reads:

*The development of superhuman machine intelligence is probably the greatest threat to the continued existence of humanity. There are other threats that I think are more certain to happen (for example, an engineered virus with a long incubation period and a high mortality rate) but are unlikely to destroy every human in the universe in the way that SMI [Superhuman Machine Intelligence] could.*

Altman's response distorts the meaning of his original post, and instead runs with the misunderstanding of Blumenthal.

Similarly, in [his more recent writing](#) Altman has pulled back from his previous position on AGI risks and continues to downplay the risks by only addressing AI's impact to the labor market:

*As we have seen with other technologies, there will also be downsides, and we need to start working now to maximize AI's benefits while minimizing its harms. As one example, we expect that this technology can cause a significant change in labor markets (good and bad) in the*

*coming years, but most jobs will change more slowly than most people think, and I have no fear that we'll run out of things to do (even if they don't look like "real jobs" to us today).*

Altman simply changed his tune whenever it helped him, shifting from extinction risk to labor, claiming this was what he meant all along.

Anthropic has also explicitly raced and pushed the state-of-the-art after reassuring everyone that they would prioritize safety. Historically, Anthropic [asserted](#) that their focus on safety means that they wouldn't advance the frontier of capabilities:

*We generally don't publish this kind of work because we do not wish to advance the rate of AI capabilities progress. In addition, we aim to be thoughtful about demonstrations of frontier capabilities (even without publication).*

Yet they [released](#) the Claude 3 family of models, noting themselves that:

*Opus, our most intelligent model, outperforms its peers on most of the common evaluation benchmarks for AI systems, including undergraduate level expert knowledge (MMLU), graduate level expert reasoning (GPQA), basic mathematics (GSM8K), and more. It exhibits near-human levels of comprehension and fluency on complex tasks, leading the frontier of general intelligence.*

Anthropic's leaked [pitch deck](#) in 2023 is another example, with the AGI company stating that it plans to build models "orders of magnitude" larger than competitors. They write: "These models could begin to automate large portions of the economy," and "we believe that companies that train the best 2025/26 models will be too far ahead for anyone to catch up in subsequent cycles." These statements evidence that despite arguments about restraint and safety, Anthropic is just as motivated to race for AGI as their peers.

DeepMind CEO Demis Hassabis has been similarly inconsistent, [arguing that AGI risks are legitimate and demanding regulation and a slowdown](#), while simultaneously [leading DeepMind's effort](#) to catch up to ChatGPT and Claude. Elon Musk has also [consistently argued](#) that AI poses serious risks, despite being the driving force between both OpenAI and xAI.

These contradictory actions lead to huge public confusion, with some onlookers even arguing that [concerns about x-risk are a commercial hype strategy](#). But the reality is the opposite: the utopists understand that AGI will be such an incredibly powerful technology that they are willing to cut corners to reach it first. Today, even if these actors can signal concern, history gives us no reason to believe they would ever prioritize safety over racing.

## Turning care into acceleration

Recall that uncertainty and fear are essential components of FUD, and the whole industry playbook: by emphasizing uncertainty on one hand, and fear on the other, the status quo can be maintained.

In this case, utopists have managed to turn both the uncertainty about the details of AGI risks, and the fear of it being done by the “wrong” people, into two more excuses to maintain the status quo of racing as fast as possible.

First, they have been leveraging the uncertainty around the risks from AI as an argument to race to AGI, just so it’s possible to iterate and be empirical about them:

- OpenAI’s [Planning for AGI](#) argues that because it’s hard to anticipate AI capabilities, it’s best to move fast and keep iterating and releasing models:
 

*“We currently believe the best way to successfully navigate AI deployment challenges is with a tight feedback loop of rapid learning and careful iteration. Society will face major questions about what AI systems are allowed to do, how to combat bias, how to deal with job displacement, and more. The optimal decisions will depend on the path the technology takes, and like any new field, most expert predictions have been wrong so far. This makes planning in a vacuum very difficult.”*
- Similarly, Anthropic’s [Core Views on Safety](#) argues that safety is a reason to race to (or past) the capabilities frontier:
 

*“Unfortunately, if empirical safety research requires large models, that forces us to confront a difficult trade-off. We must make every effort to avoid a scenario in which safety-motivated research accelerates the deployment of dangerous technologies. But we also cannot let excessive caution make it so that the most safety-conscious research efforts only ever engage with systems that are far behind the frontier, thereby dramatically slowing down what we see as vital research. Furthermore, we think that in practice, doing safety research isn’t enough – it’s also important to build an organization with the institutional knowledge to integrate the latest safety research into real systems as quickly as possible.”*
- Elon Musk has used this argument too, [justifying his creation of xAI](#) with the claim that the only way to make AGI “good” was to be a participant:
 

*“I’ve really struggled with this AGI thing for a long time and I’ve been somewhat resistant to making it happen,” he said. “But it really seems that at this point it looks like AGI is going to happen so there’s two choices, either be a spectator or a participant. As a spectator, one can’t do much to influence the outcome.”*

Then, fear of the wrong people developing AGI, have been exploited as a further reason to race even faster:

- This is [the gist of the “entente strategy” proposed by Anthropic CEO Dario Amodei](#) and endorsed by many influential members of Effective Altruism, such as think tank

RAND, discussed above, as well as the language Sam Altman is increasingly using to stoke nation-state competition:

*“That is the urgent question of our time. The rapid progress being made on artificial intelligence means that we face a strategic choice about what kind of world we are going to live in: Will it be one in which the United States and allied nations advance a global AI that spreads the technology’s benefits and opens access to it, or an authoritarian one, in which nations or movements that don’t share our values use AI to cement and expand their power?”*

- And in one of the most egregious cases, Leopold Aschenbrenner’s [Situational Awareness](#), which captures much of the utopists’ view, argues that racing and abandoning safety is the only way to ensure that noble actors win, and can then take the time to proceed thoughtfully:

*“Only if we preserve a healthy lead of democratic allies will we have the margin of error for navigating the extraordinarily volatile and dangerous period around the emergence of superintelligence. And only American leadership is a realistic path to developing a nonproliferation regime to avert the risks of self-destruction superintelligence will unfold.”*

This has the same smell as [Big Tobacco reinforcing the uncertainty of science](#) as a way to sell products that kill people, and [Big Tech stoking the fear of losing innovation](#) as a way to ensure they can keep their monopolies and kill competition. These are not thoughtful arguments that examine the pros and cons and end up deciding to race; these are instead rationalizations of the desire and decision to race, playing on sacred notions like science and democracy to get what they want.

## Capturing and neutralizing regulation and research

The AGI race has also seen systematic (and mostly successful) attempts to capture and neutralize the two forces that might slow down the race: AI regulation and AI safety research.

### Capturing AI regulation

Utopists consistently undermine regulation attempts, emphasizing the risks of slowing or stopping AI development by appealing to geopolitical tensions or suggesting that the regulation is premature and will stifle innovation.

OpenAI, Google, and Anthropic all opposed core provisions of [SB 1047](#), one of the few recent proposals that could have effectively regulated AI, emphasizing the alleged costs to competitiveness and that the industry is too nascent. OpenAI [argued](#) that this kind of legislation should happen at the federal level, not at the state level, and thus opposed the law because it might be unproductive:

*However, the broad and significant implications of AI for U.S. competitiveness and national security require that regulation of frontier models be shaped and implemented at the federal level. A federally-driven set of AI policies, rather than a patchwork of state laws, will foster innovation and position the U.S. to lead the development of global standards. As a result, we join other AI labs, developers, experts and members of California's Congressional delegation in respectfully opposing SB 1047.*

Anthropic<sup>24</sup> [requested amendments](#) to the bill that would remove any state enforcement of safety frameworks, and to instead let companies iterate by themselves:

*What is needed in such a new environment is iteration and experimentation, not prescriptive enforcement. There is a substantial risk that the bill and state agencies will simply be wrong about what is actually effective in preventing catastrophic risk, leading to ineffective and/or burdensome compliance requirements.*

These arguments seem credible and well-intentioned on their own — there are indeed geopolitical tensions, and we should design effective legislation.

However, considered in concert, this is another FUD tactic: instead of advocating for stronger, more effective and sensible legislation, or brokering international agreements to limit geopolitical tensions, the utopists suggest self-regulation that keeps them in control. Most have published frameworks for the evaluation of existential risks: Anthropic's [Responsible Scaling Policy](#), OpenAI's [Preparedness Framework](#), and DeepMind's [Frontier Safety Framework](#).

The details of each of these policies are irrelevant, because at no point do they attempt or focus on enforcement by governments. Even worse, the documents clearly state that the AGI companies can and will edit the conditions and constraints for future AIs, however they see fit. Indeed, Anthropic hasn't missed the chance of [introducing a backdoor to racing even faster if they're not in the lead anymore](#):

*It is possible at some point in the future that another actor in the frontier AI ecosystem will pass, or be on track to imminently pass, a Capability Threshold without implementing measures equivalent to the Required Safeguards such that their actions pose a serious risk for the world. In such a scenario, because the incremental increase in risk attributable to us would be small, we might decide to lower the Required Safeguards. If we take this measure, however, we will also acknowledge the overall level of risk posed by AI systems (including*

---

<sup>24</sup> Anthropic eventually ended up [supporting the bill](#), with Dario Amodei writing a letter to Gov. Newsom that the "benefits outweighed the cost".

But this was after previously joining lobbying groups against the regulation, requesting revisions, and waiting until the bill already had a high chance of veto. Whether genuine or a case intended to create plausible deniability is uncertain, but in concert with their other actions, suspicious.

ours), and will invest significantly in making a case to the U.S. government for taking regulatory action to mitigate such risk to acceptable levels.

The utopists are exploiting fear of China and bad legislation to retain control over regulation, in turn devising proposals that let them modify boundaries as they approach them. To understand what type of regulation these companies really believe in, we can just track their actions: polite talk in public, while lobbying hard against regulation behind the scene, such as what [OpenAI did for the EU AI Act](#).

Both Big Tech and accelerationists actors are supporting the utopists in their push against any kind of binding AI regulation. Indeed, Microsoft, who bankrolls OpenAI, and Andreessen-Horowitz, who position themselves as champions of [little tech](#) and open-source, recently co-authored [a public post](#) arguing:

*As the new global competition in AI evolves, laws and regulations that mitigate AI harm should focus on the risk of bad actors misusing AI and aim to avoid creating new barriers to business formation, growth, and innovation.*

As discussed in Section 5, the impacts of these ideological actors on the policy space have been widespread, and now, unenforced reactive frameworks have become the primary governance strategy endorsed by governments and AI governance actors. For those familiar with the past two decades of Big Tech’s anti-regulation approaches, this is nothing new.

## Capturing safety research

The utopists have also found two ways to capture safety research: constraining “safety” and “alignment” work to research that can’t impede the race, and controlling the funding landscape.

OpenAI managed to redefine safety and alignment from “doesn’t endanger humanity” to “doesn’t produce anything racist or illegal.” The 2016 [OpenAI Charter](#) referenced AGI in its discussion of safety:

OpenAI’s mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity. We will attempt to directly build safe and beneficial AGI, but will also consider our mission fulfilled if our work aids others to achieve this outcome.

The May 2024 [OpenAI Safety Update strikes a different tone](#):

Our models have become significantly safer over time. This can be attributed to building smarter models which typically make fewer factual errors and are less likely to output harmful content even under adversarial conditions like jailbreaks.

Anthropic has similarly maneuvered to define safety as is convenient, [focusing its alignment and safety research efforts](#) on areas that do not actually limit racing but instead provide an edge. These include:

- Mechanistic interpretability, which tries to reverse-engineer AIs to understand how they work, which can then be used to advance and race even faster.
- Scalable oversight, which is another term for whack-a-mole approaches where the current issues are incrementally “fixed” by training them away. This incentivizes obscuring issues rather than resolving them. It also helps Anthropic build chatbots, providing a steady revenue stream.
- Evaluations, which test LLMs for dangerous capabilities. But Anthropic ultimately decides what is cause for slowdown, and their commitments are voluntary.

These approaches lead to a strategy that aims to use superintelligent AI to solve the hardest problems of AI safety, which, naturally, is an argument that is then used to justify racing to build AGI. [OpenAI](#), [Deepmind](#), [Anthropic](#), X.AI (“[accelerating human scientific discovery](#)”), and [others](#) have all proposed deferring and outsourcing questions of AI safety to more advanced AI systems.

These opinions have matriculated into the field of technical AI safety and now make up the majority of research efforts, largely because the entire funding landscape is controlled by utopists.

Most AI safety researchers concerned with extinction risk work at AGI companies which are ironically one of the few places with the funding and interest to pay for safety research. This structure benefits AGI companies. By controlling which areas of research get attention, AGI companies have successfully shifted the field of AI safety towards a paradigm that implicitly endorses building powerful AI.

Outside of AGI companies, [the main source of funding has been the Effective Altruism community](#), which has pushed for the troubling entente strategy and endorsed self-regulation through its main founding organ, [Open Philanthropy](#).

In this situation, it is expected that none of the “AI safety” actors are really pushing against the race, nor paying the costs of alignment.

## How will this go?

Because utopists are driven to be the first to build AGI to control the future, they intentionally downplay the risks and neutralize any potential regulatory obstacles. Their behavior, and the secondary motions of Big Tech players and accelerationists, further accelerating the race: everybody is trying to get ahead.

The race runs the risk of morphing from commercial to political as governments become increasingly convinced that AGI is a matter of national security and supremacy. Government intervention could override market forces and unlock significantly more funding, heightening geopolitical tensions. Political actors may be motivated to race out of fear that a competitor can deploy AGI that neutralizes all other parties.

This transition might already be in motion. The US government recently toyed with the idea of [establishing](#) national labs on AI:

The new approach won't propose the "Manhattan Project for AI" that some have urged. But it should offer a platform for public-private partnerships and testing that could be likened to a national laboratory, a bit like Lawrence Livermore in Berkeley, Calif., or Los Alamos in New Mexico. For the National Security Council officials drafting the memo, the core idea is to drive AI-linked innovation across the U.S. economy and government, while also anticipating and preventing threats to public safety.

And then, the US government started collaborating with AGI companies to [harness the power of AI for national security](#):

*The National Security Memorandum (NSM) is designed to galvanize federal government adoption of AI to advance the national security mission, including by ensuring that such adoption reflects democratic values and protects human rights, civil rights, civil liberties and privacy. In addition, the NSM seeks to shape international norms around AI use to reflect those same democratic values, and directs actions to track and counter adversary development and use of AI for national security purposes.*

This is yet another sign that the AI industry is gearing itself up to race even faster, not pivot toward safety.

## (7) A good future, if you can keep it <sup>25</sup>

We must find a way to avert extinction by AI and put humanity in full control of its technological development. This Compendium aims to be a tactical guide. Having mapped the historical trajectory and current landscape and challenges, we can now consider effective interventions.

The prerequisite for any global solution is a shared understanding of the issues. We need to generate civic engagement, build informed community opposition to the AGI race, and make the catastrophic risks from AI common knowledge. Human extinction concerns all of us, and a solution must eventually compound into global governance that can build technology deliberately and justly.

If you'd like to participate, we'd love for you to join us.<sup>26</sup>

In [Civic duty is the foundation of a response to AGI risk](#), we challenge the idea that the extraordinary risks from AGI require extraordinary solutions. Instead, we propose that what is needed today is basic civic engagement from concerned individuals and a whole lot of mundane (but important) work. This is not easy, as the civics process around technology has been undermined by Big Tech.

In [Creating a vision and a plan for a good future](#), we encourage readers to think critically about the future they want and how to get there. We then outline our high-level vision for a technologically mature society and a "Just Process" for making civilization-wide decisions about risks like AGI.

In [Actions that help reduce AGI risk](#), we propose a "bootstrapping" process to get involved immediately and learn the necessary skills to contribute. We then outline practical actions to improve AI safety communication, coordination, civics, and technical caution.

### Civic duty is the foundation of a response to AGI risk

The race to AGI is not an isolated incident, but representative of a broader societal tension between rapid technical progress and our inability to coordinate effectively to manage its consequences. While technological advancements unlock immense potential, the last 20

---

<sup>25</sup> The phrase "a republic, if you can keep it" is attributed to Benjamin Franklin and conveys the idea that the stability and success of a republic would depend heavily on the engagement of its citizens.

<sup>26</sup> If you'd like to discuss what you can do more directly, please join us alongside nonprofit ControlAI [here](#)

years have exhibited major failures in our ability to direct these innovations toward the collective good and manage their externalities: we are witnessing environmental collapse, increased partisanship and devolvement of our information sources, mental health epidemics, and trillion dollar companies that profit off of stealing human attention.

Today's landscape is a product of Big Tech's coup; technology was built faster than governments could regulate, agitated by a doctrine of fear ("if you regulate us, the economy will die"), uncertainty ("we can't regulate this tech now because the future is uncertain"), and doubt ("governments are too unfamiliar with the tech to make prudent decisions on how to regulate"). Big Tech attacked both regulatory legislation and the legislative process itself, building a public ideology that the law is intrinsically bad and that governments and public institutions are intrinsically bad, all while embedding lobbyists and allies into governments to capture power and become vital to intelligence, defense, and other public projects.

This coup has led people to believe that issues with technology are someone else's problem – governments, NGOs, the UN, whomever – and that individuals are powerless to intervene on new technical questions.

The sense of powerlessness has spilled over to AI: key decisions are made by a handful of AGI companies that have partnered up with Big Tech and captured technical safety and governance efforts. AI progress is moving so fast that newcomers have a hard time making sense of the race. And the actions that are necessary to get to a good future – such as building more stable global governance – appear far out of reach of ordinary people. Assessing all of these, concerned individuals may feel like it is hopeless to contribute.

It is a mistake to read the situation as hopeless, and plays into the hand of the actors driving the race.

**Although we are in an emergency, the work to stop AGI today is not hardcore, but an exercise of our basic civic duty.**

Public concern for climate change was activated by the collective efforts of a handful of informed citizens who cared, took the time to get informed, and slowly chipped away at Big Oil's ability to obfuscate the problem. We can learn from that playbook to mitigate AGI risk by educating the public and encouraging simple actions like talking to friends about the issues, writing on social media, and contacting local representatives. As public perception of the risks strengthens, it will catalyze other interventions like creating [AI Safety Institutes](#), [public statements](#), [international dialogues](#), [protests](#), [tracking integrity](#)

[incidents of labs](#), [formalizing boundaries for AI behavior](#), [upskilling programs on AI risk](#), [educational videos](#), and more.

Most people will not and cannot dedicate their lives to working against the threat of AGI or similarly grand problems, and this is good. The world we care to protect is not the world in which everyone is single-mindedly tackling humanity's immediate priority; it is one in which people enjoy their lives as civilians, not soldiers.

But proactive involvement from more people is necessary. While “civics” isn't the entire answer to the problem, it is the foundation. We humans have gotten ourselves into this predicament, and now it's on us to do the work to get out.

## Creating a vision and a plan for a good future

The first step to shaping the future is defining a plan for dealing with the risks of AGI and getting to the kind of future we want to live in.

There are no adults in the room writing this plan for us. Consider today's most sophisticated AI legislation, [the EU AI Act](#). Although it acknowledges “systemic risk,” it does little to manage it. AGI companies are only required to evaluate the capabilities of their models, report training information, and ensure a baseline level of cybersecurity. This does not give humanity a roadmap to a good future.

A more comprehensive proposal is offered by [A Narrow Path](#), which assumes a worldview similar to our own and investigates how to prevent superintelligence development for 20 years. They consider how to prohibit each of the risk vectors that could lead to superintelligence, such as AIs improving AIs, AIs capable of breaking out of their environment, unbounded AIs, and AIs with vast general intelligence. They then consider what is necessary to enforce that, such as strong regulation, physical kill-switches, and national regulators to monitor AI usage. They then turn to balancing the international situation, proposing an international treaty with a judicial arm and new international institutions as a way to get to stable global governance capable of preventing rogue actors from building superintelligence.

We endorse A Narrow Path's proposal in full, and recommend that you read it.

However, it is most valuable to write your own plan first. We mean this literally: open a new document, name your goal, and bullet point the actions necessary to get there.

Your plan does not need to be comprehensive or perfect, but the exercise prompts you to make your current point of view explicit and reckon with the challenges ahead. Considering even the crudest strategies to avoid AI risk (e.g. “turn off all datacenters!”) immediately raises thorny issues: how is this going to be implemented? Is it really sufficient? What if there are bad actors?

A good plan:

- Articulates a vision for the future you actually want to live in.
- Adequately grapples with the risks from superintelligence. A Narrow Path is a comprehensive plan to do this, but it is not the only way. As you develop your own strategy, you will likely find other (or better!) ways to control AI.
- Defines actionable steps. This is where individual plans diverge — only you can author next steps. A Narrow Path isn’t actionable; a lawmaker could read it and decide to pursue oversight and meaningful legislation, but they would still need to write bills themselves, push them through their jurisdictions, ensure they’re not [vetoed at the last minute](#), and so on. Your plan must define a specific list of actions to take and a way to evaluate roadblocks. You are the expert on your local situation, and civic engagement is built from local engagement.

To develop your plan, talk with friends, engage in public discourse, read and consider alternative perspectives, and generally keep working to refine your worldview. But these are all things you can do over time, and they don’t need to block you from writing a first plan.

If you’ve envisioned the future you want and written an initial plan, you’ve made it further than almost anyone else in the world on this subject and are ready to take action.

## The authors’ plan

We the authors – Connor, Gabe, Chris, Andrea, Adam – make plans the same way. What we share below is a loose sketch of our plan and how we decide which actions to take.

Each of us have different values and visions, and we do not claim to know what is best for humanity, or what is the right future. But we agree that we should not aim for a specific utopia, but rather a “Just Process,” that aims to determine what the right future is, and enables humanity to work together to reach it.

Today, humanity can build technology powerful enough to end civilization, yet we lack the ability to collectively steer this progress in a safe direction. The extinction risk posed by AGI is the ultimate expression of this imbalance. Despite widespread acknowledgment of the dangers, no single person or institution possesses the capabilities, authority, and clarity to prevent these developments. And yet, individuals with [visions](#) of [utopia](#) can expose everyone else on the planet to monumental risk by racing toward AGI.

We do not claim to know what is just, but we are confident that the current state of affairs is unjust. We want to escape this out-of-control development and build a Just Process that enables humanity to consciously choose its fate.

We work backwards from this loose vision of a future to arrive at a plan.

- We can't get to a Just Process without a global solution.
- We can't get to a global solution without improvements in coordination, science, and moral philosophy; many of the hard problems of alignment are the same problems humans need to solve to figure out how together.
- These challenges will take many years to solve, time which we do not have due to the race to AGI and the risk of extinction.
- We must therefore buy time, slowing or stopping AGI development as soon as possible.
- The only way to do so is through governance, and we endorse the proposal offered by [A Narrow Path](#).
- Building institutions that can regulate AI globally comes with challenges: nation-state competition threatens global cooperation, AGI companies are now fueling geopolitical arms races, Big Tech lobbying has made it extremely difficult to pass tech regulation and captured governance efforts. Coordinating most of the existing AI safety actors is futile as their underlying motivation endorse racing to AGI.<sup>27</sup>

This brings us to the current bottlenecks, namely the lack of public consensus around the risks from AGI, the lack of an AI safety ecosystem free from AGI companies' influence, and the lack of coordination among current actors concerned with halting the race to AGI. This was the motivation for writing this document: an attempt to articulate a worldview around

---

<sup>27</sup> See "entente" in Section 5

the risks from AGI clearly enough that we can start to build coordination among those who see the situation similarly.

This is an example of how to connect a high level vision to a broad plan, and then to specific actions.

## Actions to help reduce AGI risk

Helpful actions to reduce AGI risk derive from a good plan. If you've written your first plan, you've got all you need to start. Your plan may not be good to begin with, but taking action, reflecting, and iterating on your plan is the best way to improve it.

Artificial intelligence is a technical subject, and some existing [guides on getting involved in AI safety](#) recommend learning more about the technology, taking [AI safety courses](#), or planning for a career in technical AI safety or governance. These approaches are fine if you have the interest and ability to pursue these, but they are not necessary.

- **Write things down.** What isn't written down doesn't exist. Your mind is fallible and forgetful, put as much of it in writing as possible so you can rely on and iterate on it later.
- **Think about what you do.** Your mind is your most important tool. To strengthen it, you need to think about it, about your motivations, what you've learned, what your next plans are, etc.
- **Keep things grounded,** including, and especially, for intellectual labor. It's difficult to detect progress without getting feedback from reality. The best way to practice is usually to do.
- **Keep reasonable habits.** Spend time with your friends and family, eat healthy, get enough sleep, touch grass. When faced with enormous challenges, it can be tempting to sacrifice everything in your life to struggle against them. This is unproductive and self-destructive. It's a marathon, not a race. Keep that day job. It's much better to work with someone who gives their all 10% of the time than with someone who gives their all 110% of the time and then burns out. If thinking about the risks becomes overwhelming, consider reading about [mental health and AI alignment](#), talking to a professional, and taking a break.

To make consistent and useful progress, on both our projects and ourselves, we must be capable of contributing reliably and independently.

So let's dive in: how can you make your plan to reduce risk from AGI actionable, particularly if the goals are so grand as to demand solutions humanity has not yet come up with? Below, we argue that communication, coordination, civics, and technical caution are necessary to

reach a broader solution, and suggest shovel-ready work you could do to help address today's bottlenecks.

## Communication

One of the simplest things you can do to contribute is to communicate publicly about the risks of AGI. From posting on social media to writing in a local newspaper, thoughtful opinions that add to common knowledge of the risks is helpful.

Common knowledge establishes a basis for working together to solve problems. Before collaborating, people must agree on what the problem is. To solve the risks from AGI, society must agree on what the problem is and that the risks are real, imminent, and time sensitive.

Certain problems can only be solved when there is known consensus. Consider a town election where three candidates are running for office. Alice and Bob are well-known candidates, but are both terrible choices for office, and Charlie is a newcomer and is excellent. Individually, everyone in town wants Charlie to win, but they don't know that everyone else feels the same way and worry that their vote would be wasted. In this hypothetical, the entire town may want Charlie to win the election, but he may not be voted in because there is a *lack of common knowledge*. Public statements will help Charlie win: more people need to declare their support so that everyone knows that there is a consensus and that their vote for Charlie would be meaningful.

Common knowledge is the way that a group changes its mind. Without common knowledge of AGI risks, collective concern could still result in inaction because the subject is considered unpopular. And indeed, this partially explains the current landscape: the race to AGI continues in full force, even though [polling suggests](#) that people overwhelmingly "worry about risks from AI, favor regulations, and don't trust companies to police themselves." We need to convince humanity to collaborate on this problem.

Humanity can solve a great number of issues, but only the ones that it is paying attention to. Communication also drives saliency, making ideas noticeable and prominent.

On an individual level, most people do not think about most things. There are simply too many things to pay attention to, and it can be hard to decide which of the many critical issues of geopolitical importance deserve attention over the very real and tangible challenges in one's own personal life. Without reminders and exposure, especially from peers, it's easy to ignore an issue.

Saliency is a scarce resource. This is why advertising works, and also why it is harmful: it pulls individual and group attention away from prosocial ideas and toward meaningless ideas. And this is why the communications and lobbying strategy of Big Tech is to distract and delay. Scattering attention away or dragging out a legal case reduces saliency as people's attention wanes, and makes it harder for meaningful intervention to come together.

The type of communication that is needed today is the type that cuts through distraction, and makes the risks of AGI clear, common knowledge and salient. This is necessary to convince humanity to do something.

—

**The core message to communicate is that racing to AGI is unethical and dangerous, and that the actors doing this are harmful to society. Humanity's default response to risks of this magnitude should be caution. Today, the position is to allow private companies to keep racing until there is a problem, which is untenable as allowing private companies to build nuclear reactors until one melts down.**

This message needs to go hand in hand with mature discussion about the real risks of AGI and superintelligence. Because actors like AGI companies and Big Tech have strong incentive to build AGI, public communications are a narrative battle. These actors will continue to use a playbook that downplays or obscures the risks and makes their inventions seem societally progressive and harmless, while lobbying against any regulation that slows them down. Challenging these tactics requires understanding this strategy, and ensuring companies racing to AGI have their feet held to the fire.

Communications help raise common knowledge and improve the quality of the debate. Here are some actions you could take today:

- Share a link to this Compendium online or with friends, and provide your feedback on which ideas are correct and which are unconvincing. This is a living document, and your suggestions will shape our arguments.
- Post your views on AGI risk to social media, explaining why you believe it to be a legitimate problem (or not).
- Red-team companies' plans to deal with AI risk, and call them out publicly if they do not have a legible plan.
- Find and follow 20 social media accounts that discuss risks from AGI. Regularly engage with this content, sharing and debating it.
- Push back against the race to AGI when you hear people advocate for it, engaging in productive debate and avoiding ad hominem.

- Produce content based on AGI risk, like a video, meme, short story, game, or art. If you do this more routinely, consider how your audience engages with the content and work to increase the quality of understanding viewers have over time.
- Write an opinion piece for a local newspaper, or an op-ed for a larger publication. Highlight the battle lines of argumentation – where do the risks seem genuine, and which ideas demand more debate?
- Create websites that discuss the risks or help collect important information about the race to AGI, such as websites that:
  - Quote what leaders of AGI companies have said about the risks of AGI.
  - Quote what politicians have said about the risks of AGI.
  - Detail which AI capabilities currently exist and how fast they are developing. Visualizations and charts or explanations that can be shared and built upon by others are especially useful.
  - Explain the history of the race to AGI.
  - Track lobbying efforts of Big Tech and which issues they are paying attention to.
  - Offer basic explanations of the risks of AGI to different audiences, such as artists, youth, religious groups, and so on.
  - Collect public opinions on AI, or offer platforms for individuals to voice their concerns.
- Talk to your friends about AGI risks, and write down what they say. Track argumentation and aim to improve the quality of your thinking and theirs on the subject.
- Organize a local learning group or event, like a discussion or town hall, to bring people together to talk about the risks and what can be done.

Communications compound. Common knowledge is built with small, consistent communications. Consider deliberately scheduling in weekly time to learn and share about AI risks.

There are, of course, wider actions that can be taken here. If you have the resources and skills, you could create campaigning organizations that spread awareness of superintelligence risks, or start strategic communications organizations that draw attention to AI risks around particular key events, such as AI summits or key legislation developments. For those with these kinds of affordances who are interested in helping, please [reach out to us](#).

## Coordination

Communication about the risks from racing to AGI is an essential first step for improving the situation, but it is insufficient. Even if most people agree that there is a problem and much should be done, working individually will not get us to a global solution.

Group coordination is necessary to succeed, but it is non-trivial. There are many ways this has already failed around AI safety. The most extreme historical cases involve cult dynamics, and becoming the type of community that accidentally kick-started the race to AGI (see Section 6 on how early Singularitarians led to the formation of DeepMind, OpenAI, and Anthropic).

We need strong community builders and communities. These communities cannot make excuses for actors racing to AGI just because they are friends, as is the case with existing AI safety communities. They must instead stay laser-focused on ending the race to AGI, providing a written plan for a safe future, ideally a public one that can be improved over time.

And most importantly, these communities must become mainstream. Human extinction concerns all of us, and any issue of that scale requires involving many, many people. These groups must communicate in a way that is legible to people who do not share their very specific cultural background and teach the relevant technical concepts to a wide audience in a non-jargony way. They must engage with political parties, civil society, and other institutions.

If one economist learns about a method that will stop an impending financial crisis, she can't immediately stop the economy from crashing. There are many, many steps necessary to move between her idea and a wider solution: she must convince members of government with the authority to control financial matters; they must run calculations to assess if the method is accurate; a bill may need to be written and passed as new legislation; the treasury may need to print money; financial institutions may need to change their policies, and so on.

There is no silver bullet. We need communities to try many different strategies. And these communities must be ready to not just talk about the issues, but also engage with institutions and improve them until they are competent enough to deal with the risks.

We do not believe there is any such community, with the perspective and resources to help. We have not built it yet, partially for lack of focus (we are writing the Compendium and working on technical projects), and partially for lack of still: we have tried small attempts at community building on the side and been unsuccessful. We have [a small effort alongside](#)

[ControlAI](#) that you are welcome to join and participate in, but the kind of global community that is needed is much stronger and better established than these local efforts.

What is needed until a global community exists is to lay the foundation for one to arise. Start small, find like-minded and concerned collaborators, work alongside them, and find reliable ways to keep collaborating:

- Discuss the risks with concerned or open-minded friends and family members. Suggest that they also write actionable plans for how to get involved, and regularly get together to make progress on these plans. This is the most local community you could form, and getting experience working alongside others on the problem is critical.
- Share what you are working on publicly, including how you work and how others can work with you.
- Learn about the groups working on AI safety and get involved. Join online communities and Discord groups that take the risks from AGI seriously, such as [ControlAI](#) or [PauseAI](#), and get involved with their activity and improve their projects.
- Compare strategies between individuals and groups you're in contact with and see if there are shared projects that make sense to pursue.
- Connect existing groups or organizations working on reducing the risks from AGI. Create group chats or other communications channels, and work on finding projects that make sense to pursue across multiple organizations (e.g. a communications plan that shares common messaging).
- Participate in events about AI extinction risks such as local discussion groups, AI summits, and so on.
- Take part in upskilling programs where you can meet people who care about these risks
- Come up with concrete projects that could form the basis of collaboration with others.

These are proto-community-building efforts, designed to get more people involved in direct contributions to stopping the race to AGI.

If you have the resources to get much more involved, such as a position of authority or a company with resources, then explore operating at a higher scale. Consider how to get others involved, which groups you believe are working on the problem well and how to work alongside them, and if you or your organization have the capacity to play a leadership role in community building, which is sorely lacking today. Or consider larger [coordination projects](#) outside of AI, which improve the global commons and make working on issues like this easier.

Over time, these efforts must add up to build an AI safety ecosystem away from AGI companies and the groups they have captured, coordinating between non-profits, startups, funders, regulators, academics, governments, and concerned individuals who are opposed to the AGI race and take extinction risks seriously.

## Civics

Civics is about taking responsibility for things that are larger than you, and acting on them. It is looking at civilizational coordination problems like how we will respond to climate change or AGI, and raising your hand to be a part of this.

If you live in a democracy, you have a voice and the ability to influence key local decisions. This means not only voting based on what your local politicians want to do on AI risks, but also messaging them and your fellow citizens.

Participation is necessary for democracy's function. Democracies contain many different processes intended to give power and sovereignty to the citizens, but these only work if the citizens actually exercise them. The less citizens make their voice heard, participate, ask for things and unite around core issues, the less oversight there is on the elected officials. This lack of pressure makes lobbying and political manipulation so potent: if politicians do not face pressure from citizens, they can keep their office while disregarding risks that people care about.

Some more concrete examples of civic actions include:

- Figure out where your local government (council/city/state) stands with regard to AI extinction risk. Who cares about it there? Who doesn't care about it? Is it because they understand and disagree or because they don't understand?
- Publicize what your local government thinks and proposes to do with regard to AI extinction risk. This way, others do not need to repeat your efforts and can learn directly from your investigation. For example, if others want to vote based on who takes AI risks seriously, having the information accessible on a website would make it much easier for them to judge candidates.
- Educate your local government and fellow citizens about AI extinction risk by writing to politicians, calling them, bringing up the topic at local events, sharing education materials about AI risks, and engaging others to talk about the issue. These deeper, in-person conversations often help people change their mind, by demonstrating not just that arguments for the risks exist, but that you – a real human in front of them right now – cares about the topic.

- Vote according to the position on AI extinction risk of the candidates. Although this doesn't have to be the only factor in the vote, taking this seriously and indeed voting is the simplest and yet most important part of civics.

These actions may look most useful in constituencies that have a large sway in the race to AGI, such as California where SB 1047 was proposed. But local action is meaningful everywhere. Good policies in one jurisdiction can serve as precedent for another, and raising awareness with elected officials and civil servants trickles up and grows the common knowledge and saliency of AI risks to government in general. This is necessary to motivate any coalition to push for meaningful regulation.

If you have more influence, such as working in government, then consider the recommendations in A Narrow Path and if there is anything recommended there that could apply to your jurisdiction. At a minimum, additional statements about the risks from elected officials matter greatly. If you'd like to do even more, please [reach out](#). We would be happy to discuss what can be done in your local context.

## Technical Caution

Communication, coordination, and civics are the current bottlenecks to larger solutions to AGI risk. There will be more obstacles in the future, but solving them will require making meaningful traction on these foundational issues.

To avoid worsening the problem, we must redirect technical development. As the race to AGI continues, it is developers of the technology, both at AGI companies and in the open-source community and academia, who are increasing the risks.

It is not that any single developer or released research paper is entirely responsible, but the overall wave of research and development is what will lead to AGI, superintelligence, and eventually extinction. Each open-source project that advances AGI-like capabilities, research paper that offers AI optimizations, and product that improves the general capabilities of AI chips away at the time we have left until AGI. What is especially dangerous is not just working on capabilities, but working on and sharing capabilities publicly.

If you are participating in this kind of research, you should stop. In particular, not compounding AGI risk means that you should:

- Not work for companies participating in the AGI race.
- Not share (or even draw attention to) AGI-advancing research or secrets, such as methods to improve agency, optimizations, self-improvement techniques, and so on. This includes:

- Publishing papers
- Releasing open-source projects
- Writing blog posts
- Discussing with other technical researchers,
- and similar.
- Challenge friends, colleagues, and other technical developers who are accelerating the AGI race by working at AGI companies or releasing AGI-enabling research.

Stopping is expensive. If you do not publish your research, you will get less academic credit, less VC money, less recruiting opportunities, and so on. But compromises are a deal with the devil; if you loosen your standards of publication, you will actually get more power: more people will like you, prospects of being hired will go up, and so on. If you are Anthropic and you push capabilities towards AGI by releasing [agentic AI that can use a computer](#), you will get a lot of useful clout and money. Over time, these concessions add up to the entirety of the issue.

**If you genuinely care about the risk, you should pay the cost and actually stick to this rule: do not release AGI capabilities research or products.** If everyone abided by this rule, we'd be in a safe world.

This is not to say that you should completely stop working on AI, or stop using AI to make money as long as you do so ethically. There are hundreds of AI-enabled projects that will improve the world, like improving healthcare and automating menial labor, which do not require pushing the boundaries of AI closer to risk. AI projects that try to solve a narrow problem in order to improve people's lives are much less likely to lead to AGI.

*Note: We have conveyed throughout this document that AI safety entails challenging research and technical problems, from solving AI alignment to building technical measures to bound the capabilities of AI. Unfortunately, this is a very large subject out of scope for this document. If you are passionate about this and want to contribute to a technical solution, send us an email.*

## (8) Outro

Does this all add up to a good future? Will this work? Are we going to be ok?

The truth is that we don't know.

But we do know that if we don't do it, it won't happen.

At the end of the day, we are a bunch of chimps with brains three times too big, trying our best to make it through the day, to the next generation, and eventually to the stars.

The odds were always stacked against us — we are the species that shouldn't be. Millions and billions of years and there never was a species like humans. No other species makes up for their soft skin and lack of claws like humans do. No other species makes art, culture and science like we do.

Following the [Toba volcanic catastrophe](#), humanity was reduced to as few as 3000–10000 individuals. And yet, somehow, we made it through. Humanity has made it through predation, epidemics, natural catastrophes, and ice ages. Today we ourselves are our greatest threat.

The survival of that strange chimp was never guaranteed, and neither is it today. The century ahead of us is a challenge that strange chimps like us were never designed to face. And so, once more, we need to do the impossible, what no other species has done before:

We need to work together. As a tribe, a civilization, a species, for those strange chimps that came before us, and for those hopefully still to come.

We can build a good future, but we cannot do it alone. Humanity is the social animal — we do great things, and even greater things together.

If we don't do it, it won't happen. So let's do it.

- Connor Leahy, October 2024

# The Compendium

*By Connor Leahy, Gabriel Alfour, Chris Scammell, Andrea Miotti, Adam Shimi  
V1.3.1 - Dec 9, 2024*

*Contact: [hello@thecompendium.ai](mailto:hello@thecompendium.ai)*

*Website: [thecompendium.ai](https://thecompendium.ai)*