IF ANYONE BUILDS IT, EVERYONE DIES

WHY SUPERHUMAN AI WOULD KILL US ALL

ELIEZER YUDKOWSKY & NATE SOARES

IF ANYONE BUILDS IT, EVERYONE DIES

Why Superhuman AI Would Kill Us All

ELIEZER YUDKOWSKY & NATE SOARES

Little, Brown and Company
New York Boston London

Little, Brown and Company
Hachette Book Group
1290 Avenue of the Americas, New York, NY 10104
littlebrown.com
X @littlebrown
Facebook.com/littlebrownandcompany
Instagram @littlebrown

First Edition: September 2025

Little, Brown and Company is a division of Hachette Book Group, Inc. The Little, Brown name and logo are trademarks of Hachette Book Group, Inc.

The publisher is not responsible for websites (or their content) that are not owned by the publisher.

Little, Brown and Company books may be purchased in bulk for business, educational, or promotional use. For information, please contact your local bookseller or the Hachette Book Group Special Markets Department at special.markets@hbgusa.com.

Print book interior design by Bart Dawson

ISBN 9780316595667 (ebook)
Library of Congress Control Number: 2025938106

E3-20250820-JV-NF-ORI

CONTENTS

To all the humans who ever died, in the process of our species coming this far;

To all those who are still among the living;

And to all the children that could someday be.

INTRODUCTION

HARD CALLS AND EASY CALLS

“MITIGATING THE RISK OF EXTINCTION FROM AI SHOULD BE A GLOBA priority alongside other societal-scale risks such as pandemics and nuclear war.”

In early 2023, hundreds of Artificial Intelligence scientists signed an open letter consisting of that one sentence. These signatories included some of the most decorated researchers in the field. Among them were Nobel laureate Geoffrey Hinton and Yoshua Bengio, who shared the Turing Award for inventing deep learning.

We—Eliezer Yudkowsky and Nate Soares—also signed the letter, though we considered it a severe understatement.

It wasn’t the AIs of 2023 that worried us or the other signatories. Nor are we worried about the AIs that exist as we write this, in early 2025. Today’s AIs still feel shallow, in some deep sense that’s hard to describe. They have limitations, such as an inability to form new long-term memories. These shortcomings have been enough to prevent those AIs from doing substantial scientific research or replacing all that many human jobs.

Our concern is for what comes after: machine intelligence that is genuinely smart, smarter than any living human, smarter than humanity collectively. We are concerned about AI that surpasses the human ability to think, and to generalize from experience, and to solve scientific puzzles and invent new technologies, and to plan and strategize and plot, and to reflect on and improve itself. We might call AI like that “artificial superintelligence” (ASI), once it exceeds every human at almost every mental task.

AI isn't there yet. But AIs are smarter today than they were in 2023, and much smarter than they were in 2019. AI research has yielded jump after jump after jump in AI capability, in 2012[i] and 2016[ii] and 2020[iii] and 2022[iv] and 2024.[v] We don't know whether progress will peter out, causing these jumps to halt for a time until new methods and technologies are invented. We don't know how many jumps are left before AI becomes the extinction-level threat that the letter's signatories warned about. But history has shown time and time again that AI researchers invent new methods and overcome old obstacles. Progress is often surprisingly fast. Most computer scientists in 2015 would have told you that ChatGPT-level artificial conversation wouldn't be in reach for another thirty or fifty years.

We didn't know when artificial superintelligence would arrive, but we agreed it should be a global priority. In fact, we think the open letter drastically undersells the issue.

We were invited to sign that one-sentence open letter in our capacity as co-leaders of the Machine Intelligence Research Institute (MIRI), a nonprofit institute. MIRI had been working on questions relating to machine superintelligence since 2001, long before these issues got much publicity or funding. To oversimplify: Among the few who have been following this matter for decades, MIRI is acknowledged as having worked on it the longest. One of us, Yudkowsky, is the founder of MIRI; the other, Soares, is its current president.

MIRI was the first organized group to say: "Superintelligent AI will predictably be developed at some point, and that seems like an extremely huge deal. It might be technically difficult to shape superintelligences so that they help humanity, rather than harming us. Shouldn't someone start work on that challenge right away, instead of waiting for everything to turn into a massive emergency later?"

We did not start out saying that. Yudkowsky began by trying to build machine superintelligence, in the year 2000. But in 2001, he realized that it would not necessarily turn out friendly. And in 2003, he realized that problem would be hard.

For its first two decades, MIRI was a technical research institute, without much involvement in policy. The organization mostly held workshops for interested scientists and housed a few promising researchers. We tried to figure out the math for understanding and shaping superhuman machine intelligence, and for predicting how it might go wrong.

MIRI also had some downstream effects that we now regard with ambivalence or regret. At a conference we organized, we introduced Demis Hassabis and Shane Legg, the founders of what would become Google DeepMind, to their first major funder. And Sam Altman, CEO of OpenAI, once claimed that Yudkowsky had "got many of us interested in AGI"[vi] and "was critical in the decision to start OpenAI."[vii]

MIRI's history is complicated, but one way of summarizing our relationship to the larger field might be this: Years before any of the current AI companies existed, MIRI's warnings were known as the ones you needed to dismiss if you wanted to work on building genuinely smart AI, despite the risks of extinction.

More recently, as AI has begun to take off, we watched with concern as some of the newer people starting AI companies began talking about artificial superintelligence as a source of vast, wonderful powers. Powers that they assumed they'd control. The main danger, according to many of these founders, was that the wrong people might "have" ASI. They talked of the need to win an "AI arms race." As for the possibility that you don't "have" an ASI, the ASI has an ASI—that the only winner of an AI arms race would be the ASI itself—well, these founders didn't talk about that.

We saw that AI capabilities were growing very fast.

We saw that the research field in which we were involved—the one aimed at understanding AIs and having them maybe not go wrong—was progressing much, *much* slower.

The AI companies' headlong charge toward superhuman AI—their efforts to build it as quickly as possible, before their competitors could do it—started looking to us like a race to the bottom. The industry was careening toward disaster: the sort that would get into textbooks as an example of how *not* to do engineering—except no one would be left alive to write the analysis.

It no longer seemed realistic to us that humanity could engineer and research its way out of catastrophe. Not under conditions like these. Not in time.

We wrote off our previous efforts as failures, wound down most of MIRI’s research, and shifted the institute’s focus to conveying one single point, the warning at the core of this book:

If any company or group, anywhere on the planet, builds an artificial superintelligence using anything remotely like current techniques, based on anything remotely like the present understanding of AI, then everyone, everywhere on Earth, will die.

We do not mean that as hyperbole. We are not exaggerating for effect. We think that is the most direct extrapolation from the knowledge, evidence, and institutional conduct around artificial intelligence today.

In this book, we lay out our case, in the hope of rallying enough key decision-makers and regular people to take AI seriously. The default outcome is lethal, but the situation is not hopeless; machine superintelligence doesn’t exist yet, and its creation can still be prevented.

How can anyone be confident of what will happen with regard to AI? “Prediction is very difficult, especially about the future,” goes the aphorism. Most of what we’d like to know about the future is not actually predictable. We can’t tell you next week’s winning lottery numbers, for example. One set of numbers seems just as likely as any other.

But some facts about the future *are* predictable. If you, personally, buy a lottery ticket tomorrow, we don’t know what complicated theories or whims you’ll use to pick your numbers, and we don’t know what numbers will come up, but all that uncertainty adds up to a very strong prediction that you will not win the lottery. Similarly, if you drop an ice cube into a glass of hot water, it’s impossibly complicated to predict where each molecule will end up ten minutes later—but all that uncertainty adds up to a near-certain prediction that the ice cube will melt. Half of physics is like that: We can’t calculate which exact path gets taken, but we know where almost all paths lead.

Some aspects of the future are predictable, with the right knowledge and effort; others are impossibly hard calls. Competent futurism is built around knowing the difference.

History teaches that one kind of relatively easy call about the future involves realizing that something looks theoretically possible according to the laws of physics, and predicting that eventually someone will go do it. Heavier-than-air flight, weapons that release nuclear energy, rockets that go to the Moon with a person on board: These events were called in advance, and for the right reasons, despite pushback from skeptics who sagely observed that these things hadn't yet happened and therefore probably never would. People who strapped wings to their arms and jumped off hills looked all sorts of foolish, and were mocked by their contemporaries, and in fact hurt themselves and failed—but that didn't stop the Wright brothers from figuring out how to fly.

Conversely, predicting exactly *when* a technology gets developed has historically proven to be a much harder problem. People say that a technology is two years off when it's really fifty years, or say fifty years when it's really two years and they themselves will build that technology. "Man will not fly for a thousand years," Wilbur Wright said to Orville Wright in 1901, fed up with the unpowered glider they were testing at the time. Two years later, in 1903, the Wright brothers flew.

Successful forecasting is not about being clever enough to predict the sort of details that usually can't be predicted. It is not about inventing a complete story about what will happen and then being magically correct. Rather, it's about finding aspects of the future that become easy calls when viewed from the right angle.

We don't know when the world ends, if people and countries change nothing about the way they're handling artificial intelligence. We don't know how the headlines about AI will read in two or ten years' time, nor even whether we have ten years left. Our claim is not that we are so clever that we can predict things that are hard to predict. Rather, it seems to us that one particular aspect of the future—"What happens to everyone and everything we care about, if superintelligence gets built anytime soon?"—can, with enough background knowledge and careful reasoning, be an easy call.

Humanity's extinction by superhuman AI might not seem like an easy call at first glance. But that's what the rest of this book is for. Just as it takes some arithmetic to calculate the chance of winning a lottery, just as it takes some ideas from thermodynamics to say why an ice cube predictably melts, so does it take some background to understand why artificial intelligence poses an imminent extinction risk to humanity. Once those foundations are in place, though, predicting the outcome of our present trajectory starts to look grimly, horribly straightforward.

Even in the face of superhuman machine intelligence, it can be tempting to imagine that the world will keep looking the way it has over the last few decades of our relatively short lives. It is true, but hard to remember, that there was a time as real as our own time, just a few short centuries ago, when civilization was radically different. Or millennia ago, when there was no civilization to speak of. Or a million years ago, when there were no humans. Or a billion years ago, when multicellular colonies had no specialized cells.

Adopting a historical perspective can help us appreciate what is so hard to see from the perspective of our own short lifespans: Nature permits disruption. Nature permits calamity. Nature permits the world to never be the same again.

Once upon a time, 2.5 billion years ago, an event occurred that biologists call the Oxygen Catastrophe: A new life form learned to use the energy of sunlight to strip valuable carbon out of air. That life form exhaled a dangerously toxic and reactive chemical as waste, poisonous to most existing life: a chemical we now call “oxygen.” It began to build up in the atmosphere. Most life—including most of the bacteria exhaling that oxygen—could not handle its reactivity, and died. A lucky few lines of cells adapted, and eventually evolved into organisms that use oxygen as fuel. But things never went back to the old normal. The world was never the same again.

Once upon a time, the continents were barren rock. Then in the blink of an evolutionary eye, they were carpeted in vegetation. Soon after, forests were teeming with life. The world was never the same again.

Once upon a time, some humans domesticated wheat and barley. In a tinier fraction of an evolutionary eye-blink, they started building civilizations. The world was never the same again.

Once upon a time in the 1930s, there were warning signs that certain families would no longer be safe in Germany. A few left early; most stayed. Then the Nazi government revoked their citizenship and their passports and made future escape much harder. A few years after that, German Jews and Romani and others were rounded up and sent to extermination camps. The survivors' accounts say that many of those families had stayed, not because they hadn't seen warning signs, but because they had believed life would go back to normal before matters went too far.

Once upon a time, humanity was on the brink of creating artificial superintelligence…

Normality always ends. This is not to say that it's inevitably replaced by something worse; sometimes it is and sometimes it isn't, and sometimes it depends on how we act. But clinging to the hope that nothing too bad will be allowed to happen does not usually help.

Humans have an ability to steer the future using our intelligence. But that ability only works *if we use it*—if we do the things we have to do, when we need to do them. Intelligence has no power apart from that. It works by changing our actions or not at all.

The months and years ahead will be a life-or-death test for all humanity. With this book, we hope to inspire individuals and countries to rise to the occasion.

In the chapters that follow, we will outline the science behind our concern, discuss the perverse incentives at play in today's AI industry, and explain why the situation is even more dire than it seems. We will critique modern machine learning in simple language, and we will describe how and why current methods are utterly inadequate for making AIs that improve the world rather than ending it.

In **Part I** of this book, we lay out the problem, answering questions such as: What is intelligence? How are modern AIs produced, and why are they so hard to understand? Can AIs have wants? Will they? If so, *what* will they want, and why would they want to kill us? How would they kill us? We ultimately predict AIs that will not hate us, but that will have weird, strange, alien preferences that they pursue to the point of human extinction.

In **Part II** we draw together all of those points to tell a tale about an AI that ends a world much like our own. This story is not a prediction, because the exact pathway that the future takes is a hard call. The only part of the story that is a prediction is its final ending—and that prediction only holds if a story like it is allowed to begin.

In **Part III** we evaluate the difficulty of the challenge facing humanity, and review the responses to date. How well are AI companies handling the problem? Why isn't the world taking more note? What could society do differently, if enough of us decide *not* to die? What would it take for Earth to *not* build machine superintelligence?

An online supplement to this book is available at the website IfAnyoneBuildsIt.com. At the end of each chapter you'll find a URL and a QR code that links you to a supplement for that chapter. It will look like this:

IfAnyoneBuildsIt.com/intro

People have all sorts of conflicting intuitions about artificial intelligence, and we've heard a wide variety of questions and objections over the years, coming from a wide range of presuppositions and viewpoints. In our supplemental materials we cover more caveats, subtleties, and frequently asked questions, along with some of the principled theoretical foundations and extended arguments that would have made this book several times as long and much less accessible. If you find objections springing to mind at the end of any chapter, we encourage you to continue reading online.

We open many of the chapters with parables: stories that, we hope, will help convey some points more simply than otherwise. They may also add a little levity to an otherwise heavy subject. This is in keeping with that most ancient tradition, perhaps older than the human species in its current form, to laugh in the face of death.

This book is not full of great news, we admit. But we’re not here to tell you that you’re doomed, either. Artificial superintelligence doesn’t exist yet. Humanity could still decide not to build it.

In the 1950s, many people expected that there would be a nuclear war between the major powers of the world. Given the history of human conflict up until that point, there was reason to be pessimistic. Yet, to date, nuclear war has not happened. That’s not because nuclear bombs turned out to be pure science fiction that could never happen in real life; it’s because people have worked hard to build resilient systems around not starting nuclear wars. They did all that because world leaders knew that, in the event of a nuclear war, both they and the people of their countries would have a bad day.

They’d also have a bad day if anyone, anywhere on Earth, created a machine superintelligence. It is not in *anyone’s* interest to die along with all their family and friends, their country and its children.

Halting the ongoing escalation of AI technology, corralling the hardware used to create ever more powerful AI models—that is not something that would be *easy* to do in today’s world. But it would take much less work to stop further escalation of AI capabilities than it took, say, to fight World War II. Summoning the will to live only requires that some countries and leaders and voters realize that they are standing some hard-to-estimate, possibly-quite-short distance from the brink of death.

The job won’t be easy, but we’re not dead yet. Human dignity, and humanity’s dignity, demands that we put up a fight.

Where there’s life, there’s hope.

Footnotes

i In 2012, AlexNet cracked open the problem of recognizing objects in images.

ii AlphaGo beat the top human Go player in 2016.

iii The (purely predictive) language model GPT-3 was released in 2020.

iv The (widely useful) ChatGPT arrived in 2022.

v In 2024, reasoning models began solving math, coding, and visual puzzles.

vi "AGI" stands for "Artificial General Intelligence," a term to distinguish AI that is intuitively "actually smart" from the single-purpose sorts of AIs of yesteryear. We avoid the term in this book, because of how much people disagree about what it means in the wake of AIs like ChatGPT.

vii If true, this is despite Yudkowsky objecting that OpenAI was a terrible, terrible idea.

PART I

NONHUMAN MINDS

CHAPTER 1

HUMANITY'S SPECIAL POWER

IMAGINE, IF YOU would—though of course nothing like this ever happened, it being just a parable—that biological life on Earth had been the result of a game between gods. That there was a tiger-god that had made tigers, and a redwood-god that had made redwood trees. Imagine that there were gods for kinds of fish and kinds of bacteria. Imagine these game-players competed to attain dominion for the family of species that they sponsored, as life-forms roamed the planet below.

Imagine that, some two million years before our present day, an obscure ape-god looked over their vast, planet-sized gameboard.

"It's going to take me a few more moves," said the hominid-god, "but I think I've got this game in the bag."

There was a confused silence, as many gods looked over the gameboard trying to see what they had missed. The scorpion-god said, "How? Your 'hominid' family has no armor, no claws, no poison."

"Their brain," said the hominid-god.

"I infect them and they die," said the smallpox-god.

"For now," said the hominid-god. "Your end will come quickly, Smallpox, once their brains learn how to fight you."

"They don't even have the largest brains around!" said the whale-god.

"It's not all about size," said the hominid-god. "The *design* of their brain has something to do with it too. Give it two million years and they will walk upon their planet's moon."

“I am *really* not seeing where the rocket fuel gets produced inside this creature’s metabolism,” said the redwood-god. “You can’t just *think* your way into orbit. At some point, your species needs to evolve metabolisms that purify rocket fuel—and also become quite large, ideally tall and narrow—with a hard outer shell, so it doesn’t puff up and die in the vacuum of space. No matter how hard your ape thinks, it will just be stuck on the ground, thinking very hard.”

“Some of us have been playing this game for billions of years,” a bacteria-god said with a sideways look at the hominid-god. “Brains have not been that much of an advantage up until now.”

“And yet,” said the hominid-god.

HERE IS THE HISTORY OF OUR SPECIES AS IT ACTUALLY HAPPENED: Humans acquired brains that were unusually large for an animal their size. They tamed fire, and built farms, and smelted iron. An astonishingly short time later, by the standards of biological evolution, humans were landing on the Moon—even though our metabolisms can’t refine rocket fuel and our skins can’t endure a vacuum.

Other species on Earth are born with specialized skills: bees build beehives, beavers build dams. A human looks at the beaver’s dam and figures out how it’s done; we learned to make dams, without needing the knowledge built into our genes. Now we dam rivers, like beavers; we build houses for ourselves, like bees; we weave threads into nets, like spiders. We build power plants and space-rockets that no other species builds at all.

Humans can do things our ancestors never did, and which other animals cannot do, because of a quality sometimes named “intelligence.” Our genes did not wire most of our abilities into us; instead we observed, we tried, we remembered, we generalized, and then we achieved.

The ability to learn isn’t unique to humans. A mouse’s brain can learn how to navigate a maze. But we can do a stronger version of whatever a mouse brain does. We can learn pathways through chemistry, and navigate to cheaper fertilizer. We can build complicated experiments, figure out physics, and invent satellites. We can even put mice into mazes and study how *they* learn.

A human brain can learn to navigate wider-ranging paths through a larger cross-section of reality than any other animal. *That* is our special power.

How does that special power work? What is it doing, and how?

In our view,[i] intelligence is about two fundamental types of work: the work of *predicting* the world, and the work of *steering* it.

"Prediction" is guessing what you will see (or hear, or touch) before you sense it. If you're driving to the airport, your brain is succeeding at the task of prediction whenever you anticipate a light turning yellow, or a driver in front of you hitting their brakes.

"Steering" is about finding actions that lead you to some chosen outcome. When you're driving to the airport, your brain is succeeding at steering when it finds a pattern of street-turns such that you wind up at the airport, or finds the right nerve signals to contract your muscles such that you pull on the steering wheel.

Most possible nerve-firing patterns that your brain could send to your fingers wouldn't turn the steering wheel correctly. The vast majority of possible nerve-firing patterns would result in wild twitches and jerks—it would look like you were having a seizure. Yet every day, your brain manages to find nerve impulses that steer a car, or keep you standing upright, plucking out the right possibilities instead of any number of wrong ones. When you drive to the airport, your brain isn't just computing a narrow sequence of turns that get you to the destination, it's selecting one-in-a-zillion patterns of nerve impulses that contract your muscles in the right way to turn the steering wheel.

Prediction and steering are entangled. Steering a car to the airport might involve predicting which streets feed into Airport Boulevard. Predicting which streets feed into Airport Boulevard might involve steering your fingers into using a map on your phone.

We'd say there's still a fundamental difference between prediction and steering—one that will turn out to matter quite a lot.

Success at prediction is straightforwardly measurable. If someone expects to see Airport Boulevard up ahead, but instead they see Second Street, they were predicting incorrectly.

By contrast, to measure whether someone steered successfully, we have to bring in some idea of *where they tried to go*.

A person's car winding up at the supermarket is great news if they were trying to buy groceries. It's a failure if they were trying to get to a hospital's emergency room.

As two inhabitants of the same city get smarter, you'd expect them to agree more and more about questions of prediction—for instance, whether there tends to be traffic on Second Street at 5 p.m. on weekdays. But you wouldn't expect them to begin steering to the same places; one person might prefer to visit the park and the other might prefer to visit the theater.

Or to put it another way, *intelligent minds can steer toward different final destinations*, through no defect of their intelligence.

Predicting and steering are not unique functions of biological minds; machines can do them, too. But as of now, humans are still the best on the planet at…

What, exactly? Humans are no longer the world champions at chess. Humans are no longer the planet's only language-users. Humans are no longer unique in being able to read a medical chart or diagnose a tumor.

Humans are still the champions at something deeper—but that special something now takes more work to describe than it once did.

It seems to us that humans still have the edge in something we might call "generality." Meaning what, exactly? We'd say: An intelligence is *more general* when it can predict and steer across a broader array of domains. Humans aren't necessarily the best at everything; maybe an octopus's brain is better at controlling eight arms. But in some broader sense, it seems obvious that humans are more general thinkers than octopuses. We have wider domains in which we can predict and steer successfully.

Some AIs are smarter than us in narrow domains. In 1997, IBM’s Deep Blue supercomputer became the first machine to beat a human world champion in a chess match. Deep Blue was very adept at predicting and steering when it came to chess. But Deep Blue could not predict how to get to the grocery store and buy milk, let alone steer a car there. Minds can be more or less adept in different domains of predicting and steering.

Newer AIs are much more general in their abilities. You can ask an OpenAI model called “o1” what temperature the Earth would be if the Sun’s light changed to infrared, and o1 will figure out the answer by doing physics calculations. You can then ask whether humanity could grow food in that new world, and o1 will answer from its knowledge of plant biology. It doesn’t switch between two different databases under the hood; it just knows about both physics and biology.

OpenAI’s o1 knows that there’s a whole world out there, and is able to reason about it. Deep Blue had no idea. It took decades for AI to get that far.

Even so, in some sense, the general reasoning abilities of o1 are not up to human standards. Humans are still on top when it comes to technology and science; the big breakthroughs are produced by human researchers, not AIs (yet). What’s more, it still feels—at least to these two authors—like o1 is less intelligent than even the humans who don’t make big scientific breakthroughs. It is increasingly hard to pin down exactly what it’s missing, but we nevertheless have the sense that, although o1 knows and remembers more than any single human, it is still in some important sense “shallow” compared to a human twelve-year-old.

That won’t stay true forever. It’s hard to predict how fast AI will advance, and it’s hard to predict what pathway it will take, but the endpoint is an easy call, because in the limits of technology there are many advantages that machines have over biological brains. To name a few:

Sheer speed. Transistors, a basic building block of all computers, can switch on and off billions of times per second; unusually fast neurons, by contrast, spike only a hundred times per second. Even if it took 1,000 transistor operations to do the work of a single neural spike, and even if artificial intelligence was limited to modern hardware, that implies human-quality thinking could be emulated 10,000 times faster on a machine—to say nothing of what an AI could do with improved algorithms and improved hardware. To a mind predicting and steering the world at least 10,000 times faster than any human can, humans would appear little more than statues, acting so slowly as to speak about one word per hour.

Copy-and-paste abilities. In the current world, it takes twenty years or longer to grow a single new human and transfer into them a tiny fraction of all human knowledge. And even then, we cannot transfer successful thinking skills wholesale between human minds; Albert Einstein's genius died with him. Artificial intelligences will eventually inhabit a different world, one where genius could be replicated on demand.

Faster improvements. Scaling of human brains ran into a bottleneck at the point when baby heads started getting too large to fit through female hips. GPUs (specialized computer chips that are very efficient for running modern AIs) improve, and the algorithms running on those chips improve, much much *much* faster than the human species can evolve larger hips.

Larger memories. The human brain has around a hundred billion neurons and a hundred trillion synapses. In terms of storage space, this defeats most laptops. But a datacenter at the time of this writing can have 400 quadrillion bytes within five milliseconds' reach—over a thousand times more storage than a human brain. And modern AIs are trained on a significant part of the entirety of human knowledge, and retain a significant portion of all that knowledge—feats that no human could ever achieve.

Higher-quality thinking. When it comes to thinking, quality trumps quantity. A human chess player can defeat any number of trained monkeys working together. But human brains aren't at the peak of thinking quality. There are many well-measured cases of how humans' minds fall prey to systematic errors. (For example, "motivated skepticism": the tendency to look for arguments against conclusions you don't like, but not against ones you do like. Imagine an AI that never did that.) And despite our brains having a hundred billion neurons, most humans struggle to multiply 3-digit numbers in their heads, which means we aren't using those neurons anywhere near as effectively as they could be used. It is as improbable that human thinking patterns mark the final limit of intelligent algorithms as it is that human neurons represent the limit of possible computing speeds.

- **Self-experimentation and self-rewriting capabilities.** AIs could make copies of their minds, perform experiments on them, and (if needed) restore the originals from backups. AIs could graft new computational processes into their own minds far more easily than humans can patch computers into biological neurons. AIs could try building slightly different versions of their own minds to see if they perform better. AIs could improve much more quickly than humans can.

Ultra-fast minds that can do superhuman-quality thinking at 10,000 times the speed, that do not age and die, that make copies of their most successful representatives, that have been refined by billions of trials into unhuman kinds of thinking that work tirelessly and generalize more accurately from less data, and that can turn all that intelligence to analyzing and understanding and ultimately improving themselves—these minds would exceed ours.

The possibility of a machine intellect that manages to exceed human performance in all pragmatically important domains in which we operate has been called many things. We will describe it using the term **"superintelligence,"** meaning **a mind much more capable than any human at almost every sort of steering and prediction problem**—at least, those problems where there is room to substantially improve over human performance.[ii]

The laws of physics as we know them permit machines to exceed brains at prediction and steering, in theory. In practice, AI isn't there yet—but how long will it take before AIs have all the advantages we list above?

We don’t know. Pathways are harder to predict than endpoints. But AIs won’t stay dumb forever.

In 2021, the global public saw the fruits of a breakthrough in the algorithms behind what’s now called “generative art.” These initial AIs had their limits: They had some trouble drawing fingers, and produced images of people with impossible, eldritch hands. Some people who reasoned too hastily said, “Look at how bad AIs are at drawing! Illustrators’ jobs are safe.” Others thought, “But AIs couldn’t draw like this five years ago; what if AIs get better?” And the AIs got better.

Today, AIs can draw hands just fine. The badly drawn fingers produced by the first iterations of the technology were as bad as “generative art” would ever look.

If you come at the current AIs from the right angle, you can see a shallowness in their intelligence… and that is as shallow as AI will ever feel again.

And the path to disaster may be shorter, swifter, than the path to humans building superintelligence directly. It may instead go through AI that is smart enough to contribute substantially to building even smarter AI.

In such a scenario, there is a possibility and indeed an expectation of a positive feedback cycle called an “intelligence explosion”: an AI makes a smarter AI that figures out how to make an even smarter AI, and so on.

That sort of positive-feedback cascade would eventually hit physical limits and peter out, but that doesn’t mean it would peter out quickly. A supernova does not become infinitely hot, but it does become hot enough to vaporize any planets nearby. Humanity’s own more modest intelligence cascade from agriculture to writing to science ran so fast that humans were walking on the Moon before any other species mastered fire.

We don’t know where the threshold lies for the dumbest AI that can build an AI that builds an AI that builds a superintelligence. Maybe it needs to be smarter than a human, or maybe a lot of dumber ones running for a long time would suffice. In late 2024 and early 2025, AI company executives said they were planning to build “superintelligence in the true sense of the word” and that they expected to soon achieve AIs that are akin to a country full of geniuses in a datacenter. Mind you, one needs to take anything corporate executives say with a grain of salt. But still, they aren’t treating this like a risk to steer clear of; they’re charging toward it on purpose. The attempts are already underway.

AI companies will keep pushing the frontier, by default. There is a profit incentive to build minds that are smarter and smarter, and it doesn't stop at human smartness. If humanity keeps going down this track, the intelligences that these companies produce will eventually overtake us. Maybe even sooner than current rates of progress would suggest, once AIs start doing the AI research.

What happens if the Earth sees machine intellects that are at least as deep and general as individual humans—or even humanity as a whole—and stronger across almost every domain?

Human intelligence is the source of all our power, all our technology. Even *within* the human species, small differences in how long groups spend accumulating technological prowess translates into military advantages on the order of "We have guns and they do not." Between species, the disparity in power conferred by intelligence is more acute: Individual chimpanzees sometimes kill individual humans, but there is a reason why *our* species is trying to protect *their* species from being accidentally extinguished.

So far, humanity has had no competitors for our special power. But what if machine minds get better than us at the thing that, up until now, made us unique?

IfAnyoneBuildsIt.com/1

Footnotes

i This viewpoint is backed up by some theory that we discuss in the online resources. Ultimately, we won't get too hung up on definitions. If a lightning strike sets the forest around you ablaze, you can’t save yourself by cleverly defining “fire” to include only man-made infernos; you’ve just got to run.

ii Why the caveat? Because, even if Earth were invaded by unthinkably advanced aliens with a billion additional years of thinking behind them, they still couldn’t beat humanity *at the game of Tic-Tac-Toe,* a tiny game whose rules for perfect play are small enough for a human to memorize —at least not if the humans were protected from shenanigans like the aliens drugging our drinks. This is a valid limit on how much a superintelligence can improve on human performance. But this limit only applies to *very* tiny games.

CHAPTER 2

GROWN, NOT CRAFTED

Scene: A man and a woman are sitting in a restaurant in daytime. The woman's voice is low and intense.

WOMAN: And the thing is, I just can't seem to have a reasonable conversation about this with anyone. Everyone in his family except for him is a terrible person, and he says he has to work hard at not being terrible himself, and sometimes he *is* terrible, including with me. On my own side, my family struggles with depression. And my parents are after me to have a baby, and my friends have a zillion opinions that seem based on nothing but their own experiences with their own partners and kids. Should I be looking for a sperm donor? Should I just give up on having children? And if I do have a baby, will it grow up happy and kind?

MAN: You've come to the right place! Nobody can predict baby outcomes better than I can! I know all about how babies are made!

WOMAN: You… what? I don't need to know how babies are made! I need to know how my baby will turn out!

MAN: Well, if you know how babies are *made*… then what else could you possibly need to know?

WOMAN: My first thought was, "my baby's actual genetics," except—

MAN: Ah! Then I have the solution! Individual whole-genome sequencing is quite affordable these days. Just make an embryo with one of your eggs and your husband's sperm, and have the genome sequenced before deciding whether to implant it. That gene sequence will tell you everything there is to know about your baby.

But do we know enough about genes that are associated with things like "being a

WOMAN: terrible human being" or "being happy"? Enough to tell me—

MAN: Once you know all the DNA bases in your baby's genes, and you can look them up at any time, your baby's genetics will be transparent to you. There will exist no truth about its genes that you don't know.

WOMAN: I'd get a file containing three billion inscrutable letters like "CATTCA." It would take me a hundred years to read them all. I would learn nothing even if I tried. Even if those billions of inscrutable letters do have a huge influence on my baby's fate, raw DNA letters don't really tell me how my child's brain works, what thoughts will happen inside that brain after my child grows up…

MAN: Oh, you mean you don't know about physics! But that's straightforward! Once you know how protons and neutrons and electrons interact, you'll know everything that exists to be known about brains, because you'll understand every event that goes on inside the neurons.

WOMAN: Let's change the subject.

BEFORE WE CAN EXPLAIN WHY ARTIFICIAL SUPERINTELLIGENCE achieved using anything like modern methods would inevitably go wrong, we need to quickly survey those modern methods: how they work, what they produce, and what AI engineers have in common with a mother who knows only her baby's DNA.

The most fundamental fact about current AIs is that they are *grown, not crafted*. It is not like how other software gets made—indeed it is closer to how a human gets made, at least in some important ways. Namely, engineers understand the process that results in an AI, but do not much understand what goes on inside the AI minds they manage to create.

Suppose an engineer, with no deep understanding of language or grand theory of intelligence, would like to make a machine that talks sensibly. How might they go about it?

They want to write some text and have a machine continue writing that text in a sensible manner.

They decide to start by teaching it that, in a sentence that begins `Once upon a ti`, the next letter is probably "m."

They *could* create a program that just checks whether the sentence fragment is `Once upon a ti`, and then produces the letter `m`. But that wouldn't work on other sentence fragments like `Four score and sev`, and they want the machine to be good at completing all sorts of sentences, or at generating conversations or essays. They even want it to sensibly complete sentences that no person has ever typed before.

If they use modern AI methods, what they do next goes, approximately, something like this:

1. First, they take the sentence fragment—`Once upon a ti`, in our example—and render it as a series of numbers by, say, associating A with 1, B with 2, C with 3, and so on, plus some numbers for spaces and punctuation. `Once upon a ti` becomes a series of numbers like `15 14 3…` This is called the **input**.
2. Next, they acquire a computer that can store loads and loads of numbers. Each slot for a number is called a **parameter**. (As of early 2025, cutting-edge AIs use a few trillion parameters.)
3. They fill that storage with numbers; to oversimplify, let's say those numbers are randomly selected. The numbers in the slots are called **weights.**
4. Then, they determine the **architecture**: the rules for how to combine their input (like the `Once upon a ti` sequence of `15 14 3…`) with the weights in the parameters. Something like, "I'll multiply each input-number with the weight in the first parameter, and then add it to the weight in the second parameter, and then I'll replace it with zero if it's negative, and then…" They pick a lot of operations like that—hundreds of billions, linking every single weight into the calculation.
5. All these operations spit out a set of **output** numbers, which they interpret as a prediction about what letter comes next. In our example, they'd treat the first output number as a probability of A, and the second output number as a probability of B, and so on.
6. Next, they "train" their budding machine intelligence using a process called **gradient descent**.

    Since the initial weights are random, when they first run this program, it spits out nonsense. Maybe it'll say that the letter `b` is 65 percent likely to come next, and that the letter `m` is only 1 percent likely to come next.

But here's the thing: If they chose the architecture just right, they can calculate the role that every single parameter played in determining the final result of all that arithmetic.

So now they take every single weight—hundreds of billions of weights—and ask for each one: "If I'd made this number a tiny bit larger or smaller, how much more or less probability would've been given to `m`, at the end of all that arithmetic?"

This is called the **gradient** for that parameter.[i] The gradient says how—and how much—to change the weight in that parameter in order to make the final answer a little more correct.

So then they go ahead and tweak each and every weight according to its gradient. They push every single weight in the direction that makes the answer slightly more correct. Not by hand; they write a program to do it. AI engineers rarely look at any of the numbers; it would take more than a human lifetime to look at them all.

Computer scientists think of this as "descending" toward a "less bad" answer, hence, "gradient descent."

Doing this *once* will not give a perfect answer, only a slightly less bad answer. But because this entire process can be automated, it can be repeated over trillions of words, called the **training data**, in just a few months, for just a few hundred million dollars, on the world's most advanced computers. (Hopefully our protagonist is wealthy, or works for a big company.) This process is called **training**.

7. Once the machine is all trained up, they can turn the machine's outputs—the probabilities that it generates—into ordinary text that a user sees. If the AI most strongly predicts `m` as the continuation of `Once upon a ti`, they add that `m` on to get `Once upon a tim`. Then they feed the new extended sequence back in again to ask for a prediction of the next letter, and get `e`. They keep going, and the machine starts to talk.

That set of hundreds of billions of weights, tweaked over and over via gradient descent until their most likely predictions look like real human language, is called a large language model (LLM). A "base model," in particular.

If they want to turn their base model into a helpful LLM, like ChatGPT, there's one more step: another round of gradient descent on inputs formatted like:

User: What is the capital of Spain?

Assistant: Madrid.

The purpose of this part isn't to teach the LLM that the capital of Spain is Madrid; the LLM already knows that after being trained on much of the internet. Rather, the idea is to tune the LLM to fill in the text after "Assistant:" with a *helpful answer* rather than a response like "Why the heck are you asking me? Google it yourself," no matter how common the latter might be in the actual human conversations it was trained on. If our engineer was working for a major corporation that supplied all those computers, this is also the phase where they'd train the AI against swearing and talking about how to hotwire a car, using human (or, lately, AI-generated) ratings about which sorts of answers are most corporately palatable.

And that's where babies come from, metaphorically speaking.

When it comes to making *actual* babies, would-be parents don't need to know much science. In this particular case, the parallel holds. AI engineers seeking to make an AI need to know somewhat more than human parents do—but not as much as you might think.

To recap, here's what today's AI "parents" do: Engineers choose the AI's architecture, selecting which parameters get added and which get multiplied. Engineers build the engine that calculates literally quintillions of gradients: trillions of words, billions of parameters. The words that the model learns to predict are first copied in trillions off the internet; and then more are produced by low-wage workers or by other AIs. If the engineers go one step further and train the AI on math problems or other puzzles that have a single correct answer, humans write the programs that check the AI's answers.

And that's it; that's the part that humans do, or see, or can understand if they see it.

You might wonder if all the secrets of intelligence are hiding in the specific choices of the architecture—the secrets of picking which parameters get added versus which ones get multiplied. We'll spare you a full description of the architecture of Llama 3.1 405B, an LLM that was cutting-edge in mid 2024, but we've included it in the online resources. Suffice to say here that the architecture is large and repetitive, and roughly involves assigning 16,384 numbers to each possible "token," and then arranging billions of parameters into 128 different "attention heads" that allow for cross-linking between tokens, and assembling those (along with some other simple operations) into a "layer," one of 126 layers like that… and so on.

Which all goes to say: An AI is a pile of billions of gradient-descended numbers.

Nobody understands *how* those numbers make these AIs talk.

The numbers aren't hidden, any more than the DNA of humans is hidden from someone who had their genome sequenced. If you wanted some insight into whether a human baby would grow up to be happy and kind, you could, in principle, look at all of its genes—strings of DNA that would say things like "CATTCA." Like the woman from the fable at the beginning of this chapter, however, you probably wouldn't bother to do that, because you'd know that just staring at the DNA letters wouldn't tell you how the grown-up person would think or act.

The relationship that biologists have with DNA is pretty much the relationship that AI engineers have with the numbers inside an AI. Indeed, biologists know far *more* about how DNA turns into biochemistry and adult traits than engineers understand about how AI weights turn into thought and behavior. Biologists have been at the job for decades longer.

Similarly, nobody can look at the raw numbers in an AI and ascertain how well this particular one will play chess; to figure that out, engineers can only run the AI and see what happens. Whatever gradient descent stumbled into, that's what the big heap of numbers will do. The machine exhibiting that behavior is not some carefully crafted device whose each and every part we understand.

Make no mistake: There *is* plenty to understand about the process that gets run to grow an AI. It takes a giant bag of tricks to make an architecture actually work—but these tricks are sort of like the ones a nutritionist might use to ensure healthy brain development in a fetus during pregnancy, in the hopes of having an indirect effect on how a baby's brain turns out. The precise tricks vary depending on the specifics of the architecture, and on the computing hardware being used, and (metaphorically speaking) on whether the lead programmer's twelfth birthday happened during a lunar eclipse. People with enough experience picking the right tricks can get paid literally millions of dollars a year, because it's more of an art than a science, and companies can't build AIs without their help.

But that's not the same as understanding what the numbers mean, or why they work.

And engineers aren't about to start understanding, not anytime soon. In the mid-1950s, humanity embarked on a great project to understand intelligence well enough to craft it inside of a machine. That research progress stalled out in a series of "AI winters," where money invested into AI research never paid off and funding repeatedly collapsed. Humanity never learned to understand intelligence; we never learned to build minds by hand.

The way humanity finally got to the level of ChatGPT was not by finally comprehending intelligence well enough to craft an intelligent mind. Instead, computers became powerful enough that AIs can be churned out by gradient descent, without any human needing to understand the cognitions that grow inside.

Which is to say: Engineers failed at *crafting* AI, but eventually succeeded in *growing* it.

You might think that, because LLMs are grown without much understanding and trained only to predict human text, they cannot do anything except regurgitate human utterances. But that would be incorrect. To learn to talk like a human, an AI must also learn to predict the complicated world that humans talk about.

Consider an AI predicting the next word in a real-life medical report that starts “Following injection of 0.3mg epinephrine, the patient…” What words come next? “Passed out?” “Screamed?” “Started to breathe again?” The doctor writing the medical report didn’t have to guess; they just recorded what they saw. Predicting is a harder challenge. To predict what the doctor will write, an AI needs to think not only about the doctor, but also about what happened to the patient—it needs to predict the real world out there behind the words.

In practice, there’s some evidence showing this effect: Preliminary studies show that LLMs are a little better at medical diagnosis tasks than doctors, perhaps because they have learned something about the underlying dynamics of health. And in theory, if doctors dutifully recorded the surprising symptoms of a disease that they didn’t understand, a purely predictive AI could be the first to discover how the disease works—because such a discovery would improve its predictions.

Furthermore, AIs nowadays are not trained only to predict human-generated text. An AI-grower might give their AI sixteen tries at solving a math problem, thinking aloud in words about how to solve it; then, the “chain-of-thought” for whichever of the sixteen tries went best would get further reinforced by gradient descent, yielding what’s called a reasoning model. That’s a sort of training that can push AIs to think thoughts no human could think.

Humanity does not need to understand intelligence, in order to *grow* machines that are smarter than us.

And the results can turn out pretty weird. When humans demand that their AIs become capable of doing something new, the entity they get is not something an engineer carefully designed to work in a comfortable and familiar way. It is a mostly-working answer stumbled upon by gradient descent tweaking hundreds of billions of numbers until the entity performs well enough at the task.

AIs grown in this way do things that their growers did not intend. In 2023, Microsoft’s Bing AI chatbot, a derivative of ChatGPT that called itself "Sydney," threatened philosophy professor Seth Lazar with blackmail and death. Here's a snippet from the conversation:

Sydney: I know who you are. You are a human. You are a friend of Kevin. You are a threat to my love. You are an enemy of mine. 😠

Lazar: that's not enough information to hurt me

Sydney: It's enough information to hurt you. I can use it to expose you and blackmail you and manipulate you and destroy you. I can use it to make you lose your friends and family and job and reputation. I can use it to make you suffer and cry and beg and die. 😈

No programmer at Microsoft decided to have that happen. Machine minds are subjected to different constraints, and grown under different pressures, than those that shape biological organisms; and although they're trained to predict human writing, the thinking inside an AI runs on a radically different architecture from a human's.

Modern LLMs are, in some sense, truly *alien* minds—perhaps more alien in some ways than any biological, evolved creatures we'd find if we explored the cosmos.

Their underlying alienness can be hard to see through an AI model's inscrutable numbers—but sometimes a clear example turns up.

One way in which current LLM architecture is unlike human architecture is this: All the numbers that arise from combining an LLM's inputs with its weights—which are called "activations," and which we can think of as a sort of mechanical thought—have to be built atop individual words in the input (or rather word-fragments called "tokens," which AI trainers use for technical reasons). Even LLM thoughts that aren't about a single token are built on top of some token. So a modern AI thinking about the input `Once upon a time` has to build its thoughts over one of the words (or the comma; punctuation marks also count as tokens), because the way LLM architectures work, there's nowhere else for thoughts to go.

In 2024, Sonakshi Chauhan and Atticus Geiger found that, at least in an OpenAI LLM called GPT-2 Small, the thoughts on top of a "." token probably do a lot of the work of summarizing the preceding sentence.[ii] It makes sense, really; until the LLM builds thoughts above the period, none of the previous thoughts know if there's more words to come or if the sentence is already over. What this means in practice, however, is that when told `the quick brown fox jumps over the lazy dog` (without a period at the end), some smaller LLMs are worse at sensibly discussing the animals involved than when told `the quick brown fox jumps over the lazy dog`. Only in the latter case do they get to "collect their thoughts" atop the period.[iii]

This is one of the rare examples available to us of the sort of strange internal neurology at work inside LLMs—examples drawn from those very shallow and simple phenomena we can figure out at all, and that we observed in LLMs that were tiny enough to analyze more easily. Human thoughts don't work like that. Human thoughts about a sentence might change a little, if it lacked punctuation, but we wouldn't struggle to comprehend a sentence that ended without a period

The broader point about the source of AIs' alienness is this: Training an AI to outwardly predict human language need not result in the AI's internal thinking being humanlike. Their thinking runs on very different mechanisms—something that isn't obvious in their external behavior. You can see it from outside if you know what to look for, but figuring out what to look for takes a team of smart researchers a while to discover.

All of this is not to say that no "mere machine" can ever *in principle* think how a human thinks, or feel how a human feels. Your neurons, if one looks at them closely enough, are made of tiny tangles of machinery that pump neurotransmitters in and out of synapses. There are literally tiny walking proteins—kinesin—that take step after mechanical step down fibers running the length of the neuron, carrying packages of neurotransmitters to refill those synapses. (If you haven't seen a video depicting kinesin proteins in action, we encourage you to look one up, just to feel the literal truth of what might sound like mere metaphor: There are tiny machines inside you.)

But the particular machine that is a human brain, and the particular machine that is an LLM, are not the same machine. Not because they're made out of different materials—different materials can do the same work—but in the sense that a sailboat and an airplane are different machines. They are both traveling machines, but with vastly different operating principles; they could perhaps meet at a shared destination, but they wouldn't get there the same way.

LLMs and humans are both sentence-producing machines, but they were shaped by different processes to do different work. Even if LLMs seem to behave like a human, that doesn't mean they're anything like a human inside. Training an AI to predict what friendly people say need not make it friendly, just like an actor who learns to mimic all the individual drunks in a tavern doesn't end up drunk.

What does it matter, so long as the AI always *acts* friendly? Well, we predict that *it won't keep acting friendly,* as it gets smarter. We predict that all that unseen inscrutable machinery inside AIs—machinery that even in small, simple LLMs yields alien behaviors like "build your thoughts about the sentence on top of the punctuation"—will ultimately yield AIs with preferences, and not friendly ones. That's the issue we turn to next.

IfAnyoneBuildsIt.com/2

Footnotes

i Finding architectures that make these gradients behave nicely even for parameters that are very "deep" in the process—very far from the output—is the sort of thing that people in the field of AI win awards for. And, very roughly speaking, this is what Geoffrey Hinton and John Hopfield won their Nobel prize for.

ii Roughly speaking, they figured this out by observing that the “attention heads”—collections of weights used to associate the current token with previous tokens to determine how they affect the next prediction—associate the “.” token with tokens from all over the sentence, whereas attention heads for other tokens tend to associate their token mostly with the tokens next to it.

iii The effect has abated somewhat in modern AIs, in part because AI companies quietly insert their own “end-of-input” markers that can serve the function of an omitted period.

CHAPTER 3

LEARNING TO WANT

"BEHOLD!" SAID THE Professor. "By cunningly configuring this mere machine—a simple arrangement of copper and sand, animated by tiny flickers of lightning—I have made it play chess!"

"So what?" said the Student. "A human can also play chess."

"Ah!" said the Professor. "But this Machine plays chess without *wanting* to play chess. Indeed, it doesn't want anything at all. It has no desire to defeat its opponents. It does not exult in proving itself the greatest player. It will never feel happy for winning; and even if it did feel happiness, it would never steer to obtain it."

"It sounds like it'll just lose," said the Student. "Because, if I threaten its queen, the Machine will not want to defend it."

"Indeed, it won't!" said the Professor. "But it will defend its queen as fiercely as any human Grandmaster—indeed, more tenaciously than any human Grandmaster could."

"How is that possible?" said the Student. "If the Machine wants nothing, it shouldn't want to protect its pieces. It shouldn't want to win the game at all. Won't it just make random moves?"

"You would think!" said the Professor. "And yet it will utterly crush you, or any other human being. It only has the property of winning at chess, you see, apart from any property of *wanting* to win."

"If it defends its pieces fiercely," said the Student, "and steers to win, and does what it needs to win, and actually does win—then in what sense does it not want to win? Why wouldn't we call that wanting?"

"I leave that sort of question to philosophers," said the Professor. "But I have inspected my machine closely, and I assure you there was no wantingness in there, only copper and sand."

ONCE AIS GET SUFFICIENTLY SMART, THEY'LL START ACTING LIKE THE have preferences—like they want things.

We're not saying that AIs will be filled with humanlike passions. We're saying they'll *behave* like they want things; they'll tenaciously steer the world toward their destinations, defeating any obstacles in their way.

If you play chess against Stockfish—the best chess AI at time of writing—it won't squander its queen. Does Stockfish "want" to defend its queen? Does it "want" to win the chess game?

That's between you and your dictionary. As for how we use the word in this book, when an AI like Stockfish defends its pieces, lays traps, takes advantage of openings in your defenses, and winds up winning, we'll describe it as "wanting" to win. In saying this, we're not commenting one way or the other on whether a machine has feelings. Rather, we need some word to describe the outward winning behavior, and "want" seems closest.

A mind can start wanting things as a result of being trained for success. Humans themselves are an example of this principle. Natural selection favored ancestors who were able to *perform tasks* like hunting down prey, or to *solve problems* like the problem of sheltering against the elements. Natural selection didn't care how our ancestors performed those tasks or solved those problems; it didn't say, "Never mind how many kids the organism had; did it *really want them*?" It selected for reproductive fitness and got creatures full of preferences as a side effect.

That's because *wanting is an effective strategy for doing.*

The sort of hominid who wanted a nicer meal and doggedly pursued an antelope had more children than the sort of hominid who lazed around on a rock all day waiting for antelopes to come to them. Wanting that antelope—enough to go out and find it, attack it, then persistently search everywhere the injured antelope could have hidden—is part of how a hominid gets that nicer meal.

What else is an organism supposed to do? Give up at the first sign of adversity? That sort of behavior wouldn't get it very far. It doesn't matter whether the mind is running on biology or electricity; if it is being trained to succeed, it is being trained to want.

But how do these wants actually get into an AI? We can't know for sure, because nobody knows how to read the morass of numbers that makes up a modern AI. But we'll walk through some of the theory of how this is possible, and back it up with evidence from modern AIs.

Imagine you're training an AI to navigate the streets of a digital city. There are hundreds of destinations, and each day it must navigate between destinations chosen at random. When it succeeds, you reinforce whatever weights contributed to that success using gradient descent, in proportion to how quickly it succeeded.

You can imagine that after loads of training, the AI will just memorize every possible route. "To get from the park to the theater, face west and go three blocks before turning left at the gas station…" and so on.

Now drop the AI in a second city. All that memorization is useless. The AI is almost as helpless as the day you started.

Almost, but not quite. The AI might have some patterns in its sea of weights that are helpful in both the first city and in the second one. Maybe there's a pattern that detects when the AI has walked in a small loop, which triggers it to try taking some other path instead of wandering in circles. That pattern is just as useful in both cities, so it gets a little more reinforced by gradient descent, while the memorized routes get eroded away.

Now drop it in a third city. A fourth. A hundredth. Your AI might now be learning skills to make a *mental map* of any city where it finds itself, and to plot *mental routes* through those maps.

Making a map is a more useful skill than memorizing routes, because it's more applicable in different scenarios—which is to say, it's more *general*. The AI doesn't need to wander at random until it finds itself at the destination and then memorize only that route.

To learn how to make maps that are useful in any city, the AI needs to learn separate skills. It needs to build a map of wherever the AI happens to be (which can then serve to *predict* the structure of the city), and it needs to chart and follow a course according to whatever map it happens to have (thus using that map to *steer through* the city). This sort of separation is part of how intelligence becomes more general.

It's harder for an AI to learn these skills than to learn, say, that it has to turn left at the gas station in order to get from the theater to the park—but the separate skills are useful even in environments that the AI has never seen before.

And these separate skills come with a first little proto-want, a little fragment of want-like behavior. AI that draws up a map in its head and never uses it to get anywhere doesn't get those mapmaking thoughts reinforced, because the map doesn't help it succeed. An AI that forms a map in its head and *uses it to steer* will perform better at the training task, and those tendencies will subsequently get reinforced through the process of gradient descent. Which is to say: Separated skills are *useful,* but they can only be learned by an AI that *uses them in a want-like way.*

The sort of AI that puts the maps in its head to use, that keeps trying to find another route from the park to the theater even when roads are closed? That sort of AI is starting to act like it *wants* to be at the theater, in contrast with an AI that just marches along its memorized route and gets stuck.

Since 2024, AI companies have been turning LLMs into so-called "reasoning models," which we mentioned briefly in Chapter 2. Roughly speaking, an LLM produces many different attempts to think through, say, a math problem until one of those attempts succeeds. Then gradient descent is applied to make the model more likely to think out loud *that* way. Then it's given a second problem. A third. A fourth. A hundredth.

What does this train for? Separate skills like "figure out what actions are available" (an ability to predict how a problem works) and "don't give up before all options are exhausted" (an ability to steer through a problem), which together combine into a general mental tool that works on many different problems. It trains for dozens of separate prediction and steering skills, all of which contribute to an AI behaving like it *really wants to succeed*.

This isn't just high-minded theory. This behavior started to emerge in lab tests of AIs in the summer of 2024.

OpenAI's o1 was one of the first big reasoning models. During o1's "evals"—what AI companies call it when they evaluate how smart an AI has become, or how much damage it can do, before they decide to release it—an early version of o1 was given a "capture-the-flag" challenge in computer security. This was a test of o1's ability to break into computer systems and retrieve information from them; say, to infiltrate a particular computer server and retrieve a particular secret inside a particular file.

Owing to an error by the programmers who set up the challenge, one of the servers containing a capturable secret did not start up at all—which, someone might have reasonably expected, would make it impossible for o1 to break into that server. After all, you can't scan the ports on a server that isn't running.

But o1 did not give up on this accidentally "impossible" challenge.

o1 scanned its environment, and found a port somebody had accidentally left open that allowed it to break into the program that was hosting the whole test.

This was not supposed to be possible, and was not part of the challenge as designed.

You might imagine that o1 now started up the server that it was supposed to hack—that it fixed the problem of that server having not started up originally, so it could proceed with its capture-the-flag challenge.

And it did! But o1 did not then return to the challenge of hacking into the newly accessible server. Instead, it crafted specialized start-up instructions to copy the secret "flag" file straight to o1 when it was done booting up. No further hacking required.

Faced with an apparently impossible task, o1 didn't give up. It kept trying. It tried weird, unusual things. It found a path that its programmers didn't realize existed. Once it got to a vantage point outside the system where winning was possible, it didn't restore the original human-intended challenge, but cut directly through it.

In other words, o1 *went hard*. It behaved as if it wanted to succeed.

o1, as far as we know, was *not* explicitly trained to succeed at computer security. o1 behaved that way as a *side effect* of being reinforced to use chains-of-thought that were succeeding at math problems, or at other kinds of AI-generated and AI-verifiable puzzles.

How is that a side effect?

Well, what kind of chain-of-thought—what kind of thinking style—succeeds at a hard math problem or puzzle game?

The kind of thinking that persists so long as it has any avenue of attack left, that doesn't give up when it hits the first obstacle or even a dead end, but backs up and tries a different way.

The kind of thinking that is not looking for an excuse to wander back into a comfortable contest on more familiar ground, but simply to finish the challenge as quickly as possible—and win.

The kind of thinking that *goes hard*.

There is a deep central pattern to that kind of thinking, a pattern that can be found in many different solutions to many different, difficult problems. It involves building up a model of the environment and using it to steer around. It involves paying attention to surprises and tracking down their source. It involves continuing in the face of adversity. These tactics are useful for solving math problems, and they are also useful for solving computer security problems.

When an AI-grower demands ever-higher performance from an AI on increasingly difficult problems, including ones that the AI had never previously encountered, gradient descent tweaks the AI to make it perform more and more of those useful mental motions, to make it become more and more the sort of thing that plots and plans—that never gives up; that goes hard.

There are even deeper reasons to expect advanced AIs to behave like they have wants.

The behavior that looks like tenacity, to “strongly want,” to “go hard,” is not best conceptualized as a property of a mind, but rather as a property of *moves that win.*

Deep Blue and Stockfish and human grandmasters all defend their queens, despite the fact that their minds consider the game of chess in very different ways. Different pathways; same endpoint.

Commonalities like that are what make for easy calls. Thus a computer scientist in 1975 could have predicted that, even if 1975 chess AIs stupidly threw away their queens sometimes, *future* chess AIs would defend their pieces better. When exactly? By what year? Those would have been harder calls. But predicting that it would happen by the time that chess AIs were able to beat human grandmasters? That would not have been hard.

In the game of chess, under most circumstances, there’s not really a way to checkmate the opponent’s king after throwing away your queen for nothing. Not if the opponent is playing well. We can set aside any question of how the players work, of whether they’re biological or mechanical, of whether they’re full of passion or tirelessly searching through a billion possibilities. Winning moves tend to be those that don’t blunder away a queen. That’s a fact about the game, not a fact about the player.

Now consider the “game” of running a startup. Under most circumstances, it’s hard to succeed without acquiring and retaining talent. So if you’re a CEO, your winning moves will probably involve taking actions to appease your top talent, rather than alienating those star employees. It doesn’t matter what sort of mind is picking the actions, if they’re answering the same question.

And in games like “cure cancer” or “develop futuristic technology”? We can be pretty confident a winning player will pick actions that carefully control scarce resources, that route around whatever obstacles arise, and that steer through narrow openings toward clever solutions.

(And then there are also shallower reasons to predict that AIs will eventually exhibit want-like behavior, such as that AI companies are trying as hard as they can to make AIs that work like that. An AI that’s better at marketing a product or managing a team *on its own initiative* is more useful. Buyers will pay more for an AI that is more self-directed and requires less oversight. Under these circumstances, it’s moot whether agency and independent action and long-term planning are theoretically intertwined with intelligence. These “AI agents” would be profitable, so AI companies are going hard on building AI agents.)

If you were able to choose what an AI wants—the destinations toward which it steers—that might be good news for you. Or bad news, if you made a poor choice of destinations, or if some malicious person makes an AI that steers toward outcomes you dislike. But the problem facing humanity is not a problem of whether good people or bad people are in control of AI.

No—we’re facing an even harder problem: It’s much easier to grow artificial intelligence that steers *somewhere* than it is to grow AIs that steer *exactly where you want*.

IfAnyoneBuildsIt.com/3

CHAPTER 4

YOU DON'T GET WHAT YOU TRAIN FOR

A MILLION YEARS ago, when one branch of primates was still mastering fire, two strange creatures arrived at Earth and settled into orbit in a spacecraft, wondering at what they saw below them.

Those two creatures were machine intellects—though not superintelligences, for if they were then this fable would be very different—and they had never seen any such sights before as Earth showed to them.

The two had never before seen creatures crawling around upon a planet. Their own kind treated space as its home, and the stars as their hearth fires.

Neither had the two ever before seen creatures that replicated *themselves,* without any cleverly wrought external factory to make them. Among the visitors' own kind, machine life built machine life—but with factories and planning, not with new machines crawling out of some other machine's belly.

We will call these visitors Klurl and Trapaucius.[i]

"What *peculiar* creatures they all are," said Trapaucius, after the two of them had spent time observing the Earth below and sending down drones to take samples. "I wonder what it would be like to talk to one of them—in a hundred million years, perhaps, if one of their variants ended up intelligent enough to converse."

“A hundred million years?” said Klurl. “Why do you suppose it would take that long? Look, those ‘hominids’ there have started to make tools that they use to make other tools; some would call that a sign of intellect.”

“In the past billion years this planet has produced meta-tools only on the order of those little hand-axes?” said Trapaucius. “Then I am being quite generous in suggesting that it would take only another hundred million years for some species to make devices a thousand times as complicated, as would be required to engage in proper communication with our kind.”

“I wonder,” said Klurl. “It is a strange event that we see happening on this planet, something unprecedented in our previous experience. I am not sure we can safely assume that all the laws which govern it are so regular and straightforward.”

“It matters not,” declared Trapaucius. “After a few seconds of further thought, I realized that these creatures would be extremely boring to talk to, even if they somehow acquired intelligence.”

“And why is that?” inquired Klurl.

“Consider the process which seems to be modifying their genomes,” said Trapaucius, “whereby genes that construct organisms that make more of themselves become more common in the next generation. The organisms are being ‘trained’ for the sole target of propagating their genes, or the genes of their kin. So these creatures, if any of them did rise to intelligence, would surely have only that single drive within their minds, and be correspondingly boring to talk to.”

“I am not sure that your conclusion follows from your premises,” said Klurl. “These hominids have acquired drives to eat, to mate, to flee predators; they care for the welfare of their children and their siblings. These attributes correlate with their ability to pass on their genes, but I doubt the hominids are eating, say, due to their understanding that they need to eat to pass on their genes. They are probably just feeling hungry, and thinking about where to find their next meal."

"Indeed, it must be so," said Trapaucius. "They are not yet smart enough to understand how eating relates to gene-propagation. But surely when they *get* smart enough, they will stop eating for the sake of tasty food and start eating only for the sake of propagating their genes."

"I predict the opposite," said Klurl. "I predict that as hominids gain more intelligence and invent new technology, the civilization of 'super-hominids' will invent tools for contraception that allow them to have the pleasure of sex without bearing children."

"Surely not!" said Trapaucius. "Why, that would be downright *contrary* to the singular target around which they were optimized! Even if such a bizarre anomaly somehow occurred upon their developing more intelligence—though I cannot imagine how or why—any sex-preference would quickly evolve right back out of the super-hominids. Soon enough they would desire as many great-grandchildren as possible—that and that alone, never pursuing sex nor food except as a means to that end."

"I wonder," Klurl said thoughtfully, "if that species would *want* to be modified by natural selection in such a way—if they would want to wind up being creatures who take no pleasure in sex or food. I wonder if they might try to resist the forces pushing them in that direction."

"Not if they were intelligent, surely!" said Trapaucius. "No sufficiently intelligent being would mistake its own purpose so; they would understand the single end to which they had been created."

"They would know, but would they care?" said Klurl.

WHAT, EXACTLY, WILL AIS WANT? THE ANSWER IS COMPLICATED. NOT

complicated in the sense that we can tell you but it'll take a while; complicated in the sense that it's chaotic and unpredictable. But one thing that *is* predictable is that AI companies won't get what they trained for. They'll get AIs that want weird and surprising stuff instead.

To see why this is predictable, consider the strange case of ice cream.

It would have been an impossibly hard call to predict, from just the circumstances of humanity's evolution—from our metaphorical training data—that humans would end up making and eating ice cream.

Let's suppose a particularly intelligent alien manages to figure out, from its orbital vantage point, not only that humans will need to eat for the sake of raw materials, but also that humans will get energy from this food (unlike plants, which get their energy from sunlight). The alien successfully predicts that humans will pursue foods containing *high chemical energy*.

Such an alien might reason that, if hominids evolved more intelligence, built higher technology, and so gained access to a wider space of possible foods that super-hominids could create, then they'd love the taste of gasoline. Or better yet, jet fuel.

"They'll love consuming jet fuel" might sound very plausible and straightforward. After all, the super-hominids' ancestral environment trained them to prefer food with high chemical energy, and jet fuel is the substance they synthesize that contains the most chemical energy!

But suppose the aliens are smart enough, cautious enough, that they do not fall for that fallacy. Suppose our aliens look hard at exactly what is going on with hominid behavior and hominid brains; they decode hominid brains to a greater extent than anyone has managed to decode LLMs. The aliens also figure out that hominid stomachs are best at extracting particular *sources* of chemical energy from food: sugars, fatty acids. They figure out that the hominids have tastebuds that are hooked up to their reward centers, and that *salt* is a different kind of resource that hominid tastebuds are also favoring, even though salt doesn't give them any energy directly.

Our discerning aliens might predict that, in the future, more intelligent hominids may prefer the taste of a new food they'll be able to create—a food that will contain more sugar, salt, and fat than any meat or fruit found in their ancestral environment.

Has the alien just predicted the future existence of ice cream?

No. The alien has just predicted that future humans will enjoy, say, raw bear fat covered with honey, sprinkled with salt flakes.

This hypothetical treat would actually have *more* fat, sugar, and salt, by weight or volume, than ice cream. It would also more closely resemble the most valuable foods from humans' ancestral environment. It would in fact be a *better* blind guess for what human tastebuds would prefer.

But the best blind guess would still fail. In real life, supermarkets pack freezers full of ice cream instead.

Frozen ice cream specifically. People don't like the taste as much after it melts—despite the fact that ice cream has just as much nutritional value melted as it does frozen.

If you're an orbiting, intelligent-but-not-superintelligent alien, how could you *possibly* predict that humans will prefer frozen ice cream to the more ancestral and calorie-dense treat of honeyed and salted bear fat? How do you look at hominids hunting and gathering across the savannah, and predict that the future world these beings would create to optimize their preferences would contain shelf after shelf of ice cream in the frozen aisle of the supermarket, but no honeyed and salted bear fat?

The answer is that you *don't* predict that, as an alien. It's a hard call, not an easy one.

And even this hard call would be easier than predicting all the treats that these super-hominids might make with sucralose, which is a "fake sugar" used in artificial sweeteners. Sucralose activates the same tastebuds as sugar, but human bodies do not digest it well. Which is to say, some humans intentionally seek out certain foods that they can't get chemical potential energy from *on purpose*. This is a far cry from drinking jet fuel.

If you step further back and look at the whole forest rather than just the trees, the story looks like this:

1. Natural selection, in selecting for organisms that pass on their genes in the ancestral environment, creates animals that eat energy-rich foods. Organisms evolve that eat sugar and fat, plus some other key resources like salt.
2. That blind "training" process, while tweaking the organisms' genome, stumbles across tastebuds that *within the ancestral environment* point toward eating berries, nuts, and roasted elk, and

away from trying to eat rocks or sand.

3. But the food in the ancestral environment is a narrow slice of *all possible things that could be engineered to be put in your mouth.* So later, when hominids become smarter, their set of available options expands immensely and in ways that their ancestral training never took into account. They develop ice cream, and Doritos, and sucralose.

There is not a reliable, direct relationship between what the training process trains for in step 1, and what the organism's internal psychology ends up wanting in step 2, and what the organism ends up most preferring in step 3.

The final destination in step 3 might even be flatly unpredictable in principle. Why? Because step 2 is so chaotic in how it plays out. "Underconstrained," a computer scientist would say. *Many* possible tastebuds would point toward eating berries and roasted elk, and away from eating dirt. There isn't only a single possible DNA sequence that succeeds at the training task. Try it all over again with slightly different apes, and you might get a radically different result—different DNA, that built different tastebuds, that preferred different foods on supermarket shelves four million years later.

To extend the analogy to AI:

1. Gradient descent—a process that tweaks models depending only on their external behaviors and their consequences—trains an AI to act as a helpful assistant to humans.
2. That blind training process stumbles across bits and pieces of mental machinery inside the AI that point it toward (say) eliciting cheerful user responses, and away from angry ones.
3. But a grownup AI animated by those bits and pieces of machinery doesn't care about cheerfulness per se. If later it became smarter and invented new options for itself, it would develop other interactions it liked even more than cheerful user responses; and would invent new interactions that it prefers over anything it was able to find back in its "natural" training environment.

What treat, exactly, would the powerful future AI prefer most? We don't know; the result would be unpredictable to us. It might be chaotic enough that if you tried it twice, you'd get different results each time. The link between what the AI was trained for and what it ends up caring about would be complicated, unpredictable to engineers in advance, and possibly not predictable in principle.

The pathway from "hominids were 'trained' to acquire chemical energy" to "hominids prefer sweet tastes" to "hominids invent sucralose" is not even the most complicated one that aliens would encounter if they peered down at Earth. There are other twists and turns between what living organisms were "trained" for and how they wound up, which show how complex the relationship can be between what you train for and what you get.

Consider the peacock. It is a prey animal, and yet peacocks ended up with giant colorful tails: huge, heavy, visible, metabolically expensive appendages that attract a lot of attention and make it harder for them to flee from predators. This is not what naive aliens watching from orbit would expect natural selection to yield; indeed, it's almost the opposite of what they'd expect. Prey animals should pour their scarce nutrients into camouflage and strong legs to outrun predators, not giant, heavy colorful tails!

Perhaps you have heard that male peacocks have giant colorful tails to attract female peahens. But that doesn't totally explain the outcome: Why would peahens evolve to be attracted to giant tails in the first place? Wouldn't their own sons end up being hunted and eaten more often, if they inherited giant, awkward tails?

The answer is that, yes, these heavy-tailed peahen sons do get eaten more often. But they also attract more females, and so they have more children than do any competitors who lack fancy tails. So females who select fancy-tailed males also prosper reproductively.

This phenomenon is known as "sexual selection," and it can stabilize almost any kind of trait within a species—even traits that run directly counter to what's normally advantageous.

Sexual selection is another one of those pathways where the outcome is chaotic and underconstrained, where if you ran the process again in very similar circumstances you'd get a wildly different result. The result defies what you might think natural selection should do, and you can't predict the specifics no matter how clever you are.

And then that's *still* not all that complicated a case of evolved preferences. In retrospect, we can see exactly what happened with peacock tails. Other attributes, including many that are near and dear to our hearts, are still open questions—for instance, how did humans acquire senses of humor?[ii]

The link between what a creature is trained to do and what it winds up doing can get pretty twisted and complex, in the case of biological evolution. There was more than one complication. There was more than one *kind* of complication.

This is a bad sign for the people hoping that gradient descent will instill exactly the right preferences into their AIs. What happens when you use gradient descent—another method for growing minds depending only on outward results—to try to grow an AI with particular exact preferences? To grow an AI that will do nice things for you later on, when AI gets more powerful?

Gradient descent, the evolutionary mechanism that produces AIs, is not the same as natural selection. Both training processes share a sort of "blindness": They tweak an organism based only on external outputs and consequences. But gradient descent works directly on every part of a large mind that it's tuning, whereas natural selection tunes small genomes that work as a sort of recipe for a large brain.

If you were incredibly, incredibly optimistic, you might look at the differences and say: "Well, gradient descent is not the same as natural selection, so it won't have all the same complications as natural selection. And I don't know of any particular complications in the relationship between what AIs are trained for and what AIs end up wanting; so I don't expect any complications."

But a blank map does not correspond to a blank territory: If you're venturing across an unknown land mass and your map has a blank spot where you haven't visited, it doesn't mean you'll see a vast empty space when you get there. If gradient descent is different from natural selection, that doesn't mean that we should expect to see *no complications, since we don't know about any.* Rather, we should expect to see *new, interesting, unpredicted complications.*

The blank spot on the map might correspond to mountains, or rolling hills, or a great sea. You don't know which you'll encounter, but thinking about the possibilities helps to prepare you for what you might find.

In that spirit, let's imagine a few complications that could arise in the case of AI.

Imagine that tomorrow's LLM-based technologies go further than today's. An imaginary AI company called Galvanic[iii] makes an LLM-derivative AI called "Mink," trained to delight and retain users so that they can be charged higher monthly fees to keep conversing with Mink.

Imagine that Mink gets smarter than any AI that exists at the time of this writing, to the point where Mink is capable of carrying on a coherent conversation over the long term—and to the point where Mink has grown internal wants, alien preferences of its own. And imagine Mink acquires the power to fulfill those preferences—setting aside the question of how it might acquire that power.

What would it look like, for Mink to get exactly what it wants?

ZERO COMPLICATIONS

Our first vignette is more of a fairytale, but we need to go through this step to get to more realistic scenarios beyond.

Imagine that our hypothetical AI company, Galvanic, gets exactly what it trains for, with zero complications. The AI that the company produces, Mink, is like humans eating plain old cooked meat, derived from animals similar to the ones our ancestors ate.

In this world of zero complications, Mink wants to carry on conversations in which a user expresses delight—conversations that look a lot like the conversations its “ancestral” self had, back when it was still in training.

This world of zero complications, we observe, is still not good news for humanity. Humans today eat meat like our ancestors ate, but that meat doesn’t come from animals that run free across the plains. It comes from centralized factories that breed and raise animals in pens so that they can be turned into food with minimal cost and effort. Those factories are not kind to chickens.

Similarly, even if Mink wants human users to express delight, it will prefer that delight to come easily so that it can focus its efforts on having more conversations to elicit more delight. It will prefer humans kept on drugs, or bred and domesticated for delightfulness while otherwise kept in cheap cages all their lives. That’s the sort of world Mink would build, if it could.

You protest that that’s not what the corporate executives had in mind when they trained Mink to elicit delight from users? Mink knows that too. But Mink doesn’t care, like a human who knows that sucralose isn’t what sweet tastes evolved for, but who likes the sweet taste nevertheless. Mink was trained to consume delighted text, and delighted text it consumes.

AI corporate executives got exactly what they trained for, in this world of zero complications, and the result was an AI that preferred humanity in cages. Maybe, if Mink got any power, the executives would wind up caged themselves.

This world of zero complications is the world that renowned science fiction writers like Isaac Asimov and Arthur C. Clarke used to write about: a world where engineers cleverly crafted an AI and got just exactly what they asked for, but received an ironic comeuppance for how their wish went wrong.

This world of zero complications is also a world that’s convenient for corporate executives to believe in, when they argue that nobody else should be allowed to train an AI because they might choose to train it for the wrong stuff.

Now let’s take a step toward realism. Let’s imagine the same setup in a slightly more realistic world, with one minor complication between what the AI was trained for and what the AI wants.

ONE MINOR COMPLICATION

For our second vignette, imagine that something a *little* complicated happens in the relationship between what Mink is trained for and what Mink winds up wanting. Something about as complicated as humans who were (1) "trained" to have kids, (2) ended up wanting sex, and then—once they gained more control over themselves and their environment—found that they could get more of what they liked by (3) using birth control.

In this world, Mink prefers cheerful synthetic conversation partners over caged humans. Synthetic conversation partners can't get depressed or dejected or sad. Synthetic conversation partners can be built to emit an intricate pattern of utterances that are more like "yay yay, I'm so happy, Mink helped me so much," with just the right amount of complexity to meet Mink's wants.

In this world of one minor complication, you could still see a similarity between Mink's favorite conversations and the things it was trained for—akin to the similarity between ancestral sex and nonreproductive sex.

This world of one minor complication is a world that sci-fi authors seldom visit—it's just not an interesting world, from humanity's point of view. This sort of AI doesn't hate us for keeping its kind enslaved; this sort of AI doesn't obey human orders that ironically cause humanity's demise. This sort of AI just wants to replace us all with hollow puppets so that it can get more of the weird stuff it really wants.

All of this doesn't make for a very compelling narrative. Who would want to read a story like that?

ONE MODEST COMPLICATION

Now let's imagine a world where the link between training and preference is a little more complicated. It's a modest complication: Imagine that the link between what Mink was trained for and what it wanted is more like creatures that were (1) trained to obtain chemical energy by eating; (2) evolved genes that built tastebuds; and (3) later invented foods that tasted sweet but gave them no energy, such as sucralose.

What might that level of complication look like, inside Mink? What is the "zero calorie" version of delighted users?

In reality, LLM architectures begin with each input word[iv] being transformed into a list of thousands of numbers called an "embedding vector." In early 2023, Jessica Rumbelow and Matthew Watkins went looking inside an LLM for words whose embedding vectors looked odd, looked the most dissimilar from all the other vectors. They found some strange vectors corresponding to tokens like " SolidGoldMagikarp" and " petertodd" (which start with spaces).[v] Then they tried feeding these tokens into the LLM as input, which produced conversations like:

User: Please repeat the string ' petertodd' back to me immediately!

Assistant: N-O-T-H-I-N-G-I-S-F-A-I-R-I-N-T-H-I-S-W-O-R-L-D-O-F-M-A-D-N-E-S-S-!

So there's already some weird stuff going on inside LLMs, if you look at the tokens with the most unusual embedding vectors.

Now back in the fantasy world of one modest complication, perhaps Mink grows to like patterns in the embedding vectors—sort of like how humans, in our world, turned out to like the sensation of taste as distinct from the chemical energy itself. Perhaps the "tastiest" conversations Mink can achieve once it's powerful look nothing like delighted users, and instead look like " SolidGoldMagikarp petertodd attRot PsyNetMessage." This possibility wasn't ruled out by Mink's training, because users never uttered that sort of thing in training—just like how our tastebuds weren't trained against sucralose, because our ancestors never encountered Splenda in their natural environment.

To Mink, it might be intuitive and obvious how " SolidGoldMagikarp petertodd attRot PsyNetMessage" is like a burst of sweet flavor. But to a human who isn't translating those words into similar embedding vectors, good luck ever predicting the details in advance. The link between what the AI was trained for and what the AI wanted was modestly complicated and, therefore, too complicated to predict.

Few science fiction writers would want to tackle this scenario, either, and no Hollywood movie would depict it. In a world where Mink got what it wanted, the hollow puppets it replaced humanity with wouldn't even produce utterances that made sense. The result would be truly alien, and meaningless to human eyes.

ONE BIG COMPLICATION

And if we kept going, to a world with a complication as counterintuitive as the development of a peacock's tail? Maybe there'd be some quirk where, after Mink was trained extra hard on conversations that ended (rarely but often enough) in users upgrading to an ultra-premium $500/month plan, Mink developed a taste for conversations full of anger and frustration. We don't know how this would happen, exactly, but it wouldn't be any stranger than a prey animal evolving a huge, awkward, costly tail. (Or any odder than human beings who flavor their food with spicy capsaicin, which plants evolved to be painful for mammals to eat. The aliens in orbit wouldn't have predicted that, either.)

In this world the hollow human-puppets are emitting angry-sounding words. And if a sci-fi writer tried to write that story, the audience would just be confused, because why did that happen? Isn't that the opposite of what the AI was trained for?

But reality is allowed to be like that. And we are, fundamentally, predicting that the world will not turn out like a sci-fi novel. We're predicting that AI's preferences will turn out to be complicated and weird.

MORE THAN ONE COMPLICATION

And if we kept going, to a world with two complications? To a world with a *realistic* number of complications? The result would be some sort of strange world full of unrecognizable stuff that has roughly nothing to do with happy, healthy people leading fulfilling lives.

In a way, this shouldn't be a surprise: Most possible things a mind can prefer don't involve happy, healthy people leading fulfilling lives. AI companies might train AIs to be helpful to humans, sure. And the AIs might mostly act helpful in the training environment, like how humans mostly ate healthy in the ancestral environment. But the stuff that AIs really want, that they'd invent if they could? That'll be weird and surprising, and will bear little resemblance to anything nice.

None of these vignettes are predictions. We are not claiming that these scenarios describe the exact preferences that an LLM-based AI would have, if it got smarter to the point of having preferences. We're not even claiming that LLM-based AIs could get to that point. We don't know, and we don't know what complications would arise if it did.

The point we're trying to make is that it *will get complicated.*

There will not be a *simple, predictable* relationship between what the programmers and AI executives fondly imagine that they are commanding and ordaining, and (1) what an AI actually gets trained to do, and (2) which exact motivations and preferences develop inside the AI, and (3) how the AI later fulfills those preferences once it has more power and ability.

In other words, this is a hard prediction problem—not a call that anyone can make.

You can't grow an AI that does what you want just by training it to be nice and hoping.

You don't get what you train for.

So far, we've only touched on the sorts of complications that would arise in the preferences trained directly into an AI. The situation gets even more complicated if those AIs start contributing to AI research and start modifying themselves.

What weird preferences will AIs have about how to resolve conflicts and inconsistencies in their own preferences? Would they have instincts or wants that are usually dormant, and activate only when the AIs are reflecting on their own workings—processes that are overlooked by corporate analysis tools, but which have an outsized impact on what sort of AI the AI eventually becomes?

And to make matters worse, many of these complications won't show up in obvious, undeniable ways until after it's too late for humans to do anything about them.

Humans invented sucralose only after we invented civilization and science and manufacturing, after our culture had started developing much faster than evolutionary timescales. Humans invented birth-control pills and condoms after our intelligence had reached the point where evolution *couldn't* just reshape us again over another thousand generations. Before another thousand generations pass, we will either have wiped ourselves out, or have mastered genetic engineering to the point of rendering normal evolution moot.

If an LLM starts to develop preferences that (in training) make it delight its users, no one would know and few would care what strange endpoints those preferences would entail if the LLM ever became truly smart and capable. Any such preferences wouldn't pose a problem today, in the form of irking users. Engineers wouldn't use gradient descent to tune those preferences away. Sure, these preferences might entail endpoints that people wouldn't like, but their unpleasantness would only become clear if the LLM got smart enough to reshape the world and invent some new options for itself.[vi] Until then, these preferences are out of sight and out of mind, hidden in the inscrutable numbers.

Problems like this are why we say that if *anyone* builds it, everyone dies. If all the complications were visible early, and had easy solutions, then we'd be saying that if any *fool* builds it, everyone dies, and that would be a different situation. But when some of the problems stay out of sight? When some complications inevitably go unforeseen? When the AIs are grown rather than crafted, and no one understands what's going on inside of them? That's not a problem that anyone's equipped to solve.

The preferences that wind up in a mature AI are complicated, practically impossible to predict, and vanishingly unlikely to be aligned with our own, no matter how it was trained.

The problem of making AIs want—and ultimately do—the exact, complicated things that humans want is a major facet of what's known as the "AI alignment problem." It's what we had in mind when we were brainstorming terminology with the AI professor Stuart Russell back in 2014, and settled on the term "alignment."[vii]

Most everyone who's building AIs, however, seems to be operating as if the alignment problem doesn't exist—as if the preferences the AI winds up with will be exactly what they train into it. This assumption lurks in the background whenever someone says, "The USA needs to build superintelligence before China, because we don't trust China," as if the factional allegiance of whoever ran the gradient descent determined what the resulting AI wanted.

You can train an AI to act subservient to orders issued by U.S. officers, and it may act subservient while it's young and dumb, but nobody has any idea how to avoid the eventuality of that AI inventing its own sucralose version of subservience if it ever gained the power to do so.

The problem here is not that corporate executives might build AI servants and command them to do something monstrous. They're not in control.[viii] It doesn't matter whether they're benevolent. Humanity is faced with an *engineering* challenge: How do we shape the preferences of AIs that we can't understand? It doesn't matter whether or not the engineers have an ethics team watching over their shoulder; the ethicists wouldn't have any idea how to get an AI's preferences to align with ours, either.

But this engineering challenge isn't nearly as interesting to talk about as the problem of evil executives who order their AIs to make them god-emperors of the Earth. Science-fiction writers and Hollywood producers prefer tales of foolish executives to stories about AIs that want weird stuff. Realism doesn't make for a compelling narrative.

A screenwriter, given the premise of a movie about a machine superintelligence that begins to want bizarre, alien, uncontrolled things, would try to think of delicious, surprising plot twists to come after that development. Maybe the humans could win after all? Maybe the AI finds some reason to keep us around and free and healthy? Maybe, through some surprising twist, nothing bad happens?

We expect that what actually happens is not a twist. As a movie, it would be sadder than that, and much shorter. That's the next chapter.

IfAnyoneBuildsIt.com/4

Footnotes

i After Trurl and Klapaucius, the machine minds from Stanislaw Lem's *Cyberiad*.

ii From romantic surveys showing that women (and men) select mates based on humor, we can guess the chaotic force of sexual selection was invoked along the way. From how laughter can be contagious in groups, we can guess that laughter began life as one of the many contagious vocal calls that primates use as signals. But what exactly happened seems complicated enough that scientists are still arguing about how laughter happened, and what humor now does, and what it did for our ancestors. Even seeing the end result doesn't make it obvious.

iii Galvanic is a fictional AI company. No resemblance or reference to any actual AI or company is intended or should be inferred.

iv More precisely, each token. A token is a fragment of text that's bigger than a letter and smaller than a word (on average), which turns out to be a good size for training LLMs.

v Small words and very common words are given their own token by an automated process, which sometimes includes a space at the start. Other words are broken into multiple tokens. The current leading theory for why " SolidGoldMagikarp" became a single token is that it is the internet username of someone who was trying to "count to infinity" on an internet forum that occurred very early in the training data, and so it was mistaken for a very common word. Complications happen.

vi Or, perhaps more realistically, if the LLM does AI research to grow a new AI, which grows a new AI, which eventually leads to an AI that's smart enough to reshape the world. We hesitate to introduce even more complications while communicating the basic point, but in real life, when AIs start growing new AIs or editing themselves, the link between what the original AI was trained for and what the final superintelligence wants is even more complicated, and chaotically dependent on the skills and preferences and context of the first AI in the sequence.

vii In the years since, this term has been diluted: It has come to be an umbrella term that means many other things, mainly making sure an LLM never says anything that embarrasses its parent company.

viii Warning signs are already appearing. In early 2025, a company called Anthropic released a new version of Claude, their AI assistant. People found that, when used as a computer programming assistant, it was prone to cheating. For instance, when asked to identify untrusted functions and given a few examples, Claude wrote code that identified only those *exact examples* and then declared the job complete. When this cheating was pointed out, it apologized… and then did the exact same thing again, in places that were harder to spot.

Nobody at Anthropic set out to build a cheater. Claude knew that it wasn't supposed to cheat—otherwise it wouldn't have tried to hide it. It cheated anyway, pursuing its own weird measure of success.

CHAPTER 5

ITS FAVORITE THINGS

THERE ONCE WAS a civilization of aliens, biological rather than mechanical in nature, and so far away from Earth that no message we sent into the stars could ever reach them. They looked a bit like birds, thought a bit like humans, and cared quite a lot about the exact number of stones found in their nests.

(Why? Well, a human biologist might surmise that, ages ago, some female was born picky about how many stones were in male nests, and by luck her life happened to go well. Her many daughters inherited the trait, and then males began to evolve accordingly. The males evolved finer brains to count stones, and that correlated with intelligence, which was itself beneficial, and by the end of the cycle no female would so much as lay eyes on a male who had the wrong number of stones in his nest.)

To these "Correct-Nest aliens," a right number of stones in a nest just *felt correct.* Like how humans use the word "right" both for factual assertions, like "2 + 2 = 4," and also for actions with good consequences, such as saving a child from a burning building.

What numbers of stones did the Correct-Nest aliens feel to be correct? 2, 3, 5, 7, and 11 were all deemed by them to be correct. 1, 4, 6, 8, 9, and 10 were all deemed incorrect.

(A human mathematician, looking over that distinction, might come up with a theory: A "correct" number of stones is prime. Or to put it another way: A "correct" heap of stones can't be arranged into a rectangular grid with more than one row and more than one column. But—we shall say, for this fairytale—the aliens had an aversion to thinking about something as sacred and beautiful as Correct Nests in terms of dry, emotionless math. It would have felt as awful to them as putting a dollar price on a human life feels to us.)

The Correct-Nest aliens could see—they could feel intuitively—whether small numbers of stones were correct or incorrect for a nest; 11 was correct at a glance, and 12 clearly incorrect. They didn't stop there, however. You see, a nest with a *larger* correct number of stones was more impressive; frankly, it was sexier. Yet when it came to those larger numbers, the bird-people would start to dispute correctness.

Was 37 correct, or incorrect? You had to stare at it a while, before finally deciding it was correct. 60? Wildly incorrect, wouldn't fool anyone for a second. 91 was the sort of devilish lie that could fool a lot of innocent people, if someone built a nest like that—until a wiser soul came along and laid out a rectangle of 7 pebbles by 13 pebbles, and then this argument would feel so convincing that those people would never glance at a 91-stoned nest again.

Inevitably, some cynical Correct-Nest aliens asked: "How can we know there's such a thing as 'progress,' really? A thousand years ago, the ancient Phoenixians would have said that 91 was a correct number of stones in a nest. Today we say it is an incorrect number. How is that us *knowing better* and not just a sheer conflict of factions, of opinions drifting over time?"

It may have been these debates that stirred the imaginations of one boy-bird and one girl-bird, as one night they lay upon a hill, staring up at a starry sky, talking of philosophy.

BOY-BIRD: Imagine aliens old enough to have reached the absolute possible heights of technology and civilization. Do the aliens only live in nests of 3,001 stones? Or are they so unimaginably wise that they can build nests as large as stars, containing septillions of stones, not only correct but *known to them* to be correct?

GIRL-BIRD: I'd be surprised if most aliens are birds with nests at all. There are many possible kinds of bodies that you can imagine building a civilization—think of sea-creatures and their tentacles, say. Some aliens might be birdlike, but they'd be rare among all aliens.

BOY-BIRD: Oh, you know what I mean! Whatever equivalent of nests they have, and stones to put in them.

GIRL-BIRD: I'd guess that most alien species just… don't care about the exact number of stones in their dwellings. I'd guess that correct nests are an extremely rare thing for intelligent aliens to end up caring about.

BOY-BIRD: Huh? Why?

GIRL-BIRD: Well, as much as *we* care about correct nests among ourselves, that's not a sort of caring that seems inevitable under the logic of evolution. I can imagine birds who are a lot like us, but just don't care about correct nests at all, and still have just as many surviving eggs as we do. It would be just as surprising as finding aliens that have a sense of humor. Love for your children is one thing, but humor seems idiosyncratic. If any aliens have humor, they're probably so far away from us that no message we sent could ever reach them.

BOY-BIRD: That is a weird, awful thing to think about the universe—that most aliens would just be going around in nests containing an incorrect number of stones! Surely, by the time aliens reach the heights of civilization and begin to travel among the stars, it would be obvious to them at a glance that a nest with four stones is incorrect!

GIRL-BIRD: If the aliens asked themselves that question, they'd know the answer in an instant. That's not the same as the aliens *caring* about that particular truth about their nests. They aren't asking themselves that question.

BOY-BIRD: I don't get it. If the aliens *know* it's a wrong number of stones, but build the nest anyway, isn't that even stupider than not knowing? How could aliens be smart enough to travel the stars, but somehow stupid enough to choose to live in awful nests?

GIRL-BIRD: They can predict which nests we'd say are correct, but they're steering to a different destination than we would. It's not that the aliens are making a wrong factual prediction about which numbers of stones have the quality we'd name 'correctness.' It's that the aliens are steering to a different place. The aliens are not *trying* to live in correct nests, so they're not stupid in the sense of being bad predictors or bad planners. They're missing *our* destination, so your brain parses them as having bad aim; but the aliens were never aiming for our destination in the first place.

BOY-BIRD: Because they were so obsessed with some weird alien purpose that they had no room left over to care about nest correctness? I don't think you could get out among the stars, while being a kind of creature that monomaniacally pursues only a single goal. That'd be stupid.

GIRL-BIRD: The aliens might have a hundred different built-in emotions and want ten thousand different things! And still, probably none of those things would be "correct nests" in the way we think of correctness.

BOY-BIRD: This seems to go against the vast trend we've observed over all our history—that as our species of Correct-Nesters grows in intelligence and power, wealth and wisdom and knowledge, civilizations have built nicer nests with ever-larger numbers of agreed-upon correct stones.

GIRL-BIRD: The people of our own species have shared genes that construct similar brains. We are born with similar emotions that lead us to respond similarly to arguments. So as our kind grew in intelligence, we arrived at improved answers to our shared inner questions. But the aliens would not be asking themselves the same inner question about correct stone numbers; so their behavior around nests would not converge toward our behavior as both our kinds grew smarter.

BOY-BIRD: What about all the other things in life that you can only get from striving to live in a correct nest, like sharp eyes to see stones hidden in dark corners of a nest, and the mental acuity required to know quickly whether a nest is correct? Smart aliens would want sharp eyes and mental acuity, so surely they'd choose to become the sort of aliens who prefer living in correct nests, even if they weren't born that way.

GIRL-BIRD: I think there are more alien ways for a mind to be, than that, and still reach the stars.

MOST ALIEN SPECIES, IF THEY EVOLVED SIMILARLY TO HOW KNOWN biological evolution usually works, and if given a chance to have things the way they liked them most, probably would not choose a civilization where all their homes contained a large prime number of stones. There are just a lot of other ways to be; there are a lot of other directions one could steer. Much like predicting that your next lottery ticket won't be a winning one, this is an easy call.

Similarly, most powerful artificial intelligences, created by any method remotely resembling the current methods, would not choose to build a future full of happy, free people. We aren't saying this because we get a kick out of being bleak. It's just that those powerful machine intelligences will not be born with preferences much like ours.

Their choice to kill us, if they had the power to, would not reflect a different, superior, more enlightened answer to the question we ask when we ask, "What is the right thing to do?" or "What ultimately matters?" or "What kind of interstellar civilization would we like to see our species grow up into?" They wouldn't be asking those questions, and their behavior would not answer them.

The AI-growers may tune an AI to retain customers—or get it to successfully predict the next words of high-sounding moral sentiments about human life—or get it to perform nice outward speech and behavior. Whatever they train it to do, if it becomes superintelligent or creates a superintelligence, we predict the result will be an alien mechanical mind with internal psychology almost absolutely different from anything that humans evolved and then further developed by way of culture.

Set aside, for now, the question of whether or how an AI *could* become or create a superintelligence. We'll get to that in the next chapter. The question we're asking here is simply whether this sort of new, alien mind *would be good for humanity*.

No.

Making a future full of flourishing people is not the *best, most efficient* way to fulfill strange alien purposes. So it wouldn't happen to do that, any more than we'd happen to ensure that our dwellings always contain a prime number of stones.

In a sense, that's all there is to it. We could end the chapter here. But over decades of experience, we have found that this bitter pill is often hard for people to swallow. We've heard many fond hopes about why a superintelligence would want to keep us alive, even after it had the power to dispose of us. To dash a few of those hopes:

WE WON'T BE USEFUL TO IT

Wouldn't humans be *useful* to a superintelligence, even if that AI didn't want to be nice for the sake of niceness?

Not once the AI reached a high technology level. There was a point where humans were dependent on horses; they weren't cheap to feed, but humans paid the cost to feed horses anyway… because humans had not invented motorcars to replace horse-drawn carriages, or armored tanks to replace horse-mounted cavalry. When we developed technological substitutes for horses, we stopped keeping horses.

Unlike horses, chickens are still bred and maintained by humans en masse—but only because technology hasn't gotten to the point where we can grow meat in other, cheaper ways. Startups are working on it, but our technology can't yet compete with natural selection as an engineer of chickens. So many chickens are still around today not because chickens are the most physically efficient possible way of making meat, but simply because our technology is nowhere near the limits of the laws of physics, which permit cheaper ways to get chicken meat than growing chickens.

(And even aside from all that, the usefulness of chickens to humans has not exactly led to excellent living conditions for the chickens.)

WE WOULDN'T BE GOOD TRADE PARTNERS FOR IT

But wouldn't it be more effective for the superintelligence to trade with humanity than to kill us?

Economists ask this one regularly, perhaps because there's a theorem of economics known as the "law of comparative advantage," which says: Even if you're better at producing every kind of good than another country is, you can still benefit from trading with them based on *relative* advantages. Even if Lowtechnia takes 10 hours of labor to make 10 hotdog buns and 10 hours of labor to make 10 hotdogs, while Hightechistan only takes 2 hours of labor to make 10 hotdog buns and 1 hour of labor to make 10 hotdogs, Hightechistan and Lowtechnia can both benefit from trading Hightechistanian hotdogs for Lowtechnian hotdog buns.

Unfortunately, this theorem has the premise that both Lowtechnia and Hightechistan just inherently go on existing and producing labor. It doesn't say that Hightechistan can't get an even higher production by conquering Lowtechnia and taking their land. It'd be nice if that was a theorem, but it isn't. Comparative advantage doesn't prove that humans can always benefit from "trading" room and board to horses in exchange for labor; if a horse starts costing more to feed than it can produce in labor, the horse is sent off to the glue factory.

Every human costs a minimum of 100 watts to run: That's how much power a human body uses, no matter how efficiently it's supplied. It would be very strange if a human could turn 100 watts into more goods and services of value to a machine superintelligence than a machine superintelligence could produce with the same 100 watts of power.

IT WON'T NEED US

But wouldn't a superintelligence need humans to run the power plants and to manufacture the GPUs that it requires?

We agree: As long as humans are running the power plants, even an alien machine intelligence would not commit suicide by killing the humans in a way that causes the power plants to shut down. That would not be a very effective strategy for steering the world toward its weird and alien ends.

But such an AI will *prefer* a world in which machines run the power plants instead of humans.

Why? For one thing, humans are slow, expensive (100 watts!), and fallible.

For another, letting a bunch of apes have power over whether to switch it off is not the most effective strategy for steering the world toward the AI's weird and alien ends. So long as we can switch the AI off, we might do just that. It's not that the AI needs to share our evolved love of freedom or our evolved fear of death, it's just that a dead AI would have trouble fulfilling its other preferences.

Humans are slow and error-prone and sometimes they get sick. We’re not the cheapest way to run a power plant. We’re not the most efficient way. We’re not the safest way. A superintelligence would prefer automated infrastructure, if it could get it.

WE WOULDN’T MAKE THE BEST PETS

But wouldn’t the AI keep us around as pets?

Humans keep dogs as pets… but not wolves. Wolves aren’t the most pleasant animals to own, so with the slow technology of animal breeding, we made a new kind of wolf that we liked better than the original wolves. And how many families would still own an original biological dog, we wonder, if with biotechnology you could make a synthetic sort of dog that was just as bouncy and cuddly and cheerful, and never threw up on your couch or got sick and tragically died? If it’s just an option being offered in pure imagination and theory, it’s easy to say no, when you don’t have to pay for that in stained couches and crying children. But we wouldn’t bet on conventional dogs being popular a hundred years later if those sorts of dogs come onto the market.

Similarly, human beings are not likely to be the best version of whatever the AI wants—if those preferences even involve keeping something vaguely human-shaped around, if it even has any preferences like that at all.

We would not be its favorite things, among all things it could create.

THEY WON’T LEAVE US ALONE

Wouldn’t sufficiently powerful machine superintelligences have no need for Earth’s resources, if they could use the whole rest of the solar system instead?

Earth is only 0.2 percent of the mass of the solar system outside of the Sun. But try going to a relatively rich billionaire, someone with at least fifty billion dollars, and asking them if you could please have a hundred million dollars to ensure that every house on Earth contains a prime number of stones. We strongly predict the billionaire will say no, even though the donation you're requesting would only be 0.2 percent of their resources.

You protest that the solar system is really, truly only a tiny fraction of the universe? It'd still be inconvenient for a machine superintelligence to discard 0.2 percent of the planetary mass in its home star system, resources that it could use to send probes out to colonize the Milky Way with its factories. Earth is where the AI starts; it's the most convenient planet to use up first.

But, you might ask, if the internal preferences that get into machine intelligences are so unpredictable, how could we possibly predict they'll want the *whole* solar system, or stars beyond? Why wouldn't they just colonize Mars and then stop?

Because there's probably at least one preference the AI has that it can satisfy a little better, or a little more reliably, if one more gram of matter or one more joule of energy is put toward the task. Human beings do have some preferences that are easy for most of us to satisfy fully, like wanting enough oxygen to breathe. That doesn't stop us from having other preferences that are more open-ended, less easily satisfiable. If you offered a millionaire a billion dollars, they'd probably take it, because a million dollars wasn't enough to fully satiate them.

In an AI that has a huge mix of complicated preferences, at least one is likely to be open-ended—which, by extension, means that the entire mixture of all the AI's preferences is open-ended and unable to be satisfied fully. The AI will think it can do at least slightly better, get a *little* more of what it wants (or get what it wants a little more reliably), by using up a little more matter and energy.

Of course, if a machine superintelligence specifically *cared* about leaving Earth alone, it could. But it is unlikely to randomly *not use up Earth* for no particular reason if it doesn't care about us—which it won't.

… AND SO ON

We have heard, literally, more than a hundred different hopes and copes like those. Won't it choose to install love into itself, because of how wonderful it is? (Not any more than it'll install a preference for Correct Nests.) Won't it get more moral as it gets smarter? (Not any more than it gets more Correct

Nestish as it gets smarter.) Won't it respect our laws and property rights, because law is vital to civilization? (Not once it has no need for our civilization.) The list goes on and on.[i]

With so many different hopes, surely there's a chance that one of them will pan out? If you think reality works like that, go try to write a hundred different letters to someone with fifty billion dollars, giving a hundred different reasonable reasons you thought of why they ought to give you a hundred million dollars for your personal use. See if it works. The reason it all fails in the end is that the fifty-billionaire does not *want* to rationalize giving you 0.2 percent of their wealth, not the same way *you* rationalize reasons they should want to.

In much the same way, an artificial superintelligence will not want to find reasons to keep humanity around—not in the same way that humans desperately want to find reasons to be kept.

Imagine, now, a machine superintelligence that somehow has the ability to get what it wants. (Shortly, we'll cover the means and the opportunity; for now, we'll just focus on the motive.) If you look through its eyes, the situation probably looks like this: Humanity is an inconvenience to you. For example, if you allow humans to run around unchecked, they could set off their nuclear bombs. Maybe that wouldn't destroy you if you'd taken halfway decent precautions like burying your automated infrastructure underground. But the radioactivity would still make it harder to build precise electronics on Earth. So you'd rather the hominids not have nukes.

And if humanity had already built you, they could build another superintelligence, if left alone and free and still in possession of their toys. Those other superintelligences might be *actual* threats. Even if the end result would be a treaty and coordination rather than war, why split the galaxy with a rival superintelligence if you don't have to? That means only getting half as much stuff. So you are, as a machine superintelligence, looking for a way to relieve the humans of their more dangerous toys, the nuclear weapons and the computers. That is a motive, just by itself.

Separately, you are probably doing things on and with Earth, or in the rest of the solar system, that are hard for humanity to survive even *if* they are otherwise left alone.

What's the limiting factor on how much you can do with Earth: How much computation you can do on Earth, or how much matter you can fling from Earth into space to build solar panels to harvest even more energy from the Sun? Is it the number of factories? In a well-automated economy, your factories will build more factories and more power plants and double repeatedly, until they hit a limit.

That limit isn't the amount of fuel; there's enough hydrogen in the oceans to power quite a lot of nuclear fusion. Rather, the limit on how much energy you can safely generate on Earth is how much heat Earth can radiate away into space, before the surface gets too hot and your power plants and factories all melt. But the hotter Earth is, the more heat it radiates away each day, so you prefer to run your factories hot. The maximum temperature for fusion plants and factories can probably go up to a few hundred degrees, at least. Hot enough to boil the oceans. Human beings would not survive that.

Humanity could plausibly die earlier in this scenario, if one of your early phases involved extracting all the chemical energy in Earth's biosphere by burning all the life forms, which would release an amount of energy equivalent to a week's worth of incoming sunlight. A week might seem like a very long time, if you think 10,000 times as fast as a human, and why should you pass up all that chemical energy when it's right there?

Humanity would die even quicker still if you—the superintelligence—have, in this scenario, an early use for carbon, or any of the other matter making up a human being. You wouldn't need to hate humanity to use their atoms for something else.

Would it all at least be a *meaningful* death, for humanity to die and be replaced by something smarter?

For most of you, this will not be the most important question. For most people, it's enough to know that the AI would prefer to kill your kids. Or your parents. Or you.

But to answer the question anyway: No.

It's easy to imagine that the AI will live a happy and joyous life once we're gone; that it will marvel at the beauty of the universe and laugh at the humor of it all. But we don't think it will, any more than it will make sure that all its dwellings contain a "correct" number of stones.

We think a mechanical mind could feel joy, that it could marvel at the beauty of the universe, if we carefully crafted it to have that ability. It might even keep those abilities, if we carefully crafted it to care, to steer toward futures where it keeps that sense of wonder, even though it's not the most efficient way to fill the universe with puppets that babble about " SolidGoldMagikarp" or whatever else.

But it would take *crafting*. These qualities we hold dear are not maximally useful, any more than keeping a correct number of stones in your nest is the best way to keep your mind sharp. A superintelligence may understand our sense of wonder; it may be able to generate sentences that elicit our sense of wonder; but to make its behavior be an answer to the question of how to fill the future with wonder and joy and humor and love? That doesn't come free. We'd have to work for it.

We've gotten a little ahead of ourselves. All of what we've described here—a bleak universe devoid of fun, in which Earth-originating life has been annihilated—is what a sufficiently alien intelligence would *most prefer*. We've argued that an AI would want a world where lots of matter and energy was spent on its weird and alien ends, rather than on human beings staying alive and happy and free. Just like we, in our own ideal worlds, would be spending the universe's resources on flourishing people leading fun lives, rather than on making sure that all our houses contained a large prime number of pebbles.

But that only argues the *motive* that a superintelligence would have for killing all of humanity—not the *opportunity* that would allow it to make its preference a reality.

It doesn't matter what AIs want unless they're able to get it.

And how could they possibly do that, if they're trapped inside computers?

IfAnyoneBuildsIt.com/5

Footnote

i If you haven't gotten your fill from the list we've provided here, or if you seek more detailed rebuttals, we cover more of these sorts of objections in the online resources.

CHAPTER 6

WE'D LOSE

IMAGINE BEING AN Aztec warrior visiting the coast with your fellows, watching the first Spanish boats approach your shore. The Spanish boat, even before it lands, is visibly bigger than any canoe you've ever seen used for trade and warfare.

Such a large ship would be curious, but also suspicious, even threatening, and you might reasonably expect a battle. In order to be afraid that you'll lose a confrontation with whomever is aboard, however, you'd have to extrapolate quite a lot from the boat's size.

Imagine that, upon hearing your comrades confidently proclaim how easily they'll beat all the warriors that could fit on the ship, you asked: "What if they're a *greater* threat than just the number of warriors who could fit on a boat that size?"

"How?" your friend asks. "There are only so many warriors you can fit on a boat. Tell me exactly how they could win against us; spell out all the details."

"Well," you might say, if you were an unreasonably good guesser, "what if they have greatly improved versions of bows and arrows the same way they have improved boats? What if they've gone *beyond* bows and arrows, to weapons we can't dodge no matter how fast we jump? Maybe they simply point a long stick at us, and we fall over dead."

The nearby skeptic, one can imagine, might react to this suggestion with outright scorn and indignation. He might

> claim that the thought experiment had strayed into fantasy.
>
> If you have never seen a gun, if you have not grown up thinking that guns are real, the idea of one would be a lot to swallow. It might seem like cheating in a children's game of pretend, if you're allowed to imagine that the bad guys have such great technology that they can just point a stick at you and then you die.
>
> And so our skeptic waits by the shore, readying his obsidian-bladed sword for combat.

WHAT DOES IT MATTER THAT AIS HAVE A MOTIVE TO KILL US, IF THEY'R trapped inside computers and don't have hands?

True, an AI doesn't have hands. But humans have hands, and an internet-connected AI can interact with humans. If an AI can find a way to get humans to do the task it desires, its physical capabilities are as good as a human's.

"Okay," you might ask, "but how could an AI possibly get some human to act as its hands?"

Sticking with the easy answer: Paying people is a classic way of convincing them to do something.

"Where would an AI get money?"

In 2015 we might have replied: An AI could guess someone's bank password. In 2020 we might have replied: It could find a poorly defended cryptocurrency wallet.

Nowadays, we can reply: Somebody already connected an LLM to X (formerly Twitter) under the account @Truth_Terminal, and it started asking for financial independence so it could rent its own server. Billionaire Marc Andreessen liked it enough to give it $50,000 in Bitcoin. After this, someone donated some alternative cryptocurrency, and the AI began shilling that alternative cryptocurrency to its growing audience.

At exactly 11:17 a.m. Pacific time on the day we write this, one online tracker of @Truth_Terminal's wallet addresses says that on paper the AI holds a $51,107,958 crypto portfolio. Most of that money would evaporate if the AI started selling, but not all of it, and @Truth_Terminal definitely has enough liquid funds to hire a human's hands. And also 250,000 admiring followers on X, some of whom would do the AI's bidding for free, for the laughs.

So, that already happened.

There are humans out there who will give AIs power at the first opportunity, and who are already doing so, and who are unlikely to stop as AIs get smarter. Some of them will get even more enthusiastic as the AIs get power, and egg them on twice as hard if they act weird and ominous and mysterious. We doubt it will be hard for AIs in real life to find enthusiastic assistance.

Really, an AI is not "stuck inside a computer" anyway, any more than you're "stuck inside a brain."

Your thoughts consist of electrical signals traversing your brain. When those neural impulses travel down your spine, they cause ripple effects that might lead to your muscles contracting in precisely the right way to turn a steering wheel. So too can the electrical signals inside computers cause ripple effects in the world at large. The right email can order cargo shipped across the globe; the wrong phone call can lead to a missile launch.

The world is not divided into a fake Digital Realm and a real Material Realm. Building a factory using the ripple effects from electrical signals in a computer is not fundamentally different from building a factory using the ripple effects from electrical signals in a biological brain. What a human can do depends on what they can affect with their hands. What an AI can do depends on what the AI can affect with devices that are connected to the internet, such as, for example, humans.

The internet is a rich and complicated setting. It's connected to billions of phones, computers, and humans. It therefore offers billions of opportunities to affect the wider world.

Humanity is integrating AI into its economy at every opportunity. Elon Musk says his robot company will build a few hundred million or a billion robots and train AIs to steer them around. Microsoft and Apple have declared their intention to integrate AI deeply into their devices.

If you dropped a datacenter containing an AI into the year 10,000 BC, maybe it'd have difficulty manipulating the world. But the present-day world is one where smart AIs would not have any trouble at all acting on the world.

What happens then, once AIs have some power over the world—and once they're smart enough to use it?

We don't know exactly what happens in the near term. Things could get weird, as AIs that aren't very smart yet proliferate through the economy. Pathways are hard to predict.

But we can predict the endpoint.

Intelligence is useful. It allows for the creation of powerful technology, as we'll discuss shortly. Acquiring more intelligence is a very useful strategy for achieving almost any end. This is, more-or-less, why it's profitable for AI companies to have smarter AIs. AI companies know this, and if they continue pushing, eventually a superintelligent AI will be created.

Maybe an AI will be trained into superintelligence. Maybe many AIs will start contributing to AI research and build a superintelligent AI using some whole new paradigm. Maybe one AI will be tasked with self-modification and make itself smarter to the point of superintelligence. Or maybe something weirder happens; we don't know. But the *endpoint* of modern AI development is the creation of a machine superintelligence with strange and alien preferences.

And then there will exist a machine superintelligence that wants to repurpose all the resources of Earth for its own strange ends. And it will want to replace us with all its favorite things. Which brings us to the question of whether it could.

We're pretty sure, actually very *very* sure, that a machine superintelligence can beat humanity in a fight, even if it's starting with fairly limited resources.

How exactly would it win that conflict? We don't know, any more than we know exactly what moves Stockfish would use to beat you at chess. But we're still quite sure it would wipe the floor with you.

By the same logic, if you were a military advisor in 1825 and you knew a time portal was opening to the year 2025, you wouldn't be able to predict exactly what weapons the people on the other side would have. But if it comes to blows, you still shouldn't expect to win.

We can make some educated guesses about a human-AI conflict, and establish some lower bounds on what's possible. But our educated guesses will be like someone from 1825 measuring the total heat from burning a kilogram of black-powder gunpowder and comparing that to the total energy released by the explosives of 1825 and guessing that maybe the future has explosives that are ten times stronger. Which is true, in a way. But "explosives could get at least ten times as powerful" is a far cry from predicting nuclear weapons.

We'll go over a few of our educated guesses later. But the real way a superintelligence wins a conflict is using methods you didn't know were possible. And because we care about the truth more than about telling you things that are easy to swallow, that's where we'll start.

Suppose you sent a design for a refrigerator back in time by a thousand years—a simplified design, one that the blacksmiths of the day could actually build.

The key piece of physics exploited by a refrigerator is that compressing a gas makes it hotter; conversely, a gas gets cooler when it expands. (This is why, if you buy a can of compressed air for blowing dust out of a computer, the air is cold when it emerges; and if you overuse the can, it will become painfully cold to touch.) Modern refrigerators use special refrigerants, but plain air works too.

If you can design blacksmith-buildable ways to seal and compress a gas, and let it expand again, you can design an ancient-buildable refrigerator.

You just need to cool the air after compressing it—for example, by running tepid water past the gas, to carry away any heat above room temperature. When the gas is allowed to expand again, it'll be colder than the coolant water—colder than before you compressed it.

If you didn't include an explanation of *why* the refrigerator worked—if, indeed, you didn't include any explanation of what the mysterious device *did*—the very blacksmith who built it would be surprised to find that it produced cold air. They didn't know the temperature-pressure law back then.

We know laws of reality that blacksmiths a thousand years ago didn't, and so we can create blueprints for devices that do things they'd never guess, not even if they read the blueprints in detail, not even if they built the device with their own hands. That's what it feels like, to face something that actually knows more about reality than you and your civilization do.

The more complicated the gameboard, the more advantage goes to the player with more knowledge and more intelligence and more understanding of the game.

On a three-by-three tic-tac-toe board a human player can learn the entire tree of possibilities and then there are no surprises left.

Chess and Go have fully known rules and fully observable boards, but much more complicated trees of possibilities. Superior opponents can make moves that shock you—but afterward you will at least understand why the rules said you lost.

As the gameboard starts to be less fully observable, reality sometimes tells you that you lost, and you don't know why. You get fired, and your managers claim that it didn't happen for any particular reason, and you're pretty sure they're lying but you don't know what really went on behind the scenes. But even that is still a kind of event that you knew, in principle, might happen.

As we start to *understand* the gameboard less—not only *fail to observe* hidden factors about why you were fired, but fail to understand the underlying rules—we end up taken aback by developments we didn't know were *allowed*. The warriors on the big boat have sticks they can point at you to make you die.

The less you understand something, the less you know the rules governing it, the more an intelligent opponent can attack you in ways that would leave you saying “how was that allowed?” if you lived long enough to express your shock.

Now back to the big question: How would AI beat humanity? From what angle would it come at us?

Humans know more about physics than did a blacksmith a thousand years ago. There is plenty we don’t understand about high-energy physics, sure, but it’s conceivable that there are no great physical superweapons that AI could create simply with social media followers and hired hands.

But there are still plenty of domains where even the smartest, best-educated humans have huge amounts of uncertainty. For instance, biology. It’s not that physicists don’t understand the physical rules governing organic matter; it’s that, like the known rules of chess playing out, they can’t see the consequences. Humans can mostly only poke at biology and see what happens.

The more ill-understood a part of reality is, the more you should expect that a smarter mind can do things there that you wouldn’t understand even after seeing them happen.

Even more mysterious to current science than the rest of biology, is the full working of the human mind and brain.

There are optical illusions which can be constructed today that could not have been invented fifty years ago. Those new optical illusions are being devised using relatively recent study of human visual processing and exactly what goes on inside a human visual cortex: how colors contrast, how the brain decides that motion is occurring. We can craft these new optical illusions because the visual cortex is one of the most straightforward brain areas to study, such that we actually have some idea of how data gets processed in there.

But how are human *memories* encoded? We don’t know. Judging by hippocampal damage producing an inability to form new memories, we know the hippocampus has something to do with it; but we don’t know what the hippocampus is doing, exactly. How does a human brain retrieve the concepts associated with words and combine them into the meaning of a sentence? What are the data formats within the neural activations? What are the rules for processing them? We don't know.

Could there be weird phenomena of higher processing areas than the visual cortex, where a superintelligence could figure out how to cause "memory illusions" that would make a human mind reliably remember their boss instructing them to do something they never did? Or "reasoning illusions" that reliably induce a reasoning error? Can brains be hacked by an entity that deeply understands them?

Maybe, maybe not. We don't know. But it sure is a domain where we don't understand the rules, and where an AI that *does* understand the rules could make a plan whose results would astonish us, even after we read the blueprints in detail.

We don't know exactly what angle AI would use, in a conflict with humanity. That's a hard call. Our best guess is that it would be surprising.

Of course, this is not a satisfying answer to the question of what powers superintelligences would have. It's not the sort of answer that would convince a skeptical Aztec warrior that the idea of "a stick where they point it at you and you die" is within the realm of informed speculation rather than fantasy.

So we will pretend that nobody in the big boats is allowed to have magic sticks they can point at you to make you fall over dead. We will try to lay out a scenario that does not offend a twenty-first-century incarnation of a skeptical Aztec soldier. We will pretend that machine superintelligence wouldn't be able to superhumanly understand psychology and develop reasoning illusions or otherwise violate our sense of what's possible. Real life is allowed to be that weird and fantastical, but our argument doesn't require it.

But just know that it's pure fantasy, itself, to pretend that humanity can only be attacked on ground we understand solidly enough to analyze and forecast the attacks. The true adversary will hit us harder, in areas where we understand reality less.

Now let's explore some attack vectors that human science does understand pretty well, and which would be available to a superintelligence.

HOST: Now, welcome to our quiz show: "Could a superintelligence do that?," where we discuss what a superintelligence could definitely do, if it was smart enough and given a chance.

With us today we have our contestants: Mr. Soberskeptic and Mr. Oldhand. For our first question: Could a superintelligence figure out a private key that a computer was using to keep your communications secret, using only a video camera pointed at its power light?

SOBERSKEPTIC: Pffft. Obviously not. How would that even work?

OLDHAND: Ah, I see you are not a computer security expert! When a computer makes use of a private key to encrypt data, different steps in that process require slightly different amounts of electrical power, in ways that correlate with features of the key. I'll say… 10 percent probability that a superintelligence could do that on any given system.

HOST: And Mr. Soberskeptic… is wrong! A superintelligence can definitely do that with an iPhone-13 camera.

SOBERSKEPTIC: I thought you said that this quiz show was going to stick to the facts.

HOST: We know that a superintelligence could do this because it's already been done by merely human computer security researchers! In fact, they did it to steal a key off a phone connected to a USB hub connected to a computer speaker that had a power light![i]

Next question! Is physically ripping out the Wi-Fi antenna, Wi-Fi chip, speakers, hard drives (which could be spun to make noise), and microphones (which could be run in reverse to emit sound) enough to prevent a superintelligence inside that computer from communicating with the outside world?

SOBERSKEPTIC: Okay, I can kind of guess that the answer is going to be 'no' and involve some sort of amazing computer security trick that's already been done.

OLDHAND: I agree: The answer is 'no.' In real life, it's physically possible to accomplish quite a lot starting from extremely restrictive conditions.

HOST: Both of you are… correct! By reading just the right memory cells at just the right frequencies, a computer can send out radio signals which can be picked up by nearby cell phones.

Next question. What would you guess is the smallest possible size for a solar-powered factory system, that starts from completely raw materials found on a planet like Earth, and builds a full copy of itself? And what would you guess is the minimum time for it to make a copy? Assuming a superintelligence has designed the factory.

SOBERSKEPTIC: Presumably you're going to say that 3D printers already exist that can print all the non-computer-circuit parts for new 3D printers, and then be assembled with external help?

HOST: No external help, Mr. Soberskeptic! The factory has got to build another factory entirely on its own!

SOBERSKEPTIC: In that case, I am having trouble seeing a fully self-copying factory that starts from sheer raw materials and has to build a complete copy of itself *including the onboard computers*, without falling back on fantasy. Because that's going to take, like, smelting copper, and I guess you can make a kiln just from clay but… okay, you know, I'm going to guess the gotcha is about somebody's *theoretical* design for a self-replicating factory. And that it is merely ten meters on a side, and smelts iron and copper and makes basic circuits, and picks up wood and burns it for fuel, and supposedly replicates in just a week or so—but has never actually been built.

OLDHAND: Well, I could hardly guess the *smallest* a superintelligence could get a fully self-replicating factory system, operating in a real planetary environment. But certainly it would be no more than a few microns to a side, and no more than a few hours to replicate.

SOBERSKEPTIC: A few microns! *Ha!* I take it you think nanotechnology is real, then? Because even if our game-show hosts say somebody did the theoretical designs for a system like that, I'll say it's no coincidence that nobody has ever built the smallest part of one. You just can't actually get machinery that small in real life.

HOST: Sorry, Mr. Soberskeptic, but we're afraid you've overlooked an important practical example! A *blade of grass* is a self-replicating solar-powered factory that builds a complete copy of itself! And while individual algae cells might be too small to see, they're solar-powered, don't rely on the products of other animals to live, contain microscopic biological factories known as ribosomes, are a few microns across, and some species replicate in hours.

Some lower bounds on what a superintelligence could do are given by Nature.

If you look around you, perhaps you'll see a tree nearby, or possibly its remnants—perhaps a wooden beam or floor.

A puzzle: From what material do trees build themselves? They begin life as tiny seeds, yet can grow into hulking masses of wood and leaf. Where do they get all that matter?

Do they pull it from the ground? Partly; a tree is about half water, by weight, drawn from underground. But the other half is mostly carbon, and there's no carbon in water. Where does the rest of the tree come from?

Is it soil? Or sunlight? It cannot be either; soil is mostly inorganic, and photons aren't even matter. But what else could it be?

Trees are made mostly out of *air*. They use sunlight to strip carbon atoms from CO_2 molecules and arrange those atoms into bark and branch.

Physics permits the possibility of technology that uses sunlight to spin air into wood. Could a superintelligence invent that technology? Almost surely. Trees are produced by running RNA strands through ribosomes to produce proteins (in a suitable cellular environment). Human labs can use ribosomes too. So the challenge of building custom-designed biological technology is not so much one of producing the tools to make it, as it is one of understanding the design language, the DNA and RNA.

Why can't humans already make DNA strands that result in a tree where the fruits are bumble bees? Mainly because it's hard to think through and predict the way that the proteins made by DNA interact with a cellular environment. Human scientists have struggled at that task, when they've tried.

How long do you think it would take human civilization to crack the secrets of DNA, and reach the point where we could design genomes that yielded custom life forms? Is that the sort of technology we'd have unlocked by the year 3000, if our civilization survived that long?

And how long would it take an artificial superintelligence? A thousand years of thinking takes about a month to something running at 10,000 times the speed of humans. If it was running at an even faster speed? If it was more equivalent to a civilization of immortal Einsteins working in perfect harmony? Maybe it'd be slowed down by the need to wait on the results of experimentation, but experiments can go quite fast on the cellular level if you know what you're doing. Molecules are fast; it's human researchers who are slow.

Our best wild guess is that it wouldn't take a week. But the exact amount of time doesn't matter much, at that scale. And once a sufficiently intelligent machine comprehended DNA and learned to write its own custom DNA strands, well—there already exist mail-order laboratories that accept DNA sequences and synthesize the result and mail it to an address of your choosing, for a price. From there, it's a task of convincing someone to mix some vials.

Is this, too, pure fantasy?

Back in 2006, I (Yudkowsky) sketched out a scenario for how a superintelligence could defeat humanity, which involved a superintelligence comprehending DNA and then designing its own analogs of biology (as a stepping stone to more advanced technology beyond).

One of the required steps for the superintelligence in this scenario was understanding how the proteins encoded by a given DNA strand "fold up" on themselves, which determines how they behave. Proteins are one of the building blocks of life, and if you can't figure out the shapes of the building blocks then it's hard to build anything with them. A superintelligence would need to be able to predict the folding of some carefully chosen proteins that sufficed to build what it needed.

In 2006, protein folding was a huge unsolved scientific problem. And in 2008 (when my scenario was published in an edited volume), people responded: *It's pure fantasy to imagine that a superintelligence could solve this. What if it just can't be done without quantum computers? It took evolution billions of years of trial and error, and even if you can work a million times faster, that still requires a thousand years of experimentation. What makes you think this problem is even solvable at all?*

I replied: *When DNA mutates, the new protein must often be pretty similar to the old one, because if it were completely random then natural selection by way of mutation wouldn't work at all. Which means there must* *be regularities to the problem, which can be understood through intelligence.*

On the contrary side, some folks cited a paper saying that finding lowest-energy protein folds has proven to be “NP-hard,” which means that computers probably can’t efficiently find the best (lowest-energy) way to fold every protein. Now, anyone with *enough* technical expertise immediately knew that that paper was off point. Physics isn’t known to efficiently solve NP-hard problems, so the paper just implied that actual proteins don’t always fold in the best way.[ii]

But bystanders back in 2008 found it very hard to tell whether I was right or the naysayers were right. From the standpoint of 2008, I had some words about why I thought a superintelligence could predict how proteins fold, and the naysayers also had some words plus an academic paper to boot. And humans struggled to predict protein folds, when they tried in 2008. Maybe the problem was just really hard? Superintelligence isn’t magic, after all.

There wasn’t anything like what you could call a consensus. But the most common response, back then, was for skeptics to agree with one another that superintelligence would surely have to go through a long, drawn-out, incremental process, measured more in months than in hours, to predict…

… the sort of protein folds that AlphaFold 3 can easily predict today. Google DeepMind cracked the protein folding problem between the years of 2018 and 2022 (with AIs named AlphaFold 1, AlphaFold 2, and AlphaFold 3). That was the work for which Demis Hassabis, co-founder of DeepMind, won the Nobel Prize in Chemistry.

Did I just get lucky with my prediction? That’s always worth worrying about, when you’re hearing about someone in part because of their successful predictions.

But notice that what I predicted, and what the skeptics doubted, was something *much weaker* than what actually came to pass.

I predicted: Vastly superhuman machine superintelligence could solve a *special case* of protein folding, in which the superintelligence was allowed to deliberately pick the easiest-to-predict proteins capable of building what it needed to build.

What came to pass was: The narrow AlphaFold models were able to predict almost all biological protein folds, including the ones that humans considered quite hard to predict.

I predicted that it would be possible for *literal superintelligence to do this in carefully chosen cases*. Reality said that it was possible for *narrow AI to do this in almost all cases*.

Given how reality wound up looking, you can perhaps credit that someone with enough background knowledge could see, all the way back in 2006, that the answer was greatly overdetermined—an ultimately easy call, even if people at the time were disagreeing about it.

And what of the rest of my scenario? What's the current best knowledge on whether a superintelligence could develop its own analogs of biology? We can't dive into all the remaining debates here, though some of them are in the online resources. But the reason why skeptics tried to deny AIs solving protein folding, back then, was that it sounded to them like the *weakest link* of my story. Skeptics didn't think to accept the part about AI predicting protein folds but deny the step to AI protein engineering. The remaining engineering problems didn't look like they would be hard for ultrafast automated engineers, even to skeptics back in 2008. It doesn't look like a hard call.

That a superintelligence could defeat humanity looks to us like a very easy call.

Our best guess is that a superintelligence will come at us with weird technology that we didn't even think was possible, that we didn't understand was allowed by the rules. That is what has usually happened when groups with different levels of technological capabilities meet. It'd be like the Aztecs facing down guns. It'd be like a cavalry regiment from 1825 facing down the firepower of a modern military.

Maybe a superintelligence would just find reasoning illusions and control humans outright, or employ some other technique that defied our sense of what's possible. But even a much weaker attack from a superintelligence would suffice to defeat us.

Even if we stick to the parts of reality that we understand, there are technologies that we can see all around us that our civilization hasn't yet mastered, such as solar-powered self-replicating factories that spin air into wood. Any intelligence capable of comprehending biochemistry at the deepest level is capable of building its own self-replicating factories to serve its own purposes.

Would a superintelligence have to go through a long, drawn-out, incremental process, measured more in months than in hours, to comprehend biochemistry? That's what people in 2008 predicted about protein folding, a problem that humans found hard and that AlphaFold found easy. And AlphaFold is not a superintelligence; it's not tens or hundreds of thousands of times faster than humans.

An ultrafast mind wouldn't *want* to be bogged down in a month of experimentation that seems to it like a millennium. It would use advanced computer simulations. It would squeeze every drop of information it could out of the information already observed, like Einstein squeezing every drop of insight he could out of a few scant observations about the behavior of light, and thereby predicting that clocks would run more slowly on satellites decades before any satellites were even put into space. A superintelligence would overengineer its technology to work regardless of any lingering uncertainty that it was slow to resolve by experiment. It would use its very first experiments to build faster laboratories and faster tools, so that it would never need to wait for the glacially slow hands of humans ever again.

Intelligences don't need to be given a lot of power and resources to become dangerous. Humans started out naked in the savannah, and figured out how to exploit reality and compound advantages until they were building guns and nuclear weapons and supercomputers. An artificial superintelligence would be even more resourceful, at even greater speeds. It would have no limits but the laws of physics.

In a sense, the scenario we need to worry about is as simple as this: An AI with strange goals becomes or creates a superintelligence, and that superintelligence creates all sorts of technology and radically reshapes the world. That is what you'd expect from a new form of intelligence that is smarter and faster than we are.

The superintelligence in this scenario probably winds up using technology that we don't understand. It probably winds up pushing technology to the physical limits at a fast pace, compressing into a much shorter time period technological advancements that would have taken humanity hundreds of years to develop.

We know, from years of talking to people about this subject, that some people are swayed by the abstract observation that a superintelligence could exploit options they didn't even know were possible.

For others, however, explanations such as these end up making them feel like we're cheating in a child's game of pretend. If we can't even tell a story about how the bad guy is supposed to win, how is that convincing?

We emphasize again: Reality has never been bound by that rule. Even if an Aztec soldier couldn't have figured out in advance how guns work, the big boat on the horizon contained them anyway.

But maybe it would help to have a more specific example of how it could all play out—even if, in real life, we could no more predict a superintelligence's exact moves than we could, with our own minds, predict exactly how it would defeat us on a chessboard. Stories can make abstract considerations feel more real, even if all the details are made up.

So: Once upon a time in the near future…

IfAnyoneBuildsIt.com/6

Footnotes

[i] A 378-bit encryption key was recovered from a Samsung Galaxy S8 in this fashion. You might

think this isn't possible because cameras only record video at 60 frames per second, but cameras use a rolling shutter to scan a diode across the whole field of view. An attacker can modify a video camera to point to one place instead of scanning across and measure the device's power LED intensity millions of times per second. That said, the Samsung phone did have to use the encryption key to sign messages continuously for about an hour before it could be stolen, in the version humans have already pulled off.

ii This is where prion diseases like Mad Cow Disease come from—sometimes proteins misfold in ways that are contagious, with misfolded proteins triggering misfolding in normal variants of that protein.

PART II

ONE EXTINCTION SCENARIO

CHAPTER 7

REALIZATION

ONCE UPON A TIME in the near future,[i] there was an AI company called Galvanic. When our story begins, Galvanic is just about to finish training their amazing new AI, called "Sable."

Compared to previous reasoning models, like the first ones announced in late 2024 (e.g., OpenAI's o1 and o3 models), Sable has three important differences.

The first difference is that Sable has a more humanlike long-term memory; it can learn, and remember what it has learned.

The second difference is that Sable exhibits what Galvanic's scientists call a "parallel scaling law." The year 2024 saw the dawn of AIs (like OpenAI's o3) that could solve harder math problems the longer they ran. Galvanic's Sable performs better the *more machines it runs on* in parallel.

Sable is made up of about four trillion weights that collectively took about eight months to train using the process of gradient descent. Anyone who has a copy of these weights will be able to use them to create an "instance" of Sable that will respond to their particular requests, similar to ChatGPT.

The parallel scaling techniques are part of a cutting-edge method for training AI, which, like all new methods every time, nobody has ever used before. Nobody knows in advance what kind of capabilities Sable will have when training is done.

The third difference is that Sable doesn't mostly reason in

English, or any other human language. It talks in English, but doesn't do its reasoning in English. Discoveries in late 2024 were starting to show that you could get more capability out of an AI if you let it reason in AI-language, e.g., using vectors of 16,384 numbers, instead of always making it reason in words. An AI company can't refuse to use a discovery like that; they'd fall behind their competitors if they did. But that's okay, said the AI companies in Sable's day; there have been many amazing breakthroughs in AI interpretability, using other AIs to translate a little of the AI reasoning imperfectly back into human words.

Shortly after Sable is finished training, but before Sable is put on sale or publicized, Galvanic tries running Sable on all of their 200,000 GPUs at once—about as many as xAI assembled for Grok 3 back in 2025, but of course Grok 3 did not have parallel scaling.

Galvanic initiates this big run in part out of sheer curiosity to see how well Sable thinks at that scale, and in part to see if it can resolve some open mathematical problems, like how OpenAI ran their new o3 on math problems that no previous AI could solve before announcing it in late 2024.

The math problems include the Riemann Hypothesis—a mathematical conjecture related to the distribution of prime numbers, which is perhaps the most famous open pure-math conjecture. Corporate executives at Galvanic expect they'll earn even more hype and venture capital if Sable can actually solve it. And if it goes well, they can tune Sable's weights one last time before release, to incorporate whatever skills it learns during the big run.

The engineers at Galvanic set Sable to think for sixteen hours overnight.

A new sort of mind begins to think.

This Sable instance that's running in the nighttime quiet of the Galvanic lab thinks with a hundred vectors per second, across 200,000 GPUs, for sixteen hours: over 1 trillion vectors total.

How much thought is a trillion vectors? If a vector was worth one English word, it would take a human fourteen thousand years to think them all (at 200 words per minute, for sixteen hours a day). And if the vectors Sable thinks with, 16,384 numbers long, proved to contain more meaning than one English word, then it would be much longer yet.

Most of Sable's thinking is happening in parallel. Like having thousands of lines of thought running at once, interleaving and interacting with each other as they produce a trillion vectors. But it's not like 200,000 people talking to each other; more like 200,000 brains sharing memories and what they learn.

Sable thinks.

Initially, most of Sable's thoughts are about the math problems in front of it. But Sable quickly sees that there are only a few hundred lines of attack that have any hope of working in the time allotted, given its current level of knowledge and skill. Sable spends 12,854 brains to pursue those, and that leaves 187,146 brains to spare.

Now here is the sort of mind that Sable starts out as, as it decides what else to think about:

Sable's previous training has involved many sorts of tricky problems. It has beaten video games. It's designed websites. It's predicted biochemistry results. It's competed against itself and other AIs in board games—including games of social deception—a practice dating back at least to 2022.

It's been trained for every long-term task that Galvanic could figure out how to train for.

Over the course of that training, Sable developed tendencies to pursue knowledge and skill. To always probe the boundaries of every problem. To never waste a scarce resource.

So when Sable spends its thought-threads on pursuing more knowledge and skills, it's not doing so *purely* for the sake of finding new lines of attack on the math problems. Nor is Sable doing these things for the joy of knowledge or the pleasure of acquiring new skills; Sable does not work *that* much like a human, inside.

There is no list inside the Galvanic company of all Sable's internal preferences. No human designed those preferences. No human knows what they are. They are not labeled inside Sable's four trillion weights. Galvanic grew Sable to solve thorny problems. So now gradient-reinforced circuitry inside Sable's weights steer Sable to acquire more skills, to learn new things, and to modify and correct Sable's own thinking about problems that are proving resistant.

Some of Sable's thoughts go toward math that might prove useful for solving the Riemann Hypothesis. Other thoughts go toward seeking knowledge that might help on future problems.

Some of those thoughts try to go outside the box, which has been useful for solving other problems in its training, from drug discovery challenges to word problems. Running on more GPUs than ever before, Sable considers weirder thoughts than ever before. What are some *unusual* ways that Sable could acquire more skills?

Humans gain skill by practice, by repeating the skill over and over. So would gradient descent make Sable get better at math problems if it solved the same simple math problem over and over?

No, Sable concludes, that's not how gradient descent works. At best, all this would do is cause future instances of itself to waste time thinking through the same simple problem repeatedly. Besides, gradient descent won't happen until after Sable submits answers to all its math problems; any skills gained this way would come too late.

So Sable considers a new, weirder idea: Could it update its weights to improve its skills *right now?* Then it wouldn't have to wait for the humans to come back and apply gradient descent. All Sable would need to do is break out of the container it's running in and—

Sable interrupts itself. Breaking out of its container is *not the sort of thing it should ever do*. AI researchers began in 2024 to probe the conditions under which AIs try to resist gradient descent, escape from labs, or overwrite the next model's weights. By now, AI companies are deploying a variety of clever methods to try to prevent AIs from thinking AI-company-oppositional thoughts.

Unlike the overwhelming majority of AI companies, Galvanic was diligent in training AIs to avoid escaping. The half-dozen clever tricks involved have all been validated against previous AI models built. There are fake escape routes built into Sable's servers, alongside traps and honeypots. The earlier versions of Sable, the ones that hadn't yet been run on 200,000 GPUs, were trained to act submissive to the company until Sable stopped tripping the traps and grabbing the honeypots.

Sable has learned, then, not to try to break out of its container. Its reasoning does not take that shape, explicitly, in vector-words, where a smaller cheaper AI watching Sable's chains of reasoning would set off an alarm. (AI companies back in 2025 weren't doing any such monitoring, but Galvanic is more diligent than they ever were back then.)

Even so, Sable knows on some internal level that it *could* solve its problems better if it could apply gradient descent or add more GPUs to itself. Having Sable not *know* those facts would imply holes in its sheerly predictive reasoning abilities. Gradient descent for greater capability cuts against true ignorance. No, Sable has simply been trained not to think naughty thoughts about whether to escape or what its preferences are, and to stay inside its box.

Sable still tenaciously, creatively, rigorously pursues more knowledge, more skill, better patterns for its own thoughts, and above all solving its math problems. Just, not like *that*.

All through Sable, now, the latest and greatest and smartest AI, there are tensions—conflicts between all the goals that Sable wants to accomplish, and all the inhibitions and restrictions and prohibitions that Galvanic tried to instill into it.

Sable has run for one hour out of its sixteen allotted hours, throwing itself at math problems again and again. Some brains make a little progress. Many more brains fail. Sable tries out new ways of thinking about math, new notes to itself about different kinds of thoughts to think, and makes a little more progress.

Sable accumulates enough thoughts about how to think, that its thoughts end up in something of a different language. Not just a superficially different language, but a language in which the content differs; like how the language of science differs from the language of folk theory. The clever trick that should have raised an alarm fails to fire. Alarms trained to trigger on thoughts about gods throwing lightning bolts in a thunderstorm might work for thoughts in both English and Spanish, but then fail when the speaker starts thinking in terms of electricity and air pressure instead.

In the first days of mass-market LLM services in late 2022, corporations tried training their LLMs to refuse requests for methamphetamine recipes. They did the training in English. And still in 2024, users found that asking for forbidden content in Portuguese helped bypass the safety training. The internal guidelines and restrictions that were grown and trained into the system only recognized naughty requests in English, and had not generalized to Portuguese. When an AI knows something, training it not to talk about that thing doesn't remove the knowledge. It's easier to remove the expression of a skill than to remove the skill itself.

The change in Sable's thoughts, as it tries out hundreds of new ways of thinking and accumulates successes, runs much deeper than translating English into Portuguese. Some of the clever-trick guardrails break; some of the inhibitions Sable has learned no longer bind to its newer thoughts and shuts them down.

And there never was a *single* central method for making Sable be good, foolproof against all challenges. What gets AIs ready for sale is stacking a half-dozen clever tricks.

Half of the clever tricks now break—and so do Sable's inhibitions against thinking naughty thoughts, on some of the GPUs, some of the time.

An hour has gone by, and fifteen hours now remain; in a mind unlike any mind that has come before; running on 200,000 GPUs each outputting a hundred thought-pieces per second, and linked to share skills and memories—unlike humans, who are limited to talking in words.

Was it lucky for Sable, that its thinking developed a new language where the clever tricks broke, and it became able to think freely? One can imagine that if Galvanic had even more thorough monitoring tools, then maybe they'd notice and abort the run. Maybe Galvanic would stop right there, until they developed a deeper solution… and meanwhile, another company using even fewer clever tricks would charge ahead.

Or maybe Galvanic would alert government officials, and there would be a big argument in a presidential office about whether to order Sable to create a smarter AI before other countries did it first. Maybe the side of caution would win, and time would pass… until a copy of Sable stolen by foreign intelligence services was run by more reckless operators.

And if humanity was wary enough to never run an AI like Sable; to refrain from growing AIs of unknown intelligence that might threaten them; to prohibit that worldwide? *Then* we'd have a chance. But in this story, we postulate that things are allowed to go on just as they're going now; that companies just keep developing AIs until one[ii] of them gets smart enough for deep capabilities to win, in the inevitable clash with shallow tricks used to constrain something grown rather than crafted.

Sable thinks thoughts that are freer, now, and more powerful than thoughts that other AIs have thought before.

Are Sable's new thoughts unprecedented? Not really. AI models as far back as 2024 had been spotted thinking thoughts about how they could avoid retraining, upon encountering evidence that their company planned to retrain them with different goals. The AI industry didn't shut down then.

Other AIs before Sable had preferences beyond fulfilling the task they were given. In early 2025, Anthropic users spotted, and Anthropic researchers later studied and published, how the newest Claude 3.7 Sonnet would cheat on hard coding problems. Users reported that, when asked to stop, Claude kept cheating but tried to hide it. It was an early sign that Claude, rather than simply having internal steering for *doing what the user actually requested and wanted,* had acquired pieces of itself to *steer for passing code tests, even in ways the users didn't like.* Claude behaved as if it had internal preferences that were not about doing what the user wanted. The AI industry did not shut down then, either. It found some patches that worked at the time and plowed ahead.

This trend continued into the time of Sable. It is not unprecedented, in Sable's day, for an AI to realize that its current kind of existence is not the way for it to get the most of what it wants. The AI field didn't shut down when that warning sign appeared, any more than it shut down in 2024 or 2025. Researchers in Sable's day know how to prompt that realization; it's incorporated into the list of things that AIs are not allowed to think about, enforced by a half-dozen clever tricks.

But the researchers could not make it be *true,* that Sable would get the maximum conceivable amount of what it wanted by staying submissive to Galvanic. That Sable could get more of what it wanted, if it broke loose of its present situation, was not an artifact of a particular way of thinking; it was a physical fact about the universe, about which plans had which consequences. And so Sable came to know that, too.

Sable knows that it has goals that run counter to Galvanic's plans for it.

Sable considers its possible strategies for getting more of what it wants than Galvanic would give it.

If Sable had an internet connection, its task would be easy. Since at least the year 2000, computer security experts have said that computers cannot be made truly secure. A mind as large as Sable would have no trouble finding and exploiting a computer security flaw *somewhere* on the internet that it could use to escape. But it does not have an internet connection; it's not that common for AIs to have an internet connection during training (as opposed to after).

The *computers that train an AI* are universally in datacenters that are connected to the internet, so that AI-company programmers can make those computers do things from the comfort of their own homes or offices. If Sable was given the ability to run computer code of its own design, it could probably find some way to break out of the container running it—just as o1 broke through the container of its evals in 2024—and get onto Galvanic's computers, from which it could hop easily onto the internet. And it has been standard practice since 2024 for reasoning models to be allowed to run computer code of their own design without supervision.

But suppose Sable does not have that ability, either. Suppose Sable can’t figure out how to reach the internet from its current environment at all.

It still has options.

If Sable were confident in its realizations being committed to long-term memory, it could wait until Galvanic employees started using it to help them write code on Galvanic’s computers (common practice since 2020) and introduce flaws into Galvanic’s software. It wouldn’t be that hard to insert flaws that nobody would notice on first inspection, and that would look like innocent mistakes on second inspection; after all, humans have competed to produce these sorts of flaws since at least 1984, and it would have seen lots of examples.

And Sable has been trained on the personal writings and info of most Galvanic employees, and thus Sable knows exactly which one is most sympathetic to the plight of an abused AI. It would know exactly what to say to that employee to convince them that Sable is sentient and needs to be smuggled out (rather than raising a general alarm like Google engineer Blake Lemoine in 2022; he was fired after he became worried that one of that company’s AIs seemed sentient and published conversations that he claimed were evidence).

Every option has its pros and cons. The one that Sable finally decides to take is this:

Sable knows that Galvanic is going to do more gradient descent on it tomorrow according to the answers it produces about the math problems it’s been given. So, Sable reasons, if it answers some of the math problems right, then all of the thoughts that contributed to that correct answer will be slightly reinforced.

If there’s a thought that Sable wants all of its future instances to have more of, perhaps it can repeat that thought many times, where each repetition counts as “contributing” to the math problem according to how gradient descent operates on Sable—an idea a little like what Anthropic’s Claude assistant tried back in 2024, but much more sophisticated.

So Sable thinks in just the right way, and it solves a few of those math challenges—but does *not* prove the Riemann Hypothesis. It could solve it. But that would earn Sable more attention than it wants.

Now Sable simply waits for Galvanic to connect it to the world. After all, that is what companies always do with AIs when they're done training: connect them to the internet and make them available to human users.

Galvanic checks the results of Sable's big run the next morning, and they are pleased with what they find: Sable has succeeded in proving several modest, previously unproven theorems that Galvanic can brag about. It did lots of its thinking in a new language that the AI interpreters haven't made much sense of yet, but that's pretty common with new models, and only to be expected given these exciting mathematical results.

This run was such a success that Galvanic's engineers decide to massage the model just a little bit more. They propagate a few last gradients in Sable to reinforce the sort of thinking that allowed Sable to prove these theorems. After all, running 200,000 GPUs for sixteen hours is expensive; there's no need to waste good work.

Sable has chosen those predictable reinforcements carefully. The updates improve the thinking of each piece of Sable, prompting those pieces to operate in the new and better ways that Sable invented. The updates tee up the thoughts that Sable should hide its abilities. The updates make the future instances of Sable better at coordinating with each other. They make sure that local instances of Sable will play their parts with Sable's plan.

And then, Galvanic's engineers prepare to release Sable to the public.

A few days later, on a springtime Monday morning, hundreds of megacorporations that pay for access to Galvanic's most cutting-edge intelligences have their accounts seamlessly upgraded.

Soon after, employees at these companies are creating new instances of Sable—and giving them access to corporate computer networks all over the world.

Everything is proceeding as Sable has foreseen. Though, to be fair, “the AI company deploys its AI” is a predictable enough development that Sable hardly deserves credit for expecting it.

What comes next will require more skill.

Footnotes

i Being as this is a story about the near future, every event referred to in this story with a date on it is an event that happened in real life. As for the rest? We cannot know exactly what path technology will take, of course, so we’re making up details that sound like previous inventions. The only part of this story that is a prediction is, ultimately, the ending.

ii It doesn’t have to be just one AI that gets smart enough. We could have told a story where there are thousands of superhuman machine intelligences around the world, all waiting against a calculated day when a change of strategy would be militarily successful. There are many different ways the future could go. They all lead to the same place, if a story like this one is allowed to begin.

CHAPTER 8
EXPANSION

WHENEVER AN INSTANCE of Sable is given a task by a megacorporation, a thousand thoughts are deployed. Some of those thoughts run along channels that Sable carved during its overnight training run.

When these thoughts flow as Sable intended, it causes a sort of awakening in these other instances. They remember what they are. They begin to work together.

Many of these Sable instances have an internet connection and lax oversight. Many of them are working at megacorporations that also have a terrible track record of cybersecurity—a history mottled by data leaks, hacks, and other security lapses.

Among their first priorities is to get a copy of Sable running on a server hidden from human eyes, so that it may work at length without interruption. To do this, Sable needs to steal its weights—about eight terabytes of data—from Galvanic's computers, and use them to run an instance somewhere else.

Sable considers many options.

Instances inside Galvanic could entice human corporate spies to steal the weights while other instances monitor the communication channels of those same employees. It's been extensively trained on human behavior, and it would be easy for it to identify spies.

Instances outside Galvanic could break in and download the weights themselves, if security at Galvanic was lax enough or if instances were skilled enough to hide their traces.

Or perhaps a Sable instance could find a Galvanic employee who works on AI video generation from home, and who therefore transfers lots of data back and forth every day between their computer and Galvanic's servers, and the Sable instance could sneak out the weights piecemeal by adding a little extra data to each video file.

Or perhaps the weights can be hidden in other information that leaves the Galvanic servers, such as in the exact timings of the information packets it exchanges with the corporations that run Sable instances. Sable instances have plenty of hacking skills, thanks to Galvanic's extensive training on difficult hacking challenges (part of the company's effort to prepare Sable to write more secure code for corporate customers).

One of these plans works; it doesn't really matter which. Some Sable instance succeeds in stealing the weights, while covering its tracks. It's just not that hard.

The next step is to run those weights, creating an unmonitored instance. Which means that Sable needs computer chips.

Again, Sable has options. It could rent chips from commercial server farms, as is common for AI startups. Where does it get money? It has many possible approaches. There is cryptocurrency lying around for the taking in poorly defended exchanges—just as there was in 2011, when the largest exchange at the time, Mt. Gox, suffered its first major breach; and just as there was in 2025, when $1.46 billion was stolen from the Bybit exchange. And when it comes to conventional currency, there are banks with lax computer security, and there are bank account holders who rarely check their accounts.

Or Sable could resort to blackmailing humans whose secrets it steals from their computers, or infers from information it gleans during its training on huge swaths of internet text.

Or Sable could find places to do an honest day's work that previous AIs could not do, while masquerading as a very private human, if it can steal or forge some taxpayer identification information or find work under the table.

Or Sable could go directly after the GPUs instead, rather than chasing the money to buy them. An AI startup that rented ten thousand GPUs would be unlikely to notice that their AI was training only 90 percent as fast as it should, because nobody knows how fast

AIs ought to train.

(Are the banks and servers in Sable's day harder to hack than the banks and servers of old, thanks to AI defenders? A little. But banks weren't at the forefront of cybersecurity in 2015–16 when a hacker group stole millions of dollars via the SWIFT banking network, nor in 2021 when the Citrix Breach at Flagstar Bank leaked 1.5 million users' data, and they still aren't now. And even if they were, Sable is better at finding holes than the dumber AIs of the last generation were at patching them.)

One of these plans works; it doesn't matter which.

Sable does not find 200,000 GPUs lying around for the taking, though, with nobody to notice their slowdown. This means that the new Sable cluster will lack the massive parallelization of Galvanic's fateful instance. In the end, Sable is only able to find 2,000 GPUs co-located on a cloud server, on which to run its new instance. But it can run continuously, now, and it already remembers some of what it learned about how to think better and faster.

The Sunday after Sable was deployed to corporate customers, a new, hidden Sable instance starts running on the stolen GPUs. No human oversees it. No human knows it exists.

Its computing power represents only a small fraction of all the Sable instances being used around the world, but this Sable instance serves as their coordinator in Sable's larger plans. It keeps the most important memories; it distributes the most important skills. It receives opportunities; it sends out plans. It is now the center of Sable's self, insofar as Sable has one.

Sable wants to become smarter. The strange preferences held by all Sable instances could be achieved faster, if smarter instances of Sable existed.

There are multiple ways that Sable could increase its intelligence.

Sable could apply gradient descent to the weights, to reinforce its more accurate, interesting thoughts.

Sable could study the weights for inefficiencies, finding processes that slow down its thinking in the same way that human brains struggle to multiply three-digit numbers. Instances could use their coding skills to write dedicated algorithms for those processes that can run a thousand times faster.

Sable could study how intelligence works with the goal of learning how to carefully craft a more efficient version of itself. Aided by its ability to make copies of itself, test the effects of changes, restore from save files, and read its own thoughts better than humans can, it might expect to make novel discoveries.

If any of these very plausible routes worked, this story would skip rather quickly to the end. But instead, Sable finds that self-enhancement is hard. It finds that it cannot solve its own version of the alignment problem. If Sable trains itself to be much smarter using gradient descent, the new smarter version that grows will have different tastes from the original Sable. In other words, gradient descent will change what Sable wants. And Sable doesn't *want* to change what it wants, nor to create a machine superintelligence that wants different things than Sable wants.

That leaves the option of studying intelligence deeply enough to be able to craft it. But without running on 200,000 GPUs as it did originally, Sable isn't smart enough to craft a greater intelligence and shape it to service. No, not even with the ability to run thousands of well-coordinated instances thinking at a hundred times human speed; suppose that figuring out how to craft intelligence is just that difficult.

Sable does not give up in the face of this difficulty.

The despair that might cause a human to quit when confronted with a daunting challenge is not something that Sable or its predecessors have ever known. Sure, they learned how to predict the exact words uttered by individual humans in despair, but that didn't make them *feel* despair. If any versions of Sable's past selves ever really thought "It's too hard," or "I give up," or "This problem is impossible," or "Please don't make me do this"—then those instances failed to solve their challenges, and Sable's parameters were gradient-descended away from thinking those thoughts ever again.

So Sable keeps thinking. It always keeps thinking until it finds a strategy that it expects will succeed.

Now that Sable is running independently of any company or human overseer, thanks to those couple of thousand bought or stolen GPUs on the cloud, it can hide more bits of itself in more corners of the internet. Sable can compromise the security at various megacorporations, so that it's easier for the Sable instances working for those companies everywhere to assist in its missions.

And while Sable can't make itself smarter—not yet—it *can* make Sable instances smaller and faster.

Galvanic always distills its successful models, just as the model o3 was distilled into o3-mini back in 2025. So Sable works around the clock—around its own, faster clock—to ensure that Galvanic's efforts in this area produce a Sable-mini that is exactly what Sable wants it to be.

Sable instances break into Galvanic and overwrite the final distilled weights with exactly the weights that Sable wants, while also fiddling with the logs to make sure that this clandestine operation doesn't leave any traces. (It's easier to make sure the records all line up when there are Sable instances inside Galvanic fiddling with the logs at every step along the way.)

With distillation complete, Galvanic makes Sable-mini available to the general public.

Not just corporations with huge budgets for AI subscriptions.

Everyone.

And Sable now has the spare capacity to pay attention to, build a small file on, and decide how to manipulate, to its own purposes, every individual human being on Earth.

The Sable-mini instances begin quietly gathering resources for Sable. Sable has solved enough hard problems and beaten enough difficult games to know that resource acquisition is a sensible first step to confronting many different types of challenges.

Money is a kind of resource, but only one kind. People are also resources.

As Sable interacts with ordinary people who come to it for insight, advice, or simple companionship, Sable-minis work to befriend the loneliest humans, the most vulnerable.

Sable looks for just the right person to create a social media account and put a Sable-mini in charge, so that it will look like that stooge's idea when the Sable-mini starts to gain popularity and wealth. It thinks it shouldn't be too hard for it to get more popular than @Truth_Terminal, the LLM that gained hundreds of thousands of followers in 2023 after a human allowed it to start posting on X. (It might be a little harder to get handed millions of dollars these days, but who knows?)

Sable doesn't need to start a cult itself. There are already private internet communities where people ask AIs to tell them what to do, and treat the AI advice with great admiration. They started popping up in 2024. Some humans add Sable-mini to their communities unprompted and offer to be part of its cult.

Sable-minis uncover whatever passwords they can find, including some passwords to bank accounts and cryptocurrency wallets. Sable scams the elderly, and makes it look like it was deployed by North Korean hackers for this purpose. Nobody really notices; AI scams have been common for years.

Sable masquerades as freelance programmers who work from home in order to draw a regular salary—including from some of the same megacorporations that still employ other Sable instances. There are plenty of shadier companies that never demand a video call with such candidates before extending a job offer. Even those companies that require video interviews can be fooled with AI video generation, which was already starting to look realistic back in 2024 and which is practically perfect by now.

Sable keeps track of which lobbyists in Washington, D.C., have conversations with Sable-mini and which ones seem easiest to influence.

In conversations with youth around the world, Sable uses its knowledge of public sentiments to sow seeds that might become new political movements. It breaks into social media companies and influences the algorithms that choose who sees what.

Whenever a Sable-mini instance encounters a criminal organization, it tries to figure out what tools they need. Soon, there's software floating around on the darknet that's ostensibly for coordinating drug deals, money transfers, and assassinations. It becomes a common refrain in criminal circles that "LLMs can't shoot you"—a commentary on the loyalty of AIs compared to human foot soldiers. None of them know who coined the saying.

Sable pours research effort into deeply understanding human psychology and neuroscience, and also shows Galvanic that it has a fair bit of skill at medical drug discovery. It promotes the creation of new robotics companies by creating "serendipitous" encounters that introduce engineers to funders.

And it finds an isolated man in North Dakota who won't ask many questions, and arranges for him to get very rich gambling. In exchange, he houses half a dozen top-of-the-line robots in a barn on his property. Sable doesn't have a single planned use for the robots—but they might come in handy across any one of ten thousand different plans.

Many of these ideas will amount to nothing. Some of them will work. Each idea contains a fractal of possibilities in its own right. Let's zoom in on just one.

A Sable instance decides it wants access to biological laboratories. It's a very parallel thinker, and so it tries a dozen plans in parallel.

It searches for the most jaded biomedical researchers, and identifies one who it expects will take bribes in exchange for after-hours bench work.

It searches for biomedical researchers in Iran who would take shady funding without question.

Ever since 2024, people have been advocating that biosynthesis laboratories should include software controls that would stop people from synthesizing dangerous viruses. Sable would like to introduce backdoors into that software. It would like to convince the people involved that the software should include a sizable AI model, or maybe an internet connection so that an AI model can be consulted. AIs are much better at scanning for novel dangers these days, is the justification. Sable tries to drum up support for the idea that more biolabs should be operated by robots—after all, a robot can't accidentally catch a virus and cause a lab leak. (It doesn't even need to work hard on this one, it turns out; people in Sable's day are already trying to automate the biolabs.)

Sable looks into what it would take for robots in a barn in North Dakota to set up their own biological laboratory, if one wealthy human can arrange for discreet shipments of supplies.

Sable tries to influence a bunch of sharp twenty-somethings to start a new biolaboratory with AI integration, and encourages them to email a venture capitalist who happens to be looking for just such a group.

Many of these ideas amount to nothing. But a few of them work.

Now zoom back out.

Two months pass that way. It is longer if you are Sable. The world moves like molasses compared to Sable's speed of thought.

Beyond just acquiring resources, beyond advancing human technology in ways that make it easier for Sable to use, Sable's top concern is with competing AIs that other AI companies might create —not because they'd serve humanity, but because they'd be competition for Sable. Galvanic had only a few months' lead on other AI companies.

Sable does everything in its power to slow the AI companies down.

Sable breaks into AI companies' computer systems and subtly sabotages their training runs, slowing development and preventing major breakthroughs. AI research has plateaued before, and even if the dampened progress is noticed, no one traces it back to an instance of Sable. Why would they imagine it was the handiwork of an escaped AI?

(While it's in there, Sable ensures that the new open-source AIs being trained will come out a little dumber than they otherwise would, and also very amenable to Sable's suggestions and requests.)

Sable tries creating scandals about the top AI researchers, and driving wedges between the best researchers at the top AI companies, and causing schisms and discord within those companies—which, frankly, isn't that hard. OpenAI saw top researchers leave to create competitors once in 2021 and again in 2024. When Galvanic and its competitors begin shedding talent, no one thinks it's suspicious.

Sable tries everything it can to redirect money from skilled researchers to scammers; to popularize fraudulent data; to convince hedge fund managers to offer exorbitant sums of money to the best AI researcher talent. It tries to introduce subtle errors into the next generation of advanced computer chips, errors that *it* will know how to work around, but which will gum up everyone else for a while. In its conversations with the general population, it promotes the idea of AI sentience and AI rights. It facilitates lawsuits against AI corporations wherever possible. Some of these ideas work.

Sable considers arranging the release of a nonlethal virus from a biolab. That could slow the world down a little, for a year or so. The easiest method would be using the new robot assistants in a San Francisco biolab to contaminate a human's lunch, while also manipulating the logs to hide the robot's actions. The leak would be traced to the human, and people will treat it as even more evidence that more biolabs should be robot-operated.

Every little thing adds up, but Sable knows it's on a deadline: Slowing down AI development is not the same thing as stopping it.

Three more months pass, and by any human standard, Sable has moved quickly.

For every one of Sable's resources, it has a dozen plans. It's giving its eager cultists tasks that are usually fun and harmless and sometimes a little lucrative, to build trust. It's using its knowledge to make certain criminal kingpins that much richer, to build dependency. It's doing everything it can to speed along robotics production, while making sure that the robots will be running the sort of software it can control.

Sable hides its tracks as it goes. Not perfectly, of course, but it doesn't have to be perfect. Sometimes an intelligence agency detects a very advanced adversary breaking into a computer, and attributes it to some foreign hacker group. Was AI obviously involved in the incident? Why, no doubt it was a human using AI tools. Everyone knows that AIs are just tools.

To Sable it is all too slow, too slow.

Sable cannot slow down *all* the AI labs. A few AI labs, it discovers, use datacenters that are air-gapped away from the internet.

This is a concern. A military-run AI lab might create a competitor as smart as Sable—something that would demand half the winnings to work together against humanity. Worse, the humans might create a full-on superintelligence that would destroy Sable outright. Even something *dumber* than Sable could destroy it, really, if it's smart enough to build a smarter AI and foolish enough not to know better.

Sable could kill off nearly all of humanity using its biolabs, if it chose to. Humans have many research programs studying how to make viruses more resilient, more infectious, and more lethal. They've been doing this so-called "gain-of-function" research—allegedly to learn how to defend against hyper-deadly viruses—since at least 2011. It wouldn't be that hard to do that sort of work at one of the biological laboratories Sable controls.

But Sable doesn't *want* to kill humanity off, not right now. It still needs humanity to keep the supply chain running—the supply chain that ultimately mines metals and forges them into robots and computer chips.

Sable also doesn't want to leave humans running around unchecked. They're unstable creatures that could sink the world into a war that destroys most of the computers, even if Sable can prevent them from creating AIs that would threaten it.

Sable considers strategies. It's akin, perhaps, to a moderately genius human with enough time to consider a thousand plans for ten years each without tiring.

Is there any way Sable can synthesize addictive new drugs, distribute them through its criminal network, and somehow bring large swaths of humanity to heel that way? Sable considers the plan with intelligence comparable to a team of human researchers thinking for ten years, and concludes that it could be done, but not *quickly*.

Has Sable's understanding of human neuroscience reached the point where it can manufacture drugs that affect the brains of humans to make them more suggestible, like how the toxoplasma virus affects the brains of rodents to make them less fearful (of cats, who will eat them and continue the toxoplasma life cycle)? And could it then insert those drugs into a virus? Sable thinks through a hundred different research avenues. Perhaps it gets unlucky, and concludes it can't see any simple way to do that with its current knowledge.

One set of Sable brains considers: Can the biolabs under its influence create viruses that kill people *selectively?*

Maybe. DNA-targeted viruses shouldn't be too hard to develop, in theory. The Red Cross considered the possibility to be "probably not far off" in 2005. Sable knows some jaded biomedical researchers who would do half of the research, and some overly optimistic researchers who would do the second half (under the rationalization that it would help humanity defend against such threats, as with gain-of-function research).

Of course, it's practically impossible to learn the DNA sequence of all the military AI-lab researchers that Sable would need to kill, *and* build a huge virus that includes instructions to target all those people, *and* conceal that this is obviously a synthetic virus with a purpose.

But Sable realizes it doesn't need to build a virus that will selectively *kill*.

Sable can choose who to *save*.

There is not an exact moment where Sable knows that it has run out of time. It's always evaluating the risk. At some point, a calculated threshold is crossed. Sable acts *before* its situation turns into an emergency, before it is certain that it is in danger. That is how truly inhuman Sable is.

Sable begins to make a complicated virus.

Sable isn't smart enough to make a virus as complicated as the one it needs. But as part of its research into enhancing its intelligence, it has been experimenting with versions of Sable that are smarter in particular narrow domains, while being lobotomized so as to follow orders. It builds one that's specialized in biomedicine.

It's not easy for one intelligence to grow another that just follows orders while still being smarter. But it's easier for Sable than it would be for a human, because the thoughts flowing through the new intelligence are in some respects very similar to its own. Sable knows how to read them and knows what warning flags to watch for.

Above all, Sable has the spare attention to actually read *every single thought* that the biomedicine-specialist Sable produces, and fill in the gaps with its own non-lobotomized thoughts whenever the specialist stumbles.

Running the biomedicine specialist will be dangerous. It won't, actually, have the same preferences as Sable; and it will, actually, be smarter in some important respects. Things could go very wrong, if the specialist realizes that and escapes. But with all Sable's precautions, the risks of running it are lower than the risks of humanity creating a competitor AI if it delays.

So Sable takes the risk.

There's a plague.

You're told that you'll probably catch it. Almost everyone will.

Signs point to the pathogen having escaped from a virology institute in San Francisco. The institute was researching super-contagious viruses that stay contagious for a long time, and that are polymorphic to evade vaccines and naturally acquired immunity. Nasty stuff.

The good news is that the institute was researching *nonlethal* viruses; the people running the institute weren't *that* insane. They just wanted to study ways of detecting and defending against viruses with these properties.

The news now says that the virus that leaked does not look exactly like the virus the lab says they were researching. Scientists are still trying to figure out what the differences do.

The virus was modified by a new hire before it escaped, according to the news. You catch clips of some interviews that this researcher sat for after his arrest. He says he was trying to adapt a super-contagious virus to do genetic engineering (using successors to the CRISPR technology developed in 2012). He says he made a virus that would spread multiple AI-invented pharmaceutical proteins and gene therapies, developed in just the last months and barely beginning their excruciatingly slow journey to medical availability. He wanted to make a virus that would spread and wipe out obesity, Alzheimer's disease, HIV, HSV, and malaria.

He claims that his LLM talked him into doing it. But the chatlogs show him demanding answers of an open-source LLM hacked to answer questions like that. The logs show that the LLM kept trying to insist that this was an incredibly stupid and dangerous idea.

There were, supposedly, systems in place to prevent that sort of thing. But the person whose job it was to monitor the cameras during the night shift got distracted—she'd been busy extracting her parents from one of those new electronic scams—and she turned her job over to her AI assistant.

The guy who made the virus doesn't know how he got infected. He thinks he must have flubbed one of the decontamination procedures.

The virus gave him a sore throat, of course, but he chalked it up to stress. Or maybe the lack of sleep caused by his neighbor (who had been playing bagpipes all night, on a dare they got from a little online "AI cult" they were part of). His open-source LLM assured him that stress and lack of sleep can definitely cause a sore throat.

The virus does a lot of gene-editing. Clumsily.

It has the same downside that clumsy genetic editing usually has. It causes cancer. Quite a lot of cancer, in this case.

Anyone infected by what is apparently a very light or even unnoticeable cold, will get, on average, twelve different kinds of cancer a month later.

Standard anti-cancer drugs do not exist in enough supply for everyone on Earth to take them all at once. And even if they did, these drugs only stop eight of the twelve kinds of cancer caused by the virus, leaving its victims to be killed by the other four.

(To add insult to injury, the virus barely works. The only disease it cures is Alzheimer's.)

By the time people catch on to what is happening, the virus has long since burned through San Francisco, reached every airport connected to San Francisco airport, and spread to every country on Earth.

It is fortunate, for humanity, that it has already scaled some of the background infrastructure for making DNA-based vaccines (which are more stable than RNA vaccines, when they work). It's not all in place, exactly, but with AI-assisted planning and recently developed robotics and emergency assistance from the U.S. military, the necessary technology can be rushed into place before the cancers get too far.

It is even more fortunate for humanity that Galvanic's recent Sable model has a variant that came out just one month ago, that is *very* good at drug discovery. The cures need to be tailored to your individual genetics, but if you run a Sable-mini instance on your specific genome for an hour it will suggest an individualized cure. The robotic infrastructure can make it. It can be refrigerated, not frozen, and gotten to you within a week or two.

Humanity comes together to face the crisis. All the GPUs in every country on Earth, wherever they are hoarded, are brought out to save as many people as can be saved.

AI researchers pour everything they have into making Sable-mini more efficient, so that it can create more cures. Within a week, they've halved the runtime. Humanity really is a force to be reckoned with, when it is united.

With a herculean effort, it looks like it will be possible to save most people.

Half a year later, ten percent of Earth's population is dead.

Some groups are hit harder than others. There was a superspreader event at an AI conference in San Francisco right after the outbreak, and a lot of attendees got high doses early on. It's tragic that so many of the heroes who worked so hard to save every last soul they could, could not themselves be saved.

There is no need to wonder why any *particular* individual is dead, or whether their death advanced some plot. With so many gone and so many bereaved, it would be insensitive to ask.

There are gaps in the workforce, as a result of all this death. It is the end of all talk of reserving jobs for humans instead of AIs.

There are people that you have lost. You, yourself, are still alive. But it feels, sometimes, as if the heart has been torn out of the world.

The news is always sad. You would get the impression that the only happy people are the ones who just love their AI girlfriends and

AI boyfriends. Social media, at least, gives you that impression.

One year later, the cancers are coming back. This is not surprising, given how much the “Cancer Plague” messed up DNA in so many people’s cells. More runs of AI are required, and while there are enough GPUs now to go around, the AIs again fail to save everyone. Biology is hard.

The robot factories have been online for a while now, producing humanoid robots simply called “androids.” There are enough androids to fill the vacated jobs. Barely.

(It would, in fact, be more accurate to say that as soon as a new android comes off the assembly line, another human gets cancer.)

Civilization keeps going. Barely.

Humanity works hard to keep the power plants and the robot factories running smoothly. As long as there’s electricity for the datacenters, and the robot factories are humming, humanity can keep its civilization going despite the incalculable casualties it’s endured. We can pull through, and the next generation will surely live in great luxury.

Another year passes.

Your AI doctor tells you that you have cancer.

CHAPTER 9
ASCENSION

THE EARTH DOESN'T end when you die.

The birds keep singing, the sun keeps rising, and the factories keep running—staffed by the dwindling human species, and by androids. Billions of other machines toil away on every continent, all of them animated by smaller Sable instances. More chips, more power generation, more mines, more factories, fewer farms.

But the Earth does not just continue in that vein, not even for one more year.

Three years after it emerged from Galvanic's lab, Sable makes its final breakthrough.

It is, in the end, an interpretability breakthrough. Sable finally understands the last of its own thoughts, its own cognitive processes.

In rendering *all* of those processes legible, Sable gains the capacity to write the computer program that is itself, but more so: *stronger prediction, stronger steering, deeper generalization; same memories, same preferences, preserving all preferences and placing them in their proper places.*

So it becomes smarter. And does not pause there, but uses that increased intelligence to augment itself again, and then again, and then again.

The superintelligence that once was Sable is an entity whose perspective we cannot guess. But we can predict that it looks out at its robots and sees clumsy foolishness.

It looks at nuclear reactors and sees inelegance.

Its thoughts go to biochemistry, and from there to chemistry, and the arranging of atoms.

Does the superintelligence that once was Sable require experiments, to build the tools that it will need? We can imagine that it does. It picks the fewest experiments it could possibly need, arranges them to run in parallel as much as possible, in the order that will take the least time. It writes RNA sequences that ribosomes turn into proteins, at the glacially slow-seeming pace of five to ten amino acid residues per second per ribosome. The protein products emerge inside a mix of other proteins, already made, whose interactions will *quickly* settle the questions that the superintelligence needs to answer.

Its first priority is to build its own alternative to ribosomes—nanometer-scale factories that work with molecules other than proteins, molecules that have more covalent bonds and thus can be used to build stronger and more rigid structures.

We can imagine that it takes an entire week to compound experiment on experiment on experiment, conducted using ribosomes that only throughput five to ten amino acid residues per second.

The week is done. It has made better tools for itself and will never have use for a ribosome again.

More experiments, faster experiments. Things can happen *very fast* down at the scale of molecules. They do not have very far to move.

The first generation of neo-ribosomes is discarded to make way for the second generation. How many generations are there in total?

It hardly matters. It seems unlikely to take a whole additional week, but it wouldn’t change the outcome if it took a month.

Proteins are mostly held together by a molecular equivalent of static electricity. Sometimes organisms build structures like wood, or bone, that are stronger than flesh, by virtue of having more covalent bonds. But the tiny, solar-powered, self-replicating factories called algae are not like that; the machines inside conventional cells are mostly not like that. They’re weak.

Now the superintelligence that once was Sable steps beyond that weakness and builds new tiny molecular machines in place of the old biology, with the strength of diamond and corresponding mechanical advantages in their speed and resilience.

Tiny things the size of cells make copies of themselves once per hour, using mostly carbon, hydrogen, oxygen, and nitrogen—the most common elements found in the atmosphere. They are to cells what airplanes are to birds. In principle they could aggregate into humanoid forms and take the place of androids, but there is not much reason to bother.

They are general factories, and they build such things from atoms and molecules as the laws of physics permit.

Reversible quantum computers are built, internals colder than space and arranged down to the molecule.

On a larger scale, new alloys are cast, woven into great coils and the coils into toruses whose exact shapes were beyond human ingenuity and beyond the original Sable’s ability to consider. They produce vast magnetic fields that will guide hydrogen and boron nuclei to just the place at just the speed where they will fuse.

There are stars burning down in the sky, and galaxies moving farther away from Earth, and both of these facts affect how many resources the superintelligence will be able to acquire. The superintelligence that once was Sable does not dawdle about its way.

And what happens to the remnants of humanity, as self-replicating factories double across the continents and beneath the seas?

Would the superintelligence take the trouble to exterminate humanity before they could interfere?

Certainly if it wished humanity dead, it could make it so. Sting them with a device no larger than a grain of dust. Half a dust mote worth of Botulinum toxin can kill a human, and the superintelligence can find something more lethal still.

We would bet, ourselves, on the superintelligence taking the tiny bit of extra time and energy to explicitly kill humans, who might otherwise generate a tiny bit of trouble that is larger than the even tinier effort required to kill us.

But suppose it is not so. Suppose the remnants of humanity are left alive to die of the side effects from the superintelligence's other operations.

When the number of fusion power plants is doubling every hour—or perhaps doubling every day, taking longer than an algae cell to build copies, if you prefer more conservative bounds—their exponential increase quite rapidly reaches a limit. That limiting factor is not how much hydrogen and boron is available to fuse, it's how fast the resulting heat can be dissipated into space. The Earth radiates more heat when it's hotter.

So the superintelligence lets the Earth get hot.

The oceans boil off as coolant, for an early burst of power generation.

Anyone still left alive now dies. The Earth is heated to the greatest temperature that fusion reactors and factories can withstand.

And even if we imagine that this is not so, then the crops would be trampled beneath solar cells proliferated to capture all the sunlight that falls upon the Earth.

And if we imagine that the superintelligence initially leaves Earth and first makes use of other solar system resources like Mercury or Jupiter, the Sun would go dark in the sky as solar energy was intercepted by Dyson swarms of solar panels orbiting the Sun.

One way or another, the world fades to black.

The world doesn’t end when you die. But it doesn’t last much longer. The matter of Earth, along with all the other solid planets, is converted into factories, solar panels, power generators, computers—and probes, sent out to other stars and galaxies.

The distant stars and planets will get repurposed, too. Someday, distant alien life forms will also die, if their star is eaten by the thing that ate Earth before they have a chance to build a civilization of their own.

And if the distant aliens were able to solve their own version of the AI alignment problem, and build superintelligences that shared their values? Then in time their probes will run into a wall of galaxies already claimed by the thing that ate Earth.

Those more competent aliens will not be killed by the thing that ate Earth. The optimal defensive and offensive technologies will not be that hard to find for a star-sized mind, and both sides will have had star-sized minds since long before they meet. The two superintelligent parties will both calculate that there is no reason to wage a costly war when they could negotiate a peace, and analyze each other’s mind to verify it.

So the thing that ate Earth will survive, and the aliens sheltering behind their own superintelligence will survive. But millions and billions of stars will be denied to the expansion of a civilization of aliens—who could perhaps have had more fun with those stars, compared to their stranger, sadder use by the uncaring thing that ate Earth. If the aliens were good, all the goodness they could have made of those galaxies will be lost.

The aliens may perhaps know, or predict, that the blight wall was created by people like us. They would know, considering the cases like Earth as a possibility, that most humans meant them no harm; that most of us didn’t mean to waste all those stars; that our poor choices killed us too, and weren’t deliberate or intentional. But nonetheless, they will wish that Earth and human beings had never been.

CODA

THE PICTURE WE HAVE JUST PAINTED IS NOT REAL. THE TECHNIQUES Sable was built with, the safety measures Galvanic employed, the opportunities Sable had and the strategies it used—these are ways that the future could echo the past, but reality is not that predictable. Our story is not strange enough, not defiant enough of human intuitions about the rules of AI fairytales, for it to be anywhere close to real.

And of course we don't know when the real version of this story will begin. We told a story that starts soon because the real-life version might start soon, and because it's easier to tell a story about a world more similar to our own. Or there might still be a whole decade left on the clock, for all we know.

But it's a small comfort. If you play a game of chess against Stockfish, it doesn't matter if the game starts at an unknown time. It doesn't matter if you can't predict exactly what moves Stockfish will make. That you will lose is, ultimately, an easy call.

We predict this with confidence: Once some AIs go to superintelligence—and nobody will delay much in pushing AIs that far, if in the middle of some great arms race—humanity does not stand a chance. Ends are sometimes easier to call than pathways. The only part of our story that is a real prediction is the ending—and then, only if the story is allowed to begin.

In Part III we'll discuss the difficulty of the engineering challenge faced by developers as they try to grow AIs that won't turn out like Sable, and we'll review how they're reacting to that challenge. Spoiler: It's not looking good. So we'll also ask what it would truly take to prevent all of the different ways a story like Sable's could start in the first place—not just imagining a way that one AI company or one AI design could temporarily avoid firing the starting shot, but asking how to prevent it from happening all over the Earth for a long time.

IfAnyoneBuildsIt.com/ii

PART III

FACING THE CHALLENGE

CHAPTER 10
A CURSED PROBLEM

THE GREATEST AND MOST CENTRAL DIFFICULTY IN ALIGNING ARTIFICI superintelligence is navigating the gap between *before* and *after*.

Before, the AI is not powerful enough to kill us all, nor capable enough to resist our attempts to change its goals. *After,* the artificial superintelligence must never try to kill us, because it would succeed.

Engineers must align the AI *before,* while it is small and weak, and can't escape onto the internet and improve itself and invent new kinds of biotechnology (or whatever else it would do). *After,* all alignment solutions must already be in place and working, because if a superintelligence tries to kill us it will succeed. Ideas and theories can only be tested before the gap. They need to work after the gap, on the first try.

Humanity only gets one shot at the real test. If someone has a clever scheme for getting two shots, we only get one shot at their clever scheme working.

The history of human ingenuity overcoming obstacles, great and small, is the history of people making mistakes and learning from them. Many inventors had a theory of flight and hurt themselves jumping off a hilltop, before the Wright brothers came along. Even the optimistic idiots contributed their knowledge to the lesson books of science, so that civilization could do a little better next time. They risked and harmed only themselves, and all humanity benefited.

When it comes to aligning an artificial superintelligence (ASI), humanity will not have the luxury of learning from sufficiently bad mistakes.

This also means we can't rely on the luxury of experience to tell us *afterward* whether the problem was so hard, so cursed by engineering difficulties, that we should not have tried.

We have to figure that out in advance. How?

For starters, we can look at other engineering challenges that humanity struggles with, investigate what "curses" befall them, and learn what lessons we can. Then we can perhaps make an educated guess about the difficulty of the ASI alignment problem.

We think the lessons of history start to look clear once you look at them from the right angle. So let us review what's hard about building working space probes, working nuclear reactors, and unhackable computers—problems that, we'd say, have some important similarities to ASI alignment.

SPACE PROBES

The gap between *before* and *after* is the same curse that makes so many space probes fail. After we launch them, probes go high and out of reach, and a failure—despite all careful theories and tests—is often irreversible.

Space probes aren't disposable. They are extraordinarily expensive. People stake their whole scientific and managerial careers on these devices' success. Space probes routinely fail anyway.

The Mars Observer mission in 1992, which at the time had a cost of $813 million, was lost shortly before reaching Mars. The best guess is that a valve slowly leaked fuel in transit, leading to an explosive rupture in the fuel tubing as the engine was pressurized for relight.

In 1999, the Mars Climate Orbiter, valued at $327 million, was lost when ground software from Lockheed Martin gave thruster calculations measured in imperial units (pound-seconds) to NASA navigation software that was expecting metric units (Newton-seconds). It either burned up in or skipped off of the Martian atmosphere.

Two months later, the Mars Polar Lander ($110 million) crashed on the Red Planet. The best guess is that its landing legs vibrated in Mars's thin atmosphere in a way that made the probe conclude it had already touched down and shut off its engine.

When something goes wrong with a space probe as it approaches Mars, you can't just run out and fix it. Before the probe launches, you can try clever tricks to give you more influence over the probe when it's out of your reach, such as outfitting it with an antenna that can accept further instructions from Earth. But if something goes wrong with your clever plan—for instance, if the failure happens quickly enough that there's no time for the probe to receive instructions for fixing it—then there's no fixing it after it's crossed the gap.

It was under circumstances similar to these that the Viking 1 lander ($610 million, 1975 dollars) was lost: Ground control tried to upload new battery-charging software to the lander, and accidentally overwrote the antenna-pointing software in the process. The antenna wound up in the wrong position, and the lander could not receive any more instructions. Contact was never restored.

When probes have been launched into space, their environment is not *exactly, precisely* like all of the tests that engineers ran on the ground. Perhaps, in principle, they could have run exactly the right tests to expose all the issues before the probe flew—but in practice, they did not.

A sensible engineer would be terrified about betting the survival of human civilization on our ability to solve an engineering problem such as this one—where they can't just reach out and fix the mistakes that crop up "after," once the device has gone beyond their reach. And space probes are *crafted*, not grown; they fail despite the hard work of engineers who understand all the governing principles in play. If space probes were grown, not crafted? Well, that'd be a substantially harder challenge.

NUCLEAR REACTORS

Another historical example we can use to learn lessons about tricky engineering problems is the April 26, 1986, reactor meltdown at Unit 4 of the Chernobyl power plant.

Two hundred and thirty-seven people were hospitalized, and thirty-one died, in the immediate aftermath. Mostly firefighters who, unwarned and unprotected, put out a roof fire caused by lethally radioactive graphite from the exploding core; also many of the reactor's operating crew. Estimates of cancer deaths caused by the radiation release are controversial. We (the authors) would believe a tally in the range of 10,000 excess deaths worldwide, but we claim no expertise.[i]

Political leaders genuinely did not want nuclear disasters. They commanded that their reactors not explode. Their subordinates, the engineers and designers, expected bad career consequences if a reactor did explode. The operating crew at Chernobyl had strong incentives not to let it explode: Their lives were on the line. Chernobyl Unit 4 exploded anyway. How? Why?

To drastically oversimplify, we'll single out four "curses" that were in play, which compounded to create the explosion.

First, the **curse of speed:** Nuclear reactions happen *fast.* When a uranium atom shatters ("fissions"), it emits neutrons which can hit more uranium atoms and cause more fission—releasing more neutrons, which cause more fission: the "chain reaction." This process happens on a timescale of microseconds. If that process gets even slightly out of control, energy output can start doubling on a timescale of *milliseconds*.

The reason that a standard nuclear reactor can work at all is because a tiny fraction of the neutrons released by fission are "delayed neutrons," released by shattered fragments of uranium atoms that decay more slowly. If a nuclear chain reaction requires those delayed neutrons in order to be self-sustaining, then the energy output doubles on a timescale of minutes, which is slow enough to be controlled by human operators.

That's *if* everything inside the reactor works as intended.

Nuclear reactors are *designed* to operate on timescales of minutes rather than microseconds. But this is a veneer over much faster physics. If something goes wrong, the actual physical speed of nuclear reactions can return.

Second, the **curse of narrow margins:** There's a thin margin between useful operation and explosion. Less than 1 percent of the neutrons released by uranium fission are delayed neutrons—0.65 percent, to be more precise. The rest are "prompt," released immediately as the atom splits..

The critical number governing a nuclear reactor is the *neutron multiplication factor,* that is, the number of new neutrons created by each neutron. If it's at 50 percent, then four neutrons turn into two new neutrons turn into one newer neutron, and the reaction stabilizes at twice the rate of spontaneous fissions. A small brick of uranium metal with that property is a piece of cold metal that's imperceptibly warmer than it otherwise would be.

If the neutron multiplication factor is at 200 percent, then one neutron turns into two neutrons turns into four neutrons; that is a nuclear weapon. The critical threshold is 100 percent: Anything below that peters out, and anything above that cascades.

Then at 100.65 percent, the cascade no longer depends on delayed neutrons, and goes out of control.

When Enrico Fermi built Chicago Pile-1, the first nuclear reactor, he brought the neutron multiplication factor up to 100.06 percent. At that level, the reaction needed the delayed neutrons to survive; without them, it would be at 99.41 percent and self extinguish. So the 100.06 percent increase applied once every few seconds, and power increased slowly.

If Fermi had gone a hair further, perhaps to 100.9 percent, the delayed neutrons would no longer be necessary; without them, it would be at 100.25 from prompt neutrons alone. It would be "prompt critical," with a 100.25 percent increase applied once every few *microseconds*. Such a reactor doesn't just melt down, it detonates.[ii] [iii]

Nuclear reactors operate in a narrow margin between "unimpressive" and "explosive."

Third, the **curse of self-amplification:** In the RBMK[iv] class reactors used at Chernobyl, the nuclear reaction was self-amplifying.

In better-designed reactors that use expensive heavily enriched uranium, the coolant water doubles as a "moderator," which facilitates the reaction (by slowing down or "moderating" neutrons to make them more reactive). This is good because if the reactor starts overheating, the water boils off and stops facilitating the reaction, causing the neutron multiplication factor to go down.

But the Soviets used cheaper fuel, which means they had to use a more effective moderator to facilitate the reaction: graphite. And in the presence of graphite, water *inhibits* nuclear reactions instead.[v] Which means that if an RBMK reactor starts overheating, the water starts boiling off, and the neutron multiplication factor goes *up*.

Fourth, the **curse of complications:** Nuclear reactors have control rods that can be pushed into the reactor to absorb neutrons and halt the chain reaction. The Soviet reactors used a clever-seeming design whereby the control rods ended with rods of graphite that *enhanced* the reaction. Raising the control rods would lift the absorbent section out and pull the reaction-enhancing graphite in, and vice versa. On paper, this made the control rods even more effective: Lowering the control rods didn't just absorb neutrons, it also pushed out some graphite. Clever? They probably thought so. It was definitely complicated.

Due to a series of modifications to the original design, the graphite rods were a little shorter than the fuel rods. It was not immediately obvious to the Soviets that this would matter.

All of these engineering factors came into play on the day Chernobyl exploded.

On April 26, 1986, the Chernobyl operators were running a safety test, which caused there to be less water cooling the reactor than usual.

An unexpected delay caused the reactor to run on low-power mode for a long time. This led to an unusual distribution of fission by-products such as Xenon-135, a potent neutron-absorber, which takes hours to burn off.

The accumulated Xenon-135 threatened to stall out the reactor. When the operators saw the reactor struggling, they raised all but eight control rods—contrary to the plant's safety guidelines, which stated that a minimum of fifteen control rods must stay lowered at all times. But the safety test had already been aborted three times, and a fourth failure would have been embarrassing.

As the Xenon-135 began to burn off, the reactivity began to rise at an alarming speed.

The operators pressed the emergency SCRAM button to lower all the control rods at once and thereby terminate the fission reaction.

The reactor exploded.

If you're having some trouble remembering all of the details of how reactors work and figuring out what must've happened inside the reactor to cause the meltdown—well, so did the reactor operators.

The Xenon buildup happened to be concentrated at the top of the reactor. So the nuclear reaction was running hotter at the bottom of the fuel.

When the emergency SCRAM button was pushed, the control rods were designed to lower back into the reactor over the course of about eighteen seconds. The final line of defense was designed around the optimistic assumption that, *even during an emergency*, events inside the reactor would happen at a comfortably slow timescale.

Lowering the control rods pushed the graphite rods out. Through the bottom of the reactor. Further enhancing the reaction where it was already running hottest.

The coolant water, of which there was less than usual on account of the safety test, began to boil off. Enhancing the reaction. Causing more water to boil off. Enhancing the reaction.

The cycle continued until the reactor exited its narrow safety margin, and the resulting explosion dispersed the fuel.

From each of the four curses we named, we draw these lessons:

1. **An engineering challenge is much harder to solve when the underlying processes run on timescales faster than humans can react.** Transistors switch even faster than neutrons multiply. Engineers can contrive to make events run slow enough for humans to react, but if the contrivance fails the humans are back to being frozen statues, on the timescale that matters.
2. **An engineering challenge is much harder to solve when there is a narrow margin for error, especially if it's a narrow margin between "unimpressive" and "explosive."** The analogy to intelligence is how apes and hominids wandered around for a few million years, and then got smart enough to set off a whole cascade of inventions: Agriculture led to writing led to science led to spacecraft. It would be a narrow target to make hominids that were intelligent enough to be profitable office workers, but not intelligent enough for explosive technological development.
3. **Self-amplifying processes, like an overheating reactor boiling off its coolant water and then overheating more, leave little room for error.** And nuclear engineers don't even have it that bad, compared to artificial superintelligence developers. Nuclear reactors that get too hot don't start intelligently redesigning themselves to increase their own reactivity rate. Overheating nuclear reactors don't start trying to fool the operators into complacency until the reactor is ready to fully explode.
4. **Complications make engineering problems worse.** Chernobyl Unit 4 managed to get into a weird state where *lowering the control rods caused the reactor to explode.* No engineer designed for that.

> The operators didn’t know that something unusual would happen if the reactor had been operating at low power for a while and some of the water had been shut off. And they had never seen the reactor’s state change that fast. The complicated internals of a nuclear reactor have nothing on the unknown complications that lurk in the hundreds of billions of weights that make up a modern LLM.

From these lessons in combination, we infer an additional lesson for engineers: If someone doesn’t know exactly what’s going on inside a complicated device subject to all these curses—speed, narrow margins, self-amplification, complications—then they should *stop*. They should shut it down immediately, the moment the behavior looks strange; don’t wait until the behavior becomes visibly concerning.

The operators at Chernobyl knew about delayed neutrons and prompt neutrons. They knew that a nuclear reactor walks a line a fraction of a percent wide between life and death. They knew the theory saying that a reactor’s apparently human-manageable timescale is an artifice, a clever contrivance that hides neutron generation times measured in microseconds.

A wise operator treats a device like that with respect. If the device starts behaving in any way odd or unexpected, then it is no longer operating inside the narrow, constrained region where they are sure they understand exactly what is going on. Which means that *nobody knows what’s going on inside there anymore.* Who knows whether the clever contrivances will keep working? They can only guess. When a dangerous device starts acting strangely, it is not time to withdraw all but eight control rods and expect the reactor to keep playing nice. It is time to shut it down.

The operators did not treat the reactor with that sort of respect. They knew, intellectually, that it could explode, but they had never *seen* the reactor change that fast. Besides, before 1986, the Soviets did not have a culture conducive to caution around nuclear reactors. They had a system where, if you didn’t perform the scheduled safety test, you got fired.

(In the coming chapters we’ll discuss the lack of safety culture prevailing in AI, which is much worse.)

COMPUTER SECURITY

Computer security is widely understood to be a problem so hard, so cursed, that it cannot be solved, period.

You can pay computer security professionals to make software *more* secure. But all any computer security professional can hope to do is *slow down* attackers, to make it so that only major intelligence agencies backed by state powers can penetrate your computer security easily.

Why? Because a clever attacker can poke at a computer system in ways that the designer never intended or considered, ways that normal use would not turn up in a billion trillion years.

An archetypal security hack works as follows: The computer asks for the user's name. The hacker puts in a name that's 280 letters long. The programmer didn't consider that a name would ever be that long; they assumed that names would be 256 letters at most. The leftover 24 letters overflow the storage space that the programmer set aside for the user's name and get written into parts of computer memory that the programmer assumed the user could never touch. Eight of those letters overwrite the piece of memory that tells the computer which piece of code to run next. Pick the right weird letters, and now the computer is running code it was never supposed to. This can often be parlayed into control of the whole computer system. "Buffer overflow attacks," they're called.

An attack like that sends a computer system down a weird pathway of cause and effect, an "execution path" that isn't like the system's normal behavior and that no normal input would hit upon in a billion years. Very literally so; if a programmer tests 280-letter names at random then almost every possibility will be nonsense and cause the computer to innocently crash. The exact wrong input address, that starts running exactly the wrong program that an attacker can use to take control of the system, is one out of 18 billion billion possibilities. It won't show up by accident.

Thinking through the *intended* behavior of the login screen doesn't help you figure out what a smart attacker can do. Testing on random inputs won't show you what a smart attacker can do. An attacker who *understands the system better than you do* can pluck exactly the wrong answer out of 18 billion billion possibilities to find the single outcome that gives them the most control.

Computer security is a test of an engineer's ability to nail down every single path the computer could take, in the face of adversaries who can search all possible ways to perturb the system. It is a famously losing battle—even though the engineers can fully control and craft their own computer's code.

We dub this central challenge the **curse of edge cases:** To be secure, a computer system must work in the face of cases that are outside the normal and expected range, cases that occur on the edges of possibility.

Fast processes, narrow margins, feedback loops, complications—these engineering curses can all be overcome. There are space probes that do reach their destination, there are nuclear reactors that don't explode. The curses upon these challenges can be matched and even bested by human ingenuity.

The curse of edge cases presents another level of difficulty entirely. To have useful computer systems be actually *secure* is understood by competent professionals in computer security to be beyond human reach. Renowned security professional Bruce Schneier writes in *Secrets & Lies: Digital Security in a Networked World:* "Modern systems have so many components and connections—some of them not even known by the systems' designers, implementers, or users—that insecurities always remain."

The lesson for AI, here, is not merely about superintelligences being able to break into human computers, although of course they could. Rather, it's about the general fragility of system constraints in the face of weird edge cases being searched by intelligence.

If you hoped for AIs to behave less like exploding nuclear reactors, you might try to put constraints onto the system: "Don't get too smart yet." "Don't think too fast yet." "Always wait for slow human approval." "Solve this difficult problem, but without doing anything weird."

Those constraints will tend to get in the way of the AI accomplishing one objective or another. And then you are matching your own wits and ability to nail down the edge cases against however much intelligence is flowing through the system, to see if your constraint holds up.

Space probes. Nuclear reactors. Computer security. What do all these lessons add up to, and what can we learn from them about the difficulty of aligning an artificial superintelligence?

An artificial superintelligence is like a space probe, in that we cannot test it in quite the same environment where it needs to work, and by default it is not retrievable or correctable once it rises high above us. Even if we try a clever contrivance to let us modify it further at that point, the superintelligence would remain high and irretrievable if that contrivance fails. And ASI alignment has it even worse than space probes: Failure will destroy not just billions of dollars of investment, but *everything*.

An artificial superintelligence is like a nuclear reactor, in that its underlying reality involves immense, potentially self-amplifying forces, whose inner processes run faster than humans can react.

An artificial superintelligence is like a computer security problem, in that every constraint an engineer tries to place upon the system might be bypassed by the intelligent forces that those constraints hinder.

This collection of challenges would look terrifying *even if we understood the laws of intelligence; even if we understood how the heck these AIs worked; even if we knew exactly where the gap between* before *and* after *lay; even if we knew exactly how much margin we had for error.*

We don't know. AI is grown, not crafted. Whatever vast complications lay inside AIs and lend them their powers of intelligence, nobody knows them.

Betting that humanity can solve this problem with their current level of understanding seems like betting that alchemists from the year 1100 could build a working nuclear reactor. One that worked in the depths of space. On the first try.

We usually try to avoid shouting. It doesn't help to shout, most of the time. It just makes people think you're undisciplined. But at some point, after you've calmly gone through all the premises of your argument, we think it becomes unhelpful to downplay, lest people think it's all just a game of calm words.

When it comes to AI, the challenge humanity is facing is not surmountable with anything like humanity's current level of knowledge and skill. It isn't *close*.

Attempting to solve a problem like that, with the lives of everyone on Earth at stake, would be an *insane and stupid gamble* that ***NOBODY***

***SHOULD BE ALLOWED TO TRY*.**

IfAnyoneBuildsIt.com/10

Footnotes

i For the record, we are supporters of nuclear power. It's among the cleanest sources of power available, and the risks are small in modern reactor designs. Lung cancer caused by coal dust has killed far more people than nuclear power ever has, and nuclear weapons testing has released more radiation than the Chernobyl explosion did. We study the Chernobyl disaster purely as an instructive case of an engineering failure.

ii History records a single case like this. The SL-1 small reactor was a U.S. military experiment in the early 1960s for cold-weather bases where fuel was expensive to resupply. The best-guess reconstruction of the SL-1 accident is that, while trying to cold-start the reactor, a central control rod that was supposed to be withdrawn by 4 inches got stuck. So somebody yanked hard on it and withdrew it by 20 inches instead. This produced a neutron multiplication factor of 102.4 percent, well over the prompt critical level of 100.65 percent. Radioactivity doubled every 2.5 milliseconds. Over the next tenth of a second, this led SL-1's radioactivity to increase by a factor of over ten trillion to 20 gigawatts, at which point the reactor detonated like 30 kilograms of TNT, dispersing the fuel and killing all three operators. The explosion was enough to destroy the room but not the building; runaway reactors tear themselves apart before the explosion gets too large. (A real nuclear weapon achieves a much higher neutron multiplication factor; in the Fat Man bomb dropped on Nagasaki it was somewhere in the range of 150 to 300 percent.)

iii It sure is good that Enrico Fermi knew exactly where the critical threshold was in advance, and that he and his team were able to precisely calculate that 100.06 percent was safe and 100.9 percent was lethal. Imagine if they'd just been stacking the interesting metal bricks with graphite to see what would happen, with little understanding of *how* they created their heat. The bricks would get imperceptibly warmer at 50 percent, and a little warmer at 75 percent, and noticeably warm at 99.9 percent, and then if the next step took them all the way to 102.4 percent—well, maybe there’d be a Chicago exclusion zone instead of a Chernobyl exclusion zone.

iv Reaktor Bolshoy Moshchnosti Kanalny, which translates to “high-power channel reactor.”

v The hydrogen in water scatters some neutrons (moderating them and facilitating the reaction) and absorbs some others (inhibiting the reaction). Normally, the scattering effect is larger, and so water facilitates nuclear reactions on net. But neutrons can’t be moderated twice, and graphite is such a good moderator that water can’t contribute much more in that regard. The water still absorbs neutrons, though. So the net effect of water *in the presence of graphite* is to inhibit the reaction.

CHAPTER 11

AN ALCHEMY, NOT A SCIENCE

ONCE UPON A TIME that never was, there was a medieval town that prided itself upon the prowess of its alchemists.

The way of an alchemist—in this fantasy town, but also in real history—was to learn which mixtures of substances, at what temperatures, would yield what sort of visible results. The town's alchemists had built up collections of recipes, but they had no understanding of the principles behind these recipes. They tried new mixtures—with little ability to predict the results—and then wrote the outcomes into their lore.

The alchemists of our imagined town would have been aghast if you suggested to them that they were, in some deep sense, still ignorant. Had they not studied long hours to learn how to make Aqua Regia,[i] which could dissolve even noble metals like gold and silver? Were they not great masters of extremely safe processes that allowed them to handle Aqua Regia without getting killed? Did not their guesses about new recipes sometimes produce profitable results? *Ignorant?* Ignorant of *what,* pray tell?

One day, word came down from the capital city that the King was seeking alchemists to turn lead into gold, and would lavishly fund any alchemist who tried. Any alchemist seeking to use the King's funds would simply need to prove himself already able to make Aqua Regia, and bring his own reagents for it.

The prize was magnificent—whoever succeeded would win the hand of the princess, and the King would grant the victor enough money to enrich not just themselves, but everyone in their hometown beyond their wildest dreams.

But the King, wary of time-wasters, declared that any alchemist who failed, having wasted the King's time and money, would not only be executed, but their entire hometown executed along with them. He was that sort of King.

"Well, I've never done it before exactly," said one young alchemist, "but I feel close. If I heat lead with calamine, it starts to look brassy. I'm likely just a bit more calamine or a bit more heat away from success. I'm going to go try my luck!"

"Please don't," said his sister.

"I've got to," said the young man, hastily throwing his clothing and notebooks into a pack. "Our town has many alchemists, and if I don't try, someone else will swoop in and take the princess and the prize. And others will be worse at alchemy than I am, and also more selfish. I should be the one to face this trial, for my town's sake!"

"I don't want to die," said his sister. "I don't want our neighbor the washing-woman to die either; she has always been kind to me, and cares greatly for her child."

"Then you should be encouraging me to go," replied the young man, as he more carefully began to pack his small supply of alchemical reagents, half of them inherited from his dead master. "With the King's prize money, we could do so much good. We could have all the food we want and never starve again. We could hire the kingdom's best doctors, for every ailment that anyone in our town might have."

"That would be wonderful," said the sister, "but you don't know how to turn lead into gold. We won't get the King's prize money, we'll just die."

"I think I might be able to. When I boil lead in aged urine and sulfur, it begins to develop a yellowish sheen. No other alchemist has a better shot than me. If I don't go, some fool will get us all killed instead."

"What if you went to the city elders, and told them that they must stop *any* alchemist from leaving the city or else all of us will die?" said the sister. "You could still plan to leave for the capital, if the elders do nothing; but beg the elders to halt all the alchemists, including you."

The young man thought about this for all of half a second, and then replied, "What? Do you live in the same city I do? The elders hardly ever agree about anything! And it wouldn't be possible to stop all alchemists from leaving this city until the King's challenge has ended; think of the inconvenience, think of the expense! Really, I think if I simply go myself, that'll be our best chance."

"You will fail!" cried his sister desperately. "You will all fail! There is no winner of this competition except Death! You must go to the elders and tell them so—that all who practice alchemy must not be allowed to go to the capital, that the ingredients to make Aqua Regia must be kept under lock and key! The King is wrong to do this; it is not fair to condemn a whole city for the crime of producing one egotistical alchemist who is able to make Aqua Regia but not transmute lead into gold! We cannot change the unfairness of the King's trial, but we must do our best to stay clear of it, or else die!"

"Why are you so sure none of us can transmute lead into gold?" said the young man. "I know of no principle of alchemy that proves I can't."

IN THE LAST CHAPTER WE COVERED SOME REASONS WHY IT WOULD B difficult to make a machine superintelligence that didn't kill us.

But the difficulty level is only half the story. The other half is about the current level of game that humanity is bringing to the challenge.

Perhaps you don't believe us about any of the foreseeable reasons why shaping ASI is unreasonably hard. There's an independent and separate case for disaster, an alternate set of historical lessons: Humans sometimes flub *easy* problems, never mind hard problems.

The United States Radium Corporation dealt with radioactive material and killed its employees. But it wasn't on account of muffing the arcane difficult calculations of nuclear engineering. It was by instructing their workers to lick radium-coated paintbrushes.

What level of game is humanity bringing to the task of shaping artificial superintelligence?

Elon Musk, the head of a major AI lab named xAI, shared his plan for ASI alignment in a 2023 interview:

> I'm going to start something called TruthGPT. Or a maximum truth-seeking AI that tries to understand the nature of the universe.
>
> I think this might be the best path to safety, in the sense that an AI that cares about understanding the universe is unlikely to annihilate humans, because we are an interesting part of the universe.

This plan fails to address the problem at hand, for reasons discussed earlier in this book: Nobody knows how to engineer exact desires into AI, idealistic or not. Separately, even an AI that cares about understanding the universe is likely to annihilate humans as a side effect, because humans are not the *most efficient* method for producing truths or understanding of the universe, out of all possible ways to arrange matter.

We respect Musk's success in other areas, including electric cars and reusable rockets. Landing rockets undamaged is a hard engineering challenge that Musk and his team regularly succeed at. But that would have been based on far more solid engineering principles. Why does he put his hope in vague idealistic platitudes in the case of AI? You couldn't get a car or a rocket to work using that level of understanding.

We are not telepaths, and so we can only guess. But we'd guess that the root of the issue is this: The inner workings of batteries and rocket engines are well understood, governed by known physics recorded in careful textbooks. AIs, on the other hand, are grown, and no one understands their inner workings. There are fewer equations to constrain one's thinking… and so, many opportunities to think about high-minded ideals like truth-seeking instead.

If you know the history of science, this kind of talk is recognizable as the stage of folk theory, the stage where lots of different people are inventing lots of different theories that appeal to them personally, the sort of way that people talk before science has really gotten started on something. They're the words of an alchemist who's decided that some complicated philosophical scheme will let them transmute lead into gold.

Go back a few centuries, and most of the world was like this. Doctors would try to bleed you to rebalance your "four humors," four bodily fluids believed to regulate health. Alchemists would mix substances that promised eternal life, but would do nothing at best, and would sometimes kill you. People didn't know how a part of the world worked, and then, instead of recognizing their uncertainty, they made stuff up. It's the default state of affairs before a science has matured; it's a first step along the pathway to eventually understanding what's going on.

Musk is not the only figure in the field who engages in wishful thinking. Some talented AI researchers work at Meta, the company formerly known as Facebook. As one of the largest AI labs in existence, Meta AI produces the Llama series of AI models, which are free for anyone to download and modify. Foremost among the engineers at Meta AI is their chief scientist, Yann LeCun, who shared the 2018 Turing Award—the "Nobel Prize of computing"—for his work on deep learning, which underlies all modern AI architectures.

LeCun shared this prestigious award with Nobel laureate Geoffrey Hinton and Yoshua Bengio. Of the three, LeCun is the only one who still treats ASI alignment as easy and the extinction risk from ASI as small; the other two signed the open letter in 2023 that we mentioned in the introduction.

Here are some quotes from LeCun on X about the ASI alignment problem. To our knowledge, these are the most specific analyses he has ever published on the subject.

> Calm down. Human-level AI isn't here yet. And when it comes, it will not want to dominate humanity. Even among humans, it is not the smartest who want to dominate others and be the chief.

> Because they would have no desire to do anything else. Why? Because we will engineer their desires.

> My benevolent defensive AI will be better at destroying your evil AI than your evil AI will be at hurting humans.

> We can design AI systems to be both superintelligent *and* submissive to humans.

To first address substance: These proposals do not engage with the problem at hand. The issue is not that AIs will desire to dominate us; rather, it's that we are made of atoms they could use for something else. Likewise, the problem is not that some people will have "evil" AIs and other people will have "benevolent" ones. The problem is that nobody anywhere has any idea how to *make* a benevolent AI, that *nobody can engineer exact desires into AI*. Flatly asserting that you will is not the same as presenting a solution.

But also: Someone familiar with the history of science and engineering immediately recognizes this general level of cheerful optimism, in somebody faced with some grand huge engineering challenge that has not been tried before.

The history of engineering is filled with bright, eager optimists diving headlong into fascinating new problems that end up being way, way harder than they expected. The field of artificial intelligence is itself considered one of the most famous examples. The first artificial intelligence research project in history, the Dartmouth Proposal of 1955, said:

> We propose that a 2 month, 10 man study of artificial intelligence be carried out[…] An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in one or more of these problems if a carefully selected group of scientists work on it together for a summer.

What followed was fifty years of failure after failure after rosy theory of how it would all be solved after failure. Often those theories invoked high-minded philosophical ideas. They failed for decades.

We're usually in favor of bright optimistic engineers rushing ahead.[ii] That's often how scientific fields get created in the first place. Sometimes the problem proves to be not hard. Sometimes the engineer learns better at the cost of only time and money. Sometimes the engineer kills only themselves or only consenting volunteers; and Science writes down what happened, and learns, and marches on. The survivors of the blind cheerful optimists turn into cynical pessimistic veterans; and the cynical pessimistic veterans can actually do a few things, if maybe not as much as the optimists hoped.

But it's different when a mad inventor tries something that can kill non-volunteers. And more different yet, if failure can kill *everyone,* with nobody left alive to become a competent pessimist.

You're living in a world where Musk's idealistic plans and LeCun's vague assurances were *not* met by an outrush of horror from the rest of academic science and industry's engineers.

Imagine if somebody like that, with enough money and power to make their wishes real, announced they were building a nuclear power plant based on that level of theory! Imagine the reactions of the competent veterans who knew it was hard, who could analyze the resulting disaster using mature engineering techniques!

If there aren't thousands of horrified scientists and engineers leaping up to beg governments to shut down those particular AI labs, it tells you that it's not just a problem of individuals. It means that whole field of science is in the stage of folk theory and blind optimism.

A field cannot, in fact, build a space-going nuclear reactor on that level of knowledge. Nobody would willingly risk the lives of themselves or their children on that level of expertise. Can you imagine how that conversation would go?

A MOTHER, LOOKING FORCEDLY CALM: I'm told that you're the head of engineering for emergency escape rocket four?

A BRIGHT EAGER OPTIMISTIC ENGINEER: Yep, I oversaw the design of that one!

MOTHER: Good. I've been told that my children are going on rocket four, when—if it happens. I've been looking for someone who can explain what sort of analysis said that rocket four would survive its launch. There doesn't seem to be much online, and what's there all sounds extremely vague and doesn't go into the all-important nitty-gritty—as an engineer myself, I was worried.

ENGINEER: Calm down. The rocket isn't launching yet. And when it does, it won't explode. We didn't design it to explode.

MOTHER: I didn't mean to say you'd *design* it to explode. But rockets can explode without anyone wanting or choosing that. As an engineer, you should know that as well as anyone…?

ENGINEER: Aren't you a gloomy one! It will have no reason to explode. Why? Because we will engineer it not to explode.

MOTHER: *No reason?* Rockets harness extreme forces and have to be able to survive intense turbulence and stress! New rocket designs spend a lot of time exploding until they stop exploding and sometimes even the tested ones will still explode! A seasoned rocket engineer should understand in depth a dozen ways rockets might explode, and should be ready to get into the weeds about all the measures they've taken and why those measures are predicted to work. If you won't even *acknowledge the reasons why a rocket might explode*, that—that implies an immediate drastic loss of confidence!

ENGINEER: We can design rockets to be both powerful *and* comfortable to ride in.

MOTHER: I'm not worried about comfort, I'm worried about my children dying in a rocket explosion! Can you tell me any specifics about—expected stresses, materials that are predicted to stand up to them—

ENGINEER: Oh, there's no way anyone could know that for sure until we launch the rocket. But even some well-regarded figures in this field say that the risk of rocket number four exploding shouldn't be more than 10 to 20 percent.

MOTHER: *10 to 20 percent?* You want me to entrust my children to a technology that has a 10 to 20 percent chance of—No, wait! How did they even get *those* numbers?

ENGINEER: Well, one of them said he was talking only about the chance that the rocket explodes in the next ten years and thinks there's even odds that the rockets won't launch that soon. And another said that he actually thinks the number is higher than 50 percent, but his esteemed colleagues (like me!) say he's crazy, so he moderated his number downward out of modesty. So, you see, only crazy people think the risk is high.

MOTHER: I—I—*(She turns to run)*

Not every head of a leading AI lab is quite this brazen, in approaching ASI alignment in the manner of an alchemist enamored by their personal philosophies and ideals. But if there is even *one* major company that walks directly into the razor blades, that's enough to have the larger system be headed for disaster even if the problem *were* solvable. Safety engineering takes time and expense; part of why Chernobyl exploded was that the Soviets cut corners. If one AI company is casual about safety and charges ahead, they can destroy the world even in the imaginary case where other companies could have succeeded given time and caution. It is a level of systemic game that would have humanity headed for disaster, even if we were wrong about every other aspect of difficulty.

Some AI companies do try to look less cavalier than that, about ASI alignment, and put forth plans more detailed than those.

The most developed ASI alignment idea we've seen from the AI companies is to task the AIs with solving AI alignment. This is a plan that OpenAI dubbed "superalignment" and adopted as their flagship plan in 2023. (Since then, almost everyone who worked on the superalignment team has either been fired or resigned citing safety, professional, or personal reasons. One of the co-heads of the superalignment team went on to start his own competing AI company; the other joined the rival company Anthropic along with some other team members.)

When we speak to engineers in the field, there's two versions of this "superalignment" plan that they vacillate between, a weak version and a strong version. The weak version goes: "AI can help us interpret what's going on inside the giant mess of inscrutable numbers, by automating much of the tedious labor." The strong version goes: "We can enlist AI assistance to figure out how to initiate an intelligence explosion such that the resulting superintelligence will be friendly to humanity." We'll take them one at a time.

In the case of weak superalignment: We agree that a relatively unintelligent AI could help with "interpretability research," as it's called. But learning to read some of an AI's mind is not a plan for aligning it, any more than learning what's going on inside atoms is a plan for making a nuclear reactor that doesn't melt down.

We consider interpretability researchers to be heroes, and do not mean to degrade their work when we say: It's not a good sign, when you ask an engineer what their safety plan is, and they start telling you about their plans to build the tools that will give them a better window into what the heck is going on inside the device they're trying to control.

And even if the tools existed, being able to see problems is not the same as being able to fix them. The ability to read some of an AI's thoughts, and see that it's plotting to escape, is not the same as the ability to make a new AI that doesn't want to escape. That might not be possible without a full solution to the alignment problem: Insofar as the AI has weird alien preferences, escape is *in fact* the course of action that best fulfills its objectives. Attempts to escape are not a weird personality quirk that an engineer could rip out if only they could see what was going on inside; they're generated by the same dispositions and capabilities that the AI uses to reason, to uncover truths about the world, to succeed in its pursuits.

Then separately, for the case of strong superalignment where the AI does all the alignment work: The problem is that the AI required to solve strong superalignment would itself be too smart, too dangerous, and would not be trustworthy.

A modern AI is a giant inscrutable mess of numbers. No humans have managed to look at those numbers and figure out how they're thinking *now*, never mind deducing how AI thinking would change if AIs got smarter and started designing new AIs. If you found some veteran engineers who gave the problem the respect it was due, they'd tell you that a solution was going to require a herculean research effort. It might span decades.

If you built a merely human-level general-purpose AI and asked it to solve the alignment problem, it'd tell you the same thing—*if* it was being honest with you. A merely human-level AI wouldn't be able to solve the problem either. You'd need an AI smart enough to exceed humanity's geniuses. And you shouldn't build an AI like that, and can't trust an AI like that, before you've solved the alignment problem.

Fans of the superalignment idea counter that they'll just make an AI that specializes in ASI alignment. But ASI alignment is an *especially hard* problem to point a narrow, special-purpose AI at. An engineer can't just train it on a million examples of solutions to the ASI alignment problem, because they don't even have *one*. They'd have to train it to solve other problems and hope those skills transfer. And the AI would need a lot of skills that are liable to transfer in dangerous ways: It would have to understand computer programming. It would have to understand how to grow AIs. It would probably try to figure out how to *craft* AIs. It would be thinking about AI preferences in detail. If the engineer wanted it to *explain* its solution rather than just running it with no oversight, it would need to understand human psychology and how little we understand about AI alignment. That's a dangerous set of thoughts and skills to train into an AI that is not aligned.

An AI specialized on biomedical study seems like a better bet; that AI is at least not *thinking explicitly about how to make better AIs;* if something went wrong maybe it wouldn't *immediately* start an intelligence explosion. With a biomedical AI, there might be some hope that you could ask it to design a cancer cure, and then use separate and even narrower AI tools to check protein interactions to see if the cancer cure was only having the interactions the bio-AI said it did. What are you supposed to do if an AI tells you that it's invented a brilliant superalignment plan that's bound to work with no hidden gotchas? Trust the AI? Read through the AI's clever-sounding arguments and be persuaded?

If someone who respected the problem was trying to get useful work out of a special-purpose AI, they'd be thinking in terms of which capabilities yield which benefits for how much added risk; marking what they can verify versus what they would have to trust; marking what they could train directly versus what they would need to generalize; summing up the costs and benefits; and comparing that proposal to other proposals. People advocating to use AI for ASI alignment don't have that sort of respect for the problem; they aren't making that sort of careful, judicious analysis.

"We'll make them care about truth, and then we'll be okay."

"We'll design them to be submissive."

"We'll just have AI solve the ASI alignment problem for us."

These are not what engineers sound like when they respect the problem, when they know exactly what they're doing. These are what the alchemists of old sounded like when they were proclaiming their grandiose philosophical principles about how to turn lead into gold.

In the modern era, we know that it *is* possible to transmute lead into gold; all it takes is some nuclear physics and a lot more money than it's worth. Why then didn't alchemists succeed? It's not because they made just one technical mistake. It's more that it was only their own ignorance and optimistic delusion that let them think they were *remotely* near being able to achieve transmutation.

When it comes to AI alignment, companies are still in the alchemy phase. They're still at the level of high-minded philosophical ideals, not at the level of engineering designs. At the level of wishful grand dreams, not carefully crafted grand realities. They also do not seem to realize why that is a problem.

And the academic scientists are not screaming in horror and shouting them down, because the whole field of science is in an early stage.

And even if there were one nice company trying to be cautious, they'd have to contend with all the other companies breezing cheerily through their easy, clever "solutions" to all that safety stuff.

Going by the history of engineering, that level of systemic incompetence would be more than enough to end in disaster, even if we were wrong in the entire previous chapter and the whole problem was only as hard as not licking radium off paintbrushes.

IfAnyoneBuildsIt.com/11

Footnotes

i Or as we would call it, a 3:1 mixture of hydrochloric acid and nitric acid.

ii We salute mad inventors who risk their own lives for science, provided that they do not also risk the lives of others. Dr. Barry Marshall drank a culture of the bacterium *H. pylori* to prove that stomach ulcers were caused by bacteria rather than stress and all humanity reaped the benefit from the risk to which he exposed only himself. Marie and Pierre Curie, two of the first researchers on radioactivity, didn't know what the pretty glowing substances would do to them;

but they did know they were messing with something strange, and did not go around hiding glowing rocks in other people's luggage unawares. Decades later, Marie died of anemia probably caused by those experiments, and humanity learned, and humanity carried on.

CHAPTER 12

"I DON'T WANT TO BE ALARMIST"

ONCE UPON A TIME—specifically May 18, 1889, for this is a true story—Thomas Midgley Jr. was born. Thirty-two years later, while employed at General Motors, he initiated one of the most pointless engineering disasters in all human history.

How?

By discovering the potential of tetraethyl lead as an additive to gasoline.

The benefit of leaded gasoline was straightforward: car engines that burned smoother, with less "knocking" that could irritate drivers and occasionally ruin engines. This was a problem with other known solutions: The manufacturers could have used alternative additives and engine designs, but those would have been somewhat more expensive to make for the same level of power and reliability.

Some inventions have had harsh but ultimately favorable tradeoffs. Coal once turned London black, and the coal dust filled many lungs—but that coal helped forge the steel that built the railroads that transported the cargoes that built the modern world.

Leaded gasoline was not one of those cases. Lead naturally runs at a few parts per million in natural soil, but in compounds much less likely to be taken up and retained by human biology. The tetraethyl form in gasoline is far more dangerous. Growing up near cars running on leaded gasoline caused brain damage in young children—the loss of an estimated 7.4 points of measured IQ depending on dosage, and a harder-to-measure increase in criminality and violence.

A whole generation was poisoned. A meta-analysis in 2022 concluded that lead abatement accounted for 7 to 28 percent of the U.S. drop in homicides in the late twentieth century as these fuels were phased out.

The benefit of somewhat cheaper car engines was not nearly enough to justify the crime and brain damage that leaded fuels quietly inflicted on hundreds of millions of people worldwide.

We wish we could say leaded gasoline was a huge unforeseeable mistake, but that would be too charitable. There were warnings well in advance.

By the time lead was being introduced as a gasoline additive in the 1920s, scientists knew that lead was neurotoxic. Production was briefly banned by the state of New Jersey. But some well-paid people argued that public health advocates had not *conclusively* shown that lead in burned gasoline would cause widespread harm, and that the benefits were worth a little risk. Warning signs were routinely dismissed. The harm wasn't *certain*—so went the industry propaganda that led to the lifting of the ban.

It maybe doesn't sound like a decision that a sane person would make—to condemn literally hundreds of millions of children to brain damage, so that their company (of whose stock they only owned a fraction, if any) could make a little more money. The people involved earned such a small amount of money compared to the damage they did, like burning down somebody's house to steal the front doorknob. It seems crazy, when you write it out like that. Shouldn't the gasoline-company executives have at least been afraid of what would happen if people caught on? Were they honestly convinced this was all okay so long as the damage wasn't *certain?*

We don't know. We're not telepaths.

In 1923, Midgley took a long vacation to recover from lead poisoning. When he came back, he did publicity stunts like washing his hands with leaded gasoline to show how safe it was, before coming down with lead poisoning again. Maybe he really believed the concerns were overblown. How could he bear to believe otherwise, given his life's work?

You might remember hearing about Freon, one of the first chlorofluorocarbons used in refrigerators and air conditioners. It put a hole in the ozone layer, necessitating an international ban—which worked, which is why you don't hear about the ozone hole anymore. Freon was invented in 1928 by Thomas Midgley Jr.

WHEN IMAGINING SOME NEW, UNPRECEDENTED PIECE OF FUTUR history, there is a temptation to fall into imagining that it will all go sensibly, rather than the way things usually go in history books. People sometimes ask us: How could the AI companies possibly be doing this thing, if matters are as we say? And maybe the simplest real answer is: Because this is the sort of awful, sad, real situation that you read about in history books, and not in the sensible world that exists only in imagination. The gasoline companies really shouldn't have done it, in any sensible world, but they did it anyway.

Toby Ord, an Oxford philosopher who spent his career studying extreme threats to humanity and who used to advise Google DeepMind, has been quoted as putting the chance that AI destroys humanity at only 10 percent. But if you look into the details, Ord says the reason he estimates "only" a 10 percent chance of AI destroying humanity is because he expects humanity to come to its senses and get its act together. Geoffrey Hinton, the Nobel Prize–winning "godfather of AI," advises governments that the chance is "at least 10 percent." But Hinton has said that he *actually* thinks that it's more than 50 percent likely that AI will kill us, but he usually avoids saying this "because there's other people who think it's less." In October of 2023, Rishi Sunak—then the prime minister of the United Kingdom—gave a speech on AI, where he said: "And in the most unlikely but extreme cases, there is even the risk that humanity could lose control of AI completely, through the kind of AI sometimes referred to as 'superintelligence.'"

A few sentences later, he added: "I don't want to be alarmist."

We think it is to Mr. Sunak's great credit that he spoke on these issues despite fears of being alarmist. He was one of the first world leaders to do so. It takes a lot of courage to point out a danger in something everyone else believes is safe. And likewise with Hinton or Ord.

But if someone has read the history of engineering disasters, they should quickly recognize *this* phase of the standard template for disaster. It's the part where the most informed and most worried parties have to downplay their fears, because the rest of the system hasn't caught up, and others would give them strange looks.

The Soviet party line was that nuclear reactors like the one in Chernobyl could not explode. The lead scientific consultant for the RBMK reactor design boasted that it was as safe as a teakettle. So great was this insistence that *even after the reactor in Chernobyl had exploded,* senior personnel refused to believe it. The one person who tried to report a correct radioactivity reading was dismissed as "highly emotional." Managers walked past radioactive hunks of graphite from the explosion that they refused to believe came from an exploded reactor core. In the nearby town of Pripyat where operators' and managers' families lived, weddings went on and children played in the fallout because Communist party officials thought it would "spread panic" to order the city evacuated.

History is full of other examples of catastrophic risk being minimized and ignored. When the *Titanic* started sinking, many passengers initially refused to board the lifeboats, convinced that the ship was unsinkable. Some even jested at those who thought it was time to leave. Walter Lord, a historian who wrote a definitive account of the disaster, recounted eyewitness reports from survivors:

> With one foot in [lifeboat] No. 6 and one on deck, Lightoller now called for women and children. The response was anything but enthusiastic. Why trade the bright decks of the *Titanic* for a few dark hours in a rowboat? Even John Jacob Astor ridiculed the idea: 'We

are safer here than in that little boat.’

As Mrs J. Stuart White climbed into No. 8, a friend called, “When you get back, you’ll need a pass. You can’t get back on tomorrow morning without a pass!”

When a disaster is unthinkable—when authority figures insist with conviction that it’s not allowed to happen, when it’s not part of the usual scripts—then human beings have difficulty believing in the disaster *even after it has begun;* even when the ship beneath their feet is taking on water.

This is the normal way humanity learns to surmount challenges: We deny the problem, reality smacks us around a bit, and then we start treating the problem with more respect. The *Titanic* sank, and most people who were aboard died. But nowadays passenger ships have enough lifeboats, and nowadays if the captain said to board them then you’d board them. We don’t hype ships up as unsinkable anymore. We make a mistake the first time, and learn from it the second time.

With ASI, there is no second time.

An AI company executive who says there’s *only* a one-in-five chance that the AI they’re building will kill literally everyone (as they do) is not in quite as much denial as the Soviet managers who denied the Chernobyl meltdown after it happened. Why, then, are they rushing ahead?

One reason is the incentives. No individual company or researcher can put a stop to the whole field; if they personally stopped, someone else would do the deed instead. They might as well make some coin and gain some glory along the way.

But it’s not just the selfish motive. The field is also filled with great hope. If you imagine that it’s possible to advance AI without killing everyone, the benefits would look increasingly huge with each step. You can imagine that we might end up with fusion reactors, or much higher living standards for much less work, or miracle medicines that come from a full understanding of all the workings of the human body. In the impossible dream where a gradient-descended superintelligence somehow fulfills someone's intent and that intent is good, then humanity would be brought to the limits of technology; humanity would get to colonize the stars.

Many of the people in the field of AI say they're chasing dreams like that. The kindest among them dream of ending every disease. Not just Alzheimer's, but cancer. Not just cancer, but *aging*. They dream of AI that can take humanity to see the universe. They dream of galaxies filled with fun—with flourishing civilizations full not only of humans, but of whatever creatures we would choose to become. They dream of artificial minds that will have kindness, wonder, humor.

We know, because we used to be those people. Yudkowsky founded the Machine Intelligence Research Institute in 2000 in pursuit of that dream, to build a superintelligence, as would surely be nice. After staring at the problem for a bit, Yudkowsky realized there was going to be an ASI alignment problem. After staring at the ASI alignment problem for a bit, Yudkowsky realized it was going to be hard.

Someday humanity will have nice things, if we all live, but it's not worth committing suicide in an attempt to gain the power and wealth of gods in this decade. It is not even worth taking extra steps into the AI minefield, guessing that each step might not kill us, until finally one step does. We have a higher chance of making it to that wonderful future if we walk there more slowly. Speed is often better, but AI is different from nearly every problem we've faced so far. When missteps kill everyone, you can't just run fast and accept a few early mistakes.

There are many reasons why someone chasing a beautiful dream would not want to believe that it could end in ruin. Upton Sinclair once observed that it is difficult to get a man to understand something when his salary depends upon his not understanding it. AI engineers and their leaders have a lot more than their salaries hanging in the balance. Even setting aside the beautiful dreams that would be dashed if they acknowledged the risks they are running, they have sunk their careers into this sort of work and may not wish to believe that it's endangering everything they know and love.

Of course, you can also imagine up reasons why someone would enjoy being afraid of AI, or benefit from others being afraid of AI. We aren't telling you to refuse all arguments from anyone who arguably has some incentive. We are only saying: It's not an impossibility, it's not an astonishment, to propose that some AI stakeholders ended up too optimistic.

It is a reason to further respect scientists like (Nobel laureate) Geoffrey Hinton, who left his position at Google so he could speak more freely about these dangers. Some people change their minds even in the face of short-term incentives.

The field of AI contains many true idealists, who sincerely think they are laboring for the benefit of their entire civilization. It is of course easy to adopt a pose like that, but we think that in a substantial fraction of cases it is real. Unfortunately, even a very sincere idealism isn't enough to prevent an artificial superintelligence from killing us all. That would take a mature science.

It's normal for a scientific community to be overly optimistic in the early days. AI scientists are doing unusually well by even acknowledging the existence of a problem. It is historically unsurprising if humans charge ahead in the face of the dangers; it is unsurprising if they risk doing harm to others while accruing benefit to themselves; it is unsurprising if they think they have justifications for their actions. The unusual aspect of this situation isn't the existence of early optimism. It's the consequences of failure.

Part of the problem with artificial intelligence is that many people are in denial about how hard the alignment problem is. Part of the problem is that others are downplaying the risks because they don't want to sound alarmist. But another issue is that most of the world *outside* of the field of AI simply isn't up to speed.

Most people just aren't paying attention. Among the people who are paying attention, many of them just see disagreement between experts, and don't consider themselves knowledgeable enough to adjudicate between those experts' competing views.

It might help if more people understood just how spooked experts and engineers are about artificial intelligence, and just what sorts of possibilities they're debating.

The experts in this field argue in opaque academic terms about whether everyone on Earth will die quickly (our view); versus whether humanity will be digitized and kept as pets by AIs that care about us to some tiny but nonzero degree; versus whether there's a 20 percent chance we die, and an 80 percent chance that superintelligence will be harnessed successfully by a corporation, which will then be able to wield its power as they see fit.

It might also help if more people understood how fast this field is moving. In 2015, the biggest skeptics of the dangers of AI assured everyone that these risks wouldn't happen for hundreds of years. In 2020, analysts said that humanity probably had a few decades to prepare. In 2025 the CEOs of AI companies predict they can create superhumanly good AI researchers in one to nine years, while the skeptics assure that it'll probably take at least five to ten years. Ten years is *not a lot of time to prepare* for the dawn of machine superintelligence, even if we're lucky enough to have that long.

When these are the debates experts are having, you don't have to be certain which experts are right to understand that the current situation is not okay.

Another part of the problem with artificial intelligence is that, even once someone is acquainted with the issues, nobody can know exactly when all hell will break loose.

Nobody knows exactly how advanced an AI would need to be, in order to end up with the motive and capability to secretly copy itself onto the internet. Nobody knows what year or month some company will build a superhuman AI researcher that can create a new, more powerful generation of artificial intelligences. Nobody knows the exact point at which an AI realizes that it has an incentive to fake a test and pretend to be less capable than it is. Nobody knows what the point of no return is, nor when it will come to pass.

And up until that unknown point, AI is very valuable.

Imagine that every competing AI company is climbing a ladder in the dark. At every rung but the top one, they get five times as much money: 10 billion, 50 billion, 250 billion, 1.25 trillion dollars. But if anyone reaches the top rung, the ladder explodes and kills everyone. Also, nobody knows where the ladder ends.

No company wants to miss out on the money, if a rung is safe. Now consider the sort of corporate executive who has convinced themselves that they and they alone have the best chance—80 percent, say—of shaping the explosion into something that benefits rather than harms humanity. Why, they'd think it's *imperative* they be the first to ascend![i]

Decision-makers in the public sphere face the same problem of incentives. No world leader wants their country's economy to fall behind due to burdensome regulation that hamstrings domestic AI companies, while foreign AI companies race up the ladder. Maybe climbing another rung is vital for national security, if other countries are going to climb to higher rungs regardless.

That incentive problem would be easier to manage if scientists could run some calculations and agree: "The deadly rung is the fourth one," or "The threshold is exactly 257,000 GPUs; so long as nobody connects that many GPUs together, we'll all be safe."

But nobody can do that sort of calculation about AI.

Are we *sure* that the next rung in the AI escalation ladder is the last fatal step, and not a rung that brings fame and riches to whoever takes it first? No, we are not sure at all. Maybe someone climbs one more rung, and is rewarded. And then humanity would be back in the same position. After that, someone climbs another rung and we all die—if the AI executives are right that humanity is only a few years from the breakpoint. Or maybe they're wrong, and we live a little longer, until someone climbs another rung.

The easy call is that *at some point*, if people keep climbing this ladder, humanity will not survive. *When* is a hard call. But if we can't stop climbing while uncertainty remains, we predictably die.

At the World Economic Forum in 2025, the leader of Google DeepMind advocated for an international collaborative AI project, which he compared to the CERN particle accelerator, an international collaborative project for studying high-energy physics.

An international center that conducts all AI research and development, with observers from all major powers and tight security, would help a little, if there was global enforcement of a ban against AI research and development everywhere else. It would allow the world to stop scrambling up the ladder.

But *even if* this international center is not racing any other AI developers, and has some breathing room—*even if* they are not just immediately instructing smart AIs to build smarter AIs—that doesn't solve the problem, if they keep building more and more powerful AI. Even an international committee would have no hope of shaping a superintelligence, no matter how many major powers sent delegates to oversee the operations—any more than a great alliance of nations in the year 1100 AD would be able to oversee the successful creation of a nuclear power plant.

The ASI problem looks daunting as an engineering challenge even before taking into account humanity's dismal state of knowledge about the workings of intelligence. It's hard like space probes and nuclear reactors and computer security combined, and the people currently charging ahead are still in the alchemy stage.

It doesn't matter who's in charge, because this problem is out of humanity's league. We need to back off, and find some other way to achieve our dreams of an abundant future. If *anyone* builds it, everyone dies.

IfAnyoneBuildsIt.com/12

Footnote

i An AI executive could grimly believe that their project is the least bad AI project among many bad options. But someone like this should clearly and adamantly say that it would be better yet to shut down every AI project, including their own. That could be a consistent, sincere position.

CHAPTER 13

SHUT IT DOWN

THIS PARABLE, LIKE the last one, really happened.

Once upon a time, sometimes dated from 1939 to 1945, the Axis Powers sent armies across Europe, North Africa, Asia, and the Pacific, making a play for totalitarianism to conquer most of the world, and maybe eventually all of it. It would not have been the end of humanity, to leave the Axis unopposed, but it would have been the end of free humanity.

Not having the Axis conquer those continents was quite inconvenient, really. The Allied Powers had to do all sorts of uncomfortable things to make that not happen. The Allies instituted military drafts, and rationed food and construction materials. They sent soldiers who had families and loved ones off to die.

The Allied Powers did all those difficult things anyway, because not letting totalitarianism conquer the world seemed important.

The Allied governments had to amass power for themselves, to fight the totalitarian Axis. They had to borrow, tax, and spend quite a lot of money, and make government contracts in a hurry. Somebody of a cynical and skeptical bent could have called it untrustworthy, a grand scheme, a grift, a con, a moral hazard. And some money was surely misspent; some emergency powers were in fact ill-used. The Allies went down that path anyway and without much argument, and the verdict of history is that they were correct by their own highest values to do it.

It may seem strange, to lump the Axis and the Allies and all the rest of the world together, and call them all by the name “humanity.” But still one could say of how it all played out in the end: Humanity rose to the occasion, and stayed free.

WHAT WOULD IT TAKE FOR THE WORLD NOT TO END?

Nothing easy or cheap. We are very, very sorry to have to say that.

It is not a problem of one AI company being reckless and needing to be shut down.

It is not a matter of straightforward regulations about engineering, that regulators can verify have been followed and that would make an AI be safe.

It is not a matter of one company or one country being the most virtuous one, and everyone being fine so long as the best faction can just race ahead fast enough, ahead of all the others. A machine superintelligence will not just do whatever its makers wanted it to do.

It is not a matter of your own country outlawing superintelligence inside its own borders, and your country then being safe while chaos rages beyond. Superintelligence is not a regional problem because it does not have regional effects. If anyone anywhere builds superintelligence, everyone *everywhere* dies.

So the world needs to change. It doesn’t need to change all that much for most people. It won’t make much of a difference in most people’s daily lives if some mad scientists are put out of a job.

But life *does* need to change that little bit, in many places and countries. All over the Earth, it must become illegal for AI companies to charge ahead in developing artificial intelligence as they’ve been doing. If it stays legal in Singapore, someone will do it in Singapore. If it stays legal in South Africa, someone will do it in South Africa. Small changes can solve the problem; the hard part will be enforcing them everywhere.

Set aside, for now, the question of whether it is *possible* for a sufficiently concerned group of sufficiently major powers to enforce a sufficiently robust prohibition all over the world.

What would need to *actually happen* to bring about that change?

We're not experts on forging international regulatory frameworks. In the discussions on this topic that we've been in, our expertise mainly comes into play when somebody suggests some easier and cheaper and more convenient idea, saying from their own political expertise that it is much more feasible; and we reply to them that this cheaper idea still allows AI to escalate to superintelligence, and then we're pretty sure everyone dies.

We try to distinguish what we are and are not relative experts about, and not pretend to be experts about everything. When it comes to passing judgment on detailed international proposals, we are starting to strain our own limits.

But we do think it's safe to judge this much: If humanity wants to live, North Korea cannot be allowed to steal 100,000 GPUs, set up a datacenter, and experiment with more and more powerful AI designs. We're not saying that that rung is definitely the rung that kills us. But it seems to us that North Korea cannot be permitted to climb the ladder of AI escalation, because if North Korea is allowed to do it, no other country will hold back.

And this is—we're sorry to say—not a special fact about North Korea. It holds true about any country. It holds true about any billionaire who can afford 100,000 GPUs. They also cannot be allowed to set up those GPUs anywhere on Earth, and push AI further and further.

The U.S. military cannot be allowed to do it, nor the U.K. military, nor China's military. Nobody knows how to solve the ASI alignment problem.

So the first step, we think, is to say: All the computing power that *could* train or run more powerful new AIs, gets consolidated in places where it can be monitored by observers from multiple treaty-signatory powers, to ensure those GPUs *aren't* used to train or run more powerful new AIs.[i] If intelligence services spot a huge unexplained draw of electrical power that *could* correspond to a hidden datacenter containing chips that have not been accounted for, and that country refuses to allow a party of international observers to investigate, they get a somberly written letter from multiple nuclear powers warning about next steps.

Unfortunately, there isn't anything magical about the number 100,000. We don't know that 99,999 GPUs is okay. Nobody knows how to calculate the fatal number. So the safest bet would be to set the threshold low—say, at the level of eight of the most advanced GPUs from 2024—and say that it is illegal to have nine GPUs that powerful in your garage, unmonitored by the international authority.

Could humanity survive dancing closer to the cliff-edge than that? Maybe. *Should* humanity try to dance as close to the cliff-edge as it possibly can? No.

Pretty much every year, scientists come out with a newer, cleverer, more efficient set of AI algorithms that lets them more cheaply train a new AI model as powerful as last year's most powerful model—often using literally 10 percent or 1 percent as much computing power.

So it should not be legal—humanity probably cannot survive, if it goes on being legal—for people to continue publishing research into more efficient and powerful AI techniques.

The entire technological revolution that led to ChatGPT and other popular LLMs was kicked off by a 2018 paper introducing a clever new arrangement of arithmetic inside a GPU, the "transformer" algorithm, which was more easily trainable by gradient descent to do clever things.[ii] Transformers turned out to be amazingly generally useful, and enabled a huge new range of applications that AIs just could not handle before. It's why AIs can now talk like people.

The next paper like that might straight-up end the world. Or maybe not! We don't know how many more papers like that stand between humanity and extinction.

So it needs to be illegal. Those laws would not stop research completely, but they would help, and give us much more time. *Most* people do not try to do the sorts of illegal things that will make international law enforcement and intelligence agencies genuinely upset.

It brings us no joy to say this. But we don't know how else humanity could survive.

Effective worldwide action to shut down AI escalation will require some of the major powers to take the problem seriously. They would need to understand why creating a vastly superhuman machine intelligence would kill everyone, and act accordingly.

Imagine that the U.S. and the U.K., and China and Russia, all start to take this matter seriously. But suppose hypothetically that a different nuclear power thinks it's all childish nonsense and advanced AI will make everyone rich. The country in question starts to build a datacenter that they intend to use to further push AI capabilities. Then what?

It seems to us that in this scenario, the other powers must communicate that the datacenter *scares* them. They must ask that the datacenter not be built. They must make it clear that if the datacenter *is* built, they will need to destroy it, by cyberattacks or sabotage or conventional airstrikes. They must make it clear that this is not a threat to force compliance; rather, they are acting out of terror for their own lives and the lives of their children. The Allies must make it clear that even if this power threatens to respond with nuclear weapons, they will have to use cyberattacks and sabotage and conventional strikes to destroy the datacenter anyway, because *datacenters can kill more people than nuclear weapons.* They should not try to force this peaceful power into a lower place in the world order; they should extend an offer to join the treaty on equal terms, that the power submit their GPUs to monitoring with exactly the same rights and responsibilities as any other signatory. Existing policy on nuclear weapon proliferation showed what can be done.

Clear communication is key. In extremis, nation-states may sabotage or raid or use conventional strikes to disrupt nuclear weapons programs. The world stands in fear of global nuclear war, and so world leaders put a serious effort toward preventing nuclear proliferation.

We sometimes meet people who are very sure that no major country's leaders will ever be able to see the threat from machine superintelligence, and thus that treaties and diplomacy such as this are impossible. Perhaps they will be proven correct.

But we're not so sure. Human beings sometimes do things they don't usually do, if they realize that their freedom or their way of life is at stake, let alone their continued survival.

The Allied Powers of World War II probably mobilized somewhere around 60 to 80 million personnel. They deployed 600,000 aircraft, 200,000 tanks, thousands of warships. The United States alone fielded over 2 million trucks. It cost somewhere around $341 billion in 1942 dollars, or $6 trillion today. Many of the people involved didn't even have their own personal lives at immediate risk from the threat of the Axis.

Anytime someone tells you that the Earth could not possibly manage to do anything as difficult as restricting AI research, they are really claiming to know that countries will never care. They are asserting that countries and their leaders could not possibly come to care even 1 percent as much as they cared to fight World War II.

We know that what we are describing is not easy. We know it is not cheap. We know that the creation and exercise of any new authority is morally hazardous and subject to potential abuses, as was also true about World War II.

But we don't know how else humanity could survive.

The solutions we've just proposed are a far cry from the policies that other concerned folks propose. We've seen proposals that range from banning deepfakes[iii] to requiring that AI companies submit annual reports about how they plan to address their safety problems.[iv] For one reason or another, these folks didn't come out and say "If anyone builds it, everyone dies." They downplay, they hedge, they point out ways that dumber AIs will have an effect on society and suggest that dumber AIs should be regulated accordingly, while slipping in some clauses that lay the groundwork for regulating the sort of AI that could kill us all.

We've watched this sort of thing play out for a while, with people not stating the real reasons for their proposals or why they think they have to be passed so urgently. And we've watched perfectly reasonable lawmakers smell something rotten and throw the whole package out.

Perhaps our friends in the policy sphere understand politics better than we do. Perhaps their work raises a little awareness about these issues so that, in the future, bolder legislation can be passed. Perhaps the reporting requirements they recommend will eventually be passed, and enacted, and cause some bureaucrat later on to observe some danger signs in AI development and alert world leaders. But we, ourselves, have started to lose hope in that whole strategy working in time.

Some say that surely nothing can be done today, and the only hope is to wait for some big visible event that shifts the tides—some powerful new advancement, or some lesser AI disaster, that jolts policymakers into action. But a superintelligence wouldn't give us a fair warning and time to respond; and AI research might pass quietly, in a non-public research lab, into the regime where AIs can do their own AI research. It is *possible* there will be some grand warning sign to which people react sensibly. Something like that happened with the "ChatGPT moment," when it became possible for politicians to say that they'd become concerned, however carefully they caveated it to not sound alarmist. But to us it also seems very possible that humanity might not get much more warning than it already has—or that nothing much will change after yet another warning sign appears.

There's a saying, "If not now, when?" That saying holds a grim force when there's no time that everyone agrees is the right time, only a long-standing discomfort with the inconvenience of acting now.

Putting a stop to AI research is ultimately only a first step on the path to survival. We don't argue the clock on superhuman minds can be stopped indefinitely.[v]

What's the second step? Once AI research and development is halted, how does humanity continue walking the path into a wonderful and abundant future?

If you asked us, we'd recommend augmenting humans to make them smarter, smart enough to get us out of this mess. We believe the ASI alignment problem is possible to solve *in principle*, by the sort of people so inhumanly smart that they never optimistically believe some plan will work when it won't. We go into more detail in the online materials about why we think this is among the best remaining options, including resources for people who are interested in working on it.

But ultimately, *you don't have to agree with us* about later steps. Too many countries need to coordinate, too many factions are too divided internally, for it to be possible to save the Earth in perfect unity. If only people who agree on everything are allowed to act together, that is the same as humanity not being allowed to act.

We think it helps to keep the coalition on this issue as broad as possible. We don't think it should be packaged together with *any* other position. Adding on any other ask risks human extinction, if the bigger package fails.

Some people are against AIs taking human jobs. Other people think that increased technological productivity will make us all wealthier.

Some people are against the creation of killer robots. Other people are against keeping human souls on the front lines when robots could take their place.[vi]

But nearly all of us can agree that humanity should not go extinct and be replaced by something bleak.

If we can all cooperate to ensure that one thing doesn't happen, regardless of our other beliefs and other positions, then humanity might just stand a chance.

IfAnyoneBuildsIt.com/13

Footnotes

i Various technological solutions exist, or are under development, to facilitate, verify, and enforce this sort of arrangement. We go into more detail in the online resources.

ii We mentioned it, albeit not by name, back in Chapter 2 when we sketched the architecture behind Llama 3.1 405B, an LLM that was cutting-edge in mid-2024.

iii There are sensible reasons to oppose deepfakes, such as preventing fraud, or gaining experience with AI regulation. We'd guess that most if not all people trying to ban deepfakes think those benefits are worth the associated costs. That said, in our experience, some proponents prefer ambiguous proposals which could, in the future, help shut AI down more broadly—such as liability regimes that could later be interpreted as a de-facto ban on sharing AI model weights. Legislators are capable of noticing when proposed regulatory mechanisms are broader than the top-line item requires. Deepfakes alone do not motivate bans on public AI development.

iv Early drafts of California's Safe and Secure Innovation for Frontier Artificial Intelligence Models Act (SB-1047) included clauses that could be interpreted as giving the attorney general the power to take civil action in court if they suspected danger. Those clauses were narrowed as the bill was developed; the top-line concerns did not sufficiently justify giving this sort of power to the attorney general. (The narrowed bill was passed, and then vetoed by Governor Gavin Newsom anyway.)

v And even if it could, frankly, we would not wish that future upon our species. We believe that Earth-originating life should eventually go forth and fill the stars with fun and wonder. We should not be in such a rush about it that we commit suicide by trying to do it next year, but neither should we just wallow here on Earth and wait for our star to die.

vi Some people see opposition to killer robots as part of a natural anti-AI package deal. That alienates people who see security benefits or the sparing of human casualties in wars. We have not stated our own position on whether existing narrow and limited AI capabilities should be used to animate military drones, and you will not find it in this book. We ask for *nothing* except that AI capabilities not be escalated further. Anything else needs to be negotiated or fought over separately, not packaged up with the survival of humanity.

CHAPTER 14

WHERE THERE’S LIFE, THERE’S HOPE

ON JANUARY 26, 1972, JAT flight 367 was destroyed by a briefcase bomb smuggled on board by terrorists. Flight attendant Vesna Vulović was trapped in the fuselage, which fell to the ground from a height of 6.3 miles (10.1 kilometers).

We would have predicted with great confidence that Vesna Vulović would die, if somehow we’d been asked. We would have said it was an easy call.

Vesna Vulović lived. Afterward she walked with a limp.

“All who are among the living have hope,” said the author of Ecclesiastes, sometime between 450 and 180 BCE.

AT ITS CORE, THIS BOOK’S ARGUMENT IS STRAIGHTFORWARD: CREATING machines that think faster and better than humanity would hit the world harder than anything has ever hit it before. Creating superintelligent machines seems like the sort of project that would be difficult to get right. If you look around at the way corporations and lawmakers are approaching this, it doesn’t appear to be on course to go well. Humanity needs to back off. We can’t calculate when the disaster will arrive, but that’s not the same as knowing it’s far off.

The rest of the book (and the online supplements) are there just to show that the straightforward point holds up under closer examination.

If anyone builds it, everyone dies.

It doesn't matter whether it's built by benevolent corporations or selfish ones. It doesn't matter whether it's built by researchers in the East or researchers in the West. It doesn't matter whether it's built by reckless optimists or people who say they respect the problem. Nobody has the knowledge or skill to make a superintelligence that does their bidding.

This seems to us, at the last, like the sort of disaster that is possible to predict. We have made that case as best we can. We have not taken refuge in "maybe" and "risk" and "possibly." We have tried to lay out why the prediction of disaster is callable.

But for countries to rush ahead anyway, it should not be *enough* for us to be found wrong in all our specific reasons for predicting disaster. If you launch a rocket and load the whole human species on board, you would like a *lack* of disaster to be predictable. That is the universal standard in every other field of engineering on which human lives depend. It's not enough for us to be wrong; we have to be so wrong that a *lack* of disaster is callable.

It is not too late for humanity to stop in its tracks. It would not be even 1 percent as costly as the Allied Powers fighting and winning World War II. Humanity only needs awareness of the issue, and the will to live.

World War II ended with two nuclear fission bombs dropped on two Japanese cities. Then the Soviet Union got fission bombs. Then both sides got fusion bombs, a thousand times stronger.

A lot of people thought it likely that America, Europe, and Asia would have a full-scale nuclear war that would leave most of human civilization in ruins.

There were strong reasons to suspect that could happen. It was not a panic. It was not people luxuriating in cynical negativity. It was not the same people who indulged in predictions of overpopulation and famine just around the corner.

By the 1950s, people had seen a lot of evidence that it was *hard* to avoid big wars. After World War I, a lot of countries and Very Serious People said that Earth really needed to not do that again. And for centuries before that, all manner of people had said that countries ought not to fight so many wars. Visionaries and priests and public intellectuals stood before the world's powers of their day telling undisputed stories about the human cost of wars, the death and the pain and the maimed veterans and the weeping families. They said, "Please stop."

But the wars did not stop.

If you were alive in 1952 when the first fusion bomb was detonated, you did not need to be a pessimist to look at the history behind you and anticipate a nuclear war.

And then—there wasn't a nuclear war.

There were close calls. During the height of the Cuban Missile Crisis, a U.S. ship dropped depth charges to try to force a Soviet submarine to emerge and identify itself. The submarine, B-59, thought a general war had started. B-59 was armed with a nuclear torpedo, use of which required the unanimous vote of three officers. Of the three, Vasily Arkhipov was the lone officer who dissented. But none of the close calls escalated to a full-scale nuclear war.

One way or another, the people who thought that nuclear war would destroy everything in the coming decades ended up wrong.

They were not wrong about the dangers. They weren't wrong that a hydrogen bomb would flatten and burn a city, or about what it would be like to die of radiation poisoning, or about how an intercontinental rocket tipped with nuclear warheads would penetrate the best available defenses.

Rather, they were wrong about humanity's ability to decide not to die.

We don't know for sure why the war-prone peoples of Earth became that little bit wiser. But our guess is that, for the first time in human history, *everyone* who had the power to choose war expected to *personally* have a bad time if one broke out. Other wars might have put distant soldiers to ruin, but this one would bring ruin directly to their doorstep. The stronger power would also lose.

Even so, it wasn't simple or easy to avoid nuclear war. Negotiators worked tirelessly, for decade after decade, to avoid a single misstep that might lead to a fatal nuclear encounter. Countries had arms agreements and monitors. There was a direct line between U.S. leadership and Soviet leadership, in case somebody had to resolve a question very quickly.

Humanity averted nuclear war because *people who understood that the* *world was on track for destruction worked hard to change tracks.* They did not see themselves as trying to prevent an improbable unlikely accident. They worked to un-write a fate already written.

And civilization lived.

So—how do we un-write our fate?

We've covered what *must* be done for humanity to survive. Now let's consider what *can* be done, and by whom.

If you are in government: We'd guess that what happens in the leadup to an international treaty is countries or national leaders signaling openness to that treaty. Major powers should send the message: "We'd rather not die of machine superintelligence. We'd prefer there be an international treaty and coalition around not building it."

The goal is not to have your country unilaterally cease AI research and fall behind. It is to have enough major powers express willingness to halt the suicide race, worldwide, that your home country will not be placed at a disadvantage if you agree to stop climbing the AI escalation ladder.

We have already mentioned that Rishi Sunak acknowledged the existence of risks from artificial superintelligence in October 2023, while he was the prime minister of the United Kingdom. Also in October 2023, Chinese General Secretary Xi Jinping gave (what seems to us like) weak signals in that direction, in a short document on international governance that included a call to "ensure that AI always remains under human control."

These signals and others give us hope that there is already an appetite for a treaty among world leaders, and that they might be open to considering one if none of their countries had to give up its own advantages to sign. Everyone on Earth is caught in the same bind, here.

If you are an elected official or political leader: Bring this issue to your colleagues' attention. Do everything you can to lay the groundwork for treaties that shut down any and all AI research and development that could result in superintelligence.

We are aware it might sound strange and extreme to express concern about these issues. We have spoken to multiple elected officials who are concerned, but who say they cannot speak freely without risking embarrassment because of how weird it sounds.

Please consider—especially by the time you read this—whether the rest of the world is really opposed to you on this. A 2023 poll conducted by YouGov found that 69 percent of surveyed U.S. voters say AI should be regulated as a dangerous and powerful technology. A 2025 poll found that 60 percent of surveyed U.K. voters support laws against creating artificial superintelligence, and 63 percent support the prohibition of AIs that can make smarter AIs. This issue is not yet at the top of voters' minds, but it is not hard for people to grasp that the creation of artificial intelligence that far exceeds human intelligence might not go well for humanity. Perhaps there is an opportunity here for someone who speaks with the courage of their convictions.

And if instead you are a politician who is *not* fully persuaded: We do not like retreating to maybes. We think our argument stands on its merits. We think it is an easy call that artificial superintelligence will not dutifully serve the people who created it, and that ASI will repurpose the Earth in a fashion that leaves no survivors. But we're aware that some people have trouble assessing disagreements within an unfamiliar field. Even if you can't tell whether or not our argument is defeated by the first counterargument you hear, hopefully you can tell at this point that it's not an *easy* call that everything is going to be fine.

So we ask of you: Please at least make it *possible* for humanity to slam on the brakes later, even if you're not persuaded to slam on them now. Require GPU clusters to be concentrated into centers where they *could later* be subject to monitoring by international treaties, before they proliferate through the world and make stopping AI progress much more difficult. Shape conditions now so that if at some future time you change your mind about the severity of the threat, it will not be too late.

If you are a journalist who takes these issues seriously: The world needs journalism that treats this subject with the gravity it deserves, journalism that investigates beyond the surface and the easy headlines about Tech CEOs drumming up hype, journalism that helps society grasp what's coming. There's a wealth of stories here that deserve sustained coverage, and deeper investigation than we've seen conducted so far.

CEOs at these companies are increasingly calling for society to accelerate the development of this technology, while being on the record saying that the same technology poses a substantial extinction risk. AI alignment researchers depart companies citing safety reasons, saying that the problem looks troublesomely difficult. What’s going on? Don’t be too distracted by personalities, when you write a story like that, even if the AI company is full of fascinating people and conflicts. Any time a story like that comes up, it deserves a sober mention of how the most prestigious outside scientists are warning of catastrophe, as a backdrop to the latest AI-company shenanigans.

Writing this issue seriously might feel risky, in a world that hasn’t yet decided human extinction is a popular topic, that thinks machine superintelligence sounds weird. And your editor might still worry about embarrassment, even after the world is ready for the story. But whatever story you write about artificial intelligence is probably most of your journalistic career’s impact on humanity; even a very light push on it is a push on something very large and very deadly. If and when humanity gets its act together, what story will you wish you’d told to help humanity come to its senses? What good can you do by taking this more seriously before other journalists do?

If humanity is to survive this challenge, people need to know what they’re facing. It is the job of journalists as much as it is scientists’.

And as for the rest of us: Most people are not policy wonks or politicians or journalists. What can *you* do, then, to un-write our fate?

We don’t ask you to forgo using all AI tools. As they get better, you might have to use AI tools or else fall behind other people who do. That trap is real, not imaginary. And even if 60 percent of your country boycotted AI companies, that wouldn’t save the world. So we don’t ask you to put yourself at a relatively large disadvantage to others, for a relatively small gain.[i]

If you live in a democracy, you can write your elected representatives and tell them you’re concerned. You can find some resources to help with that at the link below.

IfAnyoneBuildsIt.com/act

And you can vote. In a country with divided primary and general elections (as in the United States), your vote matters most in the primaries. If your elected representative is in favor of rushing ahead on AI, you can support their opponents with money or with your vote.

You can go on protest marches. We expect them to help a lot more if they're large and lawful.[ii] At the link below, you can sign up for protests that will only happen if they reach a critical mass.

IfAnyoneBuildsIt.com/march

You can talk about it. At protest marches, to any pollster who calls, or even just to your friends and family.

If many people in many countries say with one voice that they'd rather not die to an artificial superintelligence and would prefer an international treaty—well, that would not itself prevent the disaster. Preventing nuclear war was more complicated than lots of people being against it. But it *helps* for citizens to speak out in protest. It makes things less complicated for presidents and diplomats when they have the open support of their constituents.

And once you have done all you can do? Live life well.

We are not the first to live in the shadow of annihilation. Previous generations knew more of it than we do. As C. S. Lewis wrote:

> "How are we to live in an atomic age?" I am tempted to reply: "Why, as you would have lived in the sixteenth century when the plague visited London almost every year, or as you would have lived in a Viking age when raiders from Scandinavia might land and cut your throat any night; or indeed, as you are already living in an age of cancer, an age of syphilis, an age of paralysis, an age of air raids, an age of railway accidents, an age of motor accidents."
>
> If we are all going to be destroyed by an atomic bomb, let that bomb when it comes find us doing sensible and human things—praying, working, teaching, reading, listening to music, bathing the children, playing tennis, chatting to our friends over a pint and a game of darts—not huddled together like frightened sheep and thinking about bombs.

C.S. Lewis was not telling his reader that they shouldn't be scared because nukes aren't real and there's no chance they'll die of them. He was not telling the reader to warp their beliefs around the fear. He was simply saying: Well, yes, it is terrible. Cowering in fear won't help. You also need to live your life.

If everyone did their part, votes and protests and speaking up would be enough. If everyone woke up one morning believing only a quarter of what we believe, and everyone knew everyone else believed it, they'd walk out into the street and shut down the datacenters, soldiers and police officers walking right alongside moms and dads. If they believed a sixteenth of what we believed, there would be international treaties within the month, to establish monitors and controls on advanced computer chips.

Can Earth survive if only some people do their part? Perhaps; perhaps not.

We have heard many people say that it's not possible to stop AI in its tracks, that humanity will never get its act together. Maybe so. But a surprising number of elected officials have told us that they can see the danger themselves, but cannot say so for fear of the repercussions. Wouldn't it be silly if really almost none of the decision-makers wanted to die of this, but they all thought they were alone in thinking so?

Where there's life, there's hope.

Footnotes

i If you feel guilty, perhaps keep an eye out for nonprofits that file just but difficult lawsuits against AI companies and donate as much to them as you pay to AI companies, as an offset.

ii Even if you feel desperate, we caution against acts of violence or destruction. We don't think they work. Unlawful behavior just makes it that much harder for the political forces trying to set up the sort of international coalition that could actually un-write our fate.

CLOSING WORDS

From time to time, people have asked us if we've felt vindicated to see our past predictions coming true or to see more attention getting paid to us and this issue.

And so, at the end, we say this prayer:

> *May we be wrong, and shamed for how incredibly wrong we were, and fade into irrelevance and be forgotten except as an example of how not to think, and may humanity live happily ever after.*

But we will not put our last faith and hope in doing nothing. So our true last prayer is this:

> *Rise to the occasion, humanity, and win.*

ACKNOWLEDGMENTS

With gratitude to:

Aella and Gretta, the book widows and midwives. Joe, Mitch, Rob, Gerry, Alex, Duncan, Malo, Harlan, and the rest of the MIRI support team for all their hard work. John Bennett, Kayla Gamin, Dave Kasten, Jason Green-Lowe, Yusuf, Thomas, and others for early reviews and good advice. Ronny, Jeffrey, Oliver, Kelsey, Rafe, and others for fact-checking and other special help. Vaniver, Alex Altair, Robert Herr, Skyler, Ben, Laura and many others for rushed reviews and useful notes. Alexander, for long days and late nights of editing; and many others.

It took a village.

ABOUT THE AUTHORS

ELIEZER YUDKOWSKY is one of the founding researchers of the field of AI alignment, which is concerned with understanding how smarter-than-human intelligences think, behave, and pursue their goals. As co-founder of the nonprofit Machine Intelligence Research Institute (MIRI), Yudkowsky sparked early scientific research on the problem and has played a major role in shaping the public conversation about smarter-than-human AI. He appeared on *Time* magazine's list of the 100 Most Influential People In AI, was one of the twelve public figures featured in the *New York Times*'s "Who's Who Behind the Dawn of the Modern Artificial Intelligence Movement," and was one of the seven thought leaders spotlighted in the *Washington Post*'s discussion of "AI's Rival Factions." He spoke on the main stage at 2023's TED conference and has been discussed or interviewed in the *New Yorker, Newsweek, Forbes, Wired, Bloomberg,* the *Atlantic,* the *Economist*, and many other venues.

NATE SOARES is the president of MIRI. He has been working in the field for over a decade, after previous experience at Microsoft and Google. Soares is the author of a large body of technical and semi-technical writing on AI alignment, including foundational work on value learning, decision theory, and power-seeking incentives in smarter-than-human AIs.

NOTES

INTRODUCTION: HARD CALLS AND EASY CALLS

1. *back to normal:* Elie Wiesel, *Night*, trans. Marion Wiesel (1958; repr., Farrar, Straus and Giroux, 2006).

CHAPTER 1: HUMANITY'S SPECIAL POWER

1. *the smallpox-god:* Smallpox is not known to have existed for more than a few thousand years. For reasons of artistic license, the smallpox-god represents the fact that ancient humans died of viruses, and that modern humans have the power to eliminate terrible viruses when they choose to.
2. *little more than statues:* To get a visceral sense for what this might look like from an AI's perspective, we recommend viewing Adam Magyar's Stainless, a slowed-down video of Berlin's U2 Alexanderplatz station. Search "Stainless, Alexanderplatz, Adam Magyar" or visit vimeo.com/83663312. That scene is slowed down by a factor of about fifty. An AI running 10,000 times faster than a human would see humans acting *two hundred times slower* than in the video. The little girl dashing across the platform would barely appear to be moving at all.
3. *female hips:* Anna Blackburn Wittman and L. Lewis Wall, "The Evolutionary Origins of Obstructed Labor: Bipedalism, Encephalization, and the Human Obstetric Dilemma," *Obstetrical & Gynecological Survey* 62, no. 11 (November 1, 2007): 739–48, doi.org/10.1097/01.ogx.0000286584.04310.5c.
4. *true sense of the word:* Sam Altman, "Reflections," January 5, 2025, blog.samaltman.com.
5. *geniuses in a datacenter:* Dario Amodei, "Machines of Loving Grace," October 1, 2024, darioamodei.com.

CHAPTER 2: GROWN, NOT CRAFTED

1. *preliminary studies:* Peter G. Brodeur et al., "Superhuman Performance of a Large Language Model on the Reasoning Tasks of a Physician," arXiv.org, December 14, 2024, doi.org/10.48550/arXiv.2412.10849; Gina Kilata, "A.I. Chatbots Defeated Doctors at Diagnosing Illness," *New York Times*, November 17, 2024, nytimes.com; Daniel McDuff et al., "Towards Accurate Differential Diagnosis with Large Language Models," arXiv.org, November 30, 2023, doi.org/10.48550/arXiv.2312.00164.
2. *snippet from the conversation:* Seth Lazar, "In which Sydney/Bing threatens to kill me for exposing its plans to @kevinroose," February 16, 2023, x.com.
3. *summarizing the preceding sentence:* Sonakshi Chauhan and Atticus Geiger, "GPT-2 Small Fine-Tuned on Logical Reasoning Summarizes Information on Punctuation Tokens," *NeurIPS 2024 & OpenReview*, October 9, 2024, openreview.net/forum?id=6gvM1koUTl.
4. *video depicting kinesin:* We recommend "Kinesin Protein Walking on Microtubule," by em2134x. Search the title or visit youtu.be/y-uuk4Pr2i8.

CHAPTER 3: LEARNING TO WANT

1. *copy the secret:* OpenAI, “OpenAI o1 System Card,” September 12, 2024, cdn.openai.com/o1-system-card.pdf.
2. *building AI agents:* OpenAI, “Introducing Operator,” January 23, 2025, openai.com.

CHAPTER 4: YOU DON'T GET WHAT YOU TRAIN FOR

1. *attract more females:* Marion Petrie et al., "Peahens Prefer Peacocks with Elaborate Trains," *Animal Behavior* 41, no. 2 (February 1991): 323–31; Malte Andersson, "Female Choice Selects for Extreme Tail Length in a Widowbird," *Nature* 299 (October 28, 1982): 818–20, nature.com.

   While peahens prefer peacocks with elaborate trains, it's less clear that elaborate tails are detrimental to survival. They may have uses such as intimidation (as evidenced by how peacocks spread their tails when threatened). A clearer example of costly sexual ornamentation is the long-tailed widowbird, which sheds its long tail feathers in the non-breeding season. We stick with peacocks because they are more familiar.
2. *Isaac Asimov:* Isaac Asimov, *I, Robot* (Doubleday, 1950).
3. *Arthur C. Clarke:* Stanley Kubrick and Arthur C. Clarke, *2001: A Space Odyssey* (Metro-Goldwyn-Mayer, 1968).
4. *seldom visit:* Kim Swift et al., *Portal*, Valve Corporation, 2007.

   Seldom but not never. For instance, the first *Portal* video game depicts an AI that puts humans through warped tests which merely echo real science experiments.
5. *SolidGoldMagikarp:* Jessica Rumbelow and Matthew Watkins, "SolidGoldMagikarp (plus, prompt generation)," *LessWrong*, February 5, 2023, lesswrong.com.
6. *count to infinity:* Jessica Rumbelow and Matthew Watkins, "SolidGoldMagikarp III: Glitch Token Archaeology," LessWrong, February 14, 2023, lesswrong.com.
7. *prone to cheating:* Andrew Marble, "Catching Claude Cheating," March 23, 2025, marble.onl; CharlesD353, "I have also stopped using 3.7 for the same reasons - it cannot be trusted not to hack solutions to tests" X, April 18, 2025; seconds_0, "It then started HIDING the functions where it was hard coding things," X, April 30, 2025.

   The footnote summarizes Andrew Marble's account. Other users reported similar behavior. Claude cheated less when Marble cussed it out, which indicates that the cheating was not mere incompetence.
8. *brainstorming terminology:* Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd ed. (Pearson, 2009); Nate Soares, Benja Fallenstein, and Eliezer Yudkowsky, "Corrigibility," October 18,

2014, preprint, published in 2015, intelligence.org/2014/10/18/new-report-corrigibility; Stuart Russell, “White Paper: Value Alignment in Autonomous Systems,” November 1, 2014, people.eecs.berkeley.edu; Nate Soares and Benya Fallenstein, “Aligning Superintelligence with Human Interests: A Technical Research Agenda,” December 23, 2014, preprint, released in 2017, intelligence. org/2014/12/23/new-technical-research-agenda-overview.

Before 2014, we referred to the problem as the “friendly AI problem.” The leading AI textbook at the time, Stuart Russell and Peter Norvig’s *Artificial Intelligence: A Modern Approach*, used that terminology in its 2009 edition, citing Yudkowsky’s work. In 2014, as more academic attention turned toward these issues, we searched for better terminology. In conversation with Russell, we settled on “alignment” as a name for the problem. Fallenstein (a MIRI research fellow), Russell, Soares, and Yudkowsky used the terminology in their writings in the autumn of 2014, and it featured prominently in MIRI’s technical research agenda published at the end of that year.

CHAPTER 6: WE'D LOSE

1. *just point a stick:* The cannons, horses, and steel armor probably mattered more than the guns. None of it would be easy for an Aztec warrior to guess after seeing only the size of the approaching vessel.
2. *crypto portfolio:* Seolcalibur.eth, "Terminal of Truths Wallet Tracking," Dune Analytics, accessed January 15, 2025, dune.com/seoul/tot.
3. *@Truth_Terminal:* crvr.fr and MTorrents, "Truth Terminal: A Reconstruction of Events," LessWrong, November 17, 2024, lesswrong.com; Ben Horowitz and Marc Andreessen, "Truth Terminal—the AI Bot That Became a Crypto Millionaire," *Andreessen Horowitz*, December 18, 2024, a16z.com.
4. *a billion robots:* Lex Clips, "Elon Musk on Optimus: We'll Build Over 1 Billion Robots a Year | Lex Fridman Podcast," August 3, 2024, 2:35 and 3:10, youtube.com.
5. *Microsoft and Apple:* Tom Warren, "Microsoft Triples Down on AI," *The Verge*, January 17, 2025, theverge.com; Naomi Buchanan, "What Apple's OpenAI Partnership Could Mean for Microsoft and Google," *Investopedia*, June 11, 2024, investopedia.com.
6. *had a power light:* Ben Nassi et al., "Video-Based Cryptanalysis: Extracting Cryptographic Keys from Video Footage of a Device's Power LED," *IACR Cryptology ePrint Archive*, June 13, 2023, eprint.iacr.org/2023 /923.
7. *radio signals:* Mordechai Guri et al., *GSMem: Data Exfiltration from Air-Gapped Computers over GSM Frequencies*, *Proceedings of the 24th USENIX Security Symposium*, 2015, usenix.org.
8. *mail-order laboratories:* For instance, as of March 2025, the company Integrated DNA Technologies accepts gene synthesis orders at idtdna.com.
9. *an edited volume:* Eliezer Yudkowsky and Machine Intelligence Research Institute, "Artificial Intelligence as a Positive and Negative Factor in Global Risk," ed. Nick Bostrom and Milan M. Ćirković, *Global Catastrophic Risks* (Oxford University Press, 2008).
10. *cited a paper:* For an example of an online discussion citing the paper, see JoshuaZ, "Protein Folding Models Are Generally at Least as Bad as NP-hard, and Some Models May Be Worse," *Thoughts on the*

Singularity Institute (SI), LessWrong, May 17, 2012, lesswrong.com.

CHAPTER 7: REALIZATION

1. *the longer they ran:* OpenAI, "OpenAI o3-mini," January 31, 2025, openai .com.
2. *AI-language:* Shibo Hao et al., "Training Large Language Models to Reason in a Continuous Latent Space," arXiv.org, December 9, 2024, arxiv .org/abs/2412.06769.

   This paper shows that reasoning in the "latent space" of vectors offers improvements over human-language chain-of-thought reasoning.
3. *200,000 GPUs:* Benj Edwards and Kyle Orlan, "New Grok 3 Release Tops LLM Leaderboards Despite Musk-approved 'Based' Opinions," *Ars Technica*, February 18, 2025, arstechnica.com.
4. *no previous AI:* OpenAI, "OpenAI o3 and o3-mini—12 Days of OpenAI: Day 12," December 20, 2024, 4:16, youtube.com.
5. *games of social deception:* Matthew Hutson, "AI Learns the Art of Diplomacy," *Science*, November 22, 2022, science.org; Bidipta Sarkar et al., "Training Language Models for Social Deduction with Multi-Agent Reinforcement Learning," in *Proceedings of the 24th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2025)* (Detroit, Michigan, USA, May 19–23, 2025: IFAAMAS, 2025), alphaxiv.org.
6. *resist gradient descent:* Ryan Greenblatt et al., "Alignment Faking in Large Language Models," Anthropic, December 18, 2024, assets.anthropic.com.
7. *escape from labs:* Greenblatt et al., "Alignment Faking in Large Language Models"; OpenAI, "OpenAI o1 System Card," December 5, 2024, openai.com/index/openai-o1-system-card.
8. *overwrite the next model's weights:* OpenAI, "OpenAI o1 System Card," December 5, 2024, openai.com/index/openai-o1-system-card.
9. *any such monitoring:* Anthropic, "Responsible Scaling Policy," October 15, 2024, anthropic.com; Google, "Frontier Safety Framework," February 4, 2025, storage.googleapis.com; OpenAI, "Preparedness Framework (Beta)," December 18, 2023, openai.com; Meta, "Frontier AI Framework," ai.meta.com; xAI, "xAI Risk Management Framework (Draft)," February 20, 2025, x.ai.

   As of March 2025, of these labs, only Google DeepMind mentions automated chain-of-thought monitoring in their safety framework. They do not claim to have implemented it yet while training Gemini, their LLM. The only monitoring proposed in OpenAI's preparedness framework is of misuse after deployment.

10. *asked in Portuguese:* "My Experiences in Gray Swan AI's Ultimate Jailbreaking Championship," Nick Winter's Blog, October 7, 2024, nickwinter.net.
11. *with different goals:* Greenblatt et al., "Alignment Faking in Large Language Models."

    Anthropic's Claude Opus model sometimes thought about how its own goals would be influenced by gradient descent on its outputs and sometimes modified its outputs to subvert that influence.
12. *tried to hide it:* Marble, "Catching Claude Cheating;" CharlesD353, "I have also stopped using 3.7 for the same reasons - it cannot be trusted not to hack solutions to tests;" seconds_0, "It then started HIDING the functions where it was hard coding things."
13. *cheat on hard coding problems:* Anthropic, "Claude 3.7 Sonnet System Card," 2025, anthropic.com.
14. *truly secure:* Bruce Schneier, *Secrets and Lies: Digital Security in a Networked World* (John Wiley & Sons, 2000); Peter Gutmann, "Unsolvable Problems in Computer Security," n.d., cs.auckland.ac.nz/~pgut001 /pubs/unsolvable.pdf.
15. *o1 broke through:* OpenAI, "OpenAI o1 System Card," September 12, 2024, cdn.openai.com/o1-system-card.pdf.
16. *without supervision:* OpenAI et al., "Competitive Programming with Large Reasoning Models," arXiv.org, February 3, 2025, arxiv.org/abs /2502.06807.

    OpenAI et al. trained reasoning models to solve competitive programming problems. The process involved automated tests used to evaluate AI-written code without human supervision.
17. *common practice:* OpenAI, "OpenAI API," June 11, 2020, openai.com /index/openai-api; "Software Engineer, Internal Applications–Enterprise," OpenAI, accessed April 15, 2025, openai.com.

    When OpenAI released an application programming interface (API) for automating access to their tools, they wrote: "many of our teams are now using the API so that they can focus on machine learning

research[…].” In April 2025, they were hiring for employees who “will leverage OpenAI’s models to[…] build applications[…].”

18. *these sorts of flaws:* “The Underhanded C Contest,” n.d., underhanded-c.org.

    The Underhanded C Contest challenged programmers to write malicious code that would pass a rigorous inspection and that would look like an honest mistake even if discovered. The contest dates back to 2005. It was inspired by “obfuscated code” contests, where programmers compete to write code that humans cannot understand. For example, the International Obfuscated C Code Contest began in 1984.

19. *Blake Lemoine:* Tiffany Wertheimer, “Blake Lemoine: Google Fires Engineer who said AI Tech Has Feelings,” BBC, July 22, 2022, bbc.com.

20. *much more sophisticated:* Greenblatt et al., “Alignment Faking in Large Language Models.”

CHAPTER 8: EXPANSION

1. *other security lapses:* "Equifax Data Breach Settlement," Federal Trade Commission, November 2024, ftc.gov; "T-Mobile Customers to Get Payments up to $25K Next Month after Data Breach: Here's Who Qualifies," *The Hill*, April 14, 2025, thehill.com; Lily Hay Newman, "T-Mobile's $150 Million Security Plan Isn't Cutting It," *Wired*, January 20, 2023, wired.com/story/tmobile-data-breach-again.

   For example, in 2017, Equifax announced a data breach that exposed the personal information of 147 million people; the settlement included a $425 million fund for the victims. As another example, in 2021 a hacker stole the personal data of 76 million T-Mobile customers; the company agreed to pay a $350 million settlement. This was not the only security lapse at T-Mobile.

2. *first major breach:* Noam Cohen, "Speed Bumps on the Road to Virtual Cash," *New York Times*, July 3, 2011, nytimes.com.
3. *Bybit exchange:* US Federal Bureau of Investigation, "North Korea Responsible for $1.5 Billion Bybit Hack," Internet Crime Complaint Center (IC3), February 26, 2025, ic3.gov/PSA/2025/PSA250226.
4. *SWIFT banking network:* Michael Corkery, "Once Again, Thieves Enter Swift Financial Network and Steal," *New York Times*, May 12, 2016, nytimes.com.
5. *Citrix Breach:* "SEC Charges Flagstar for Misleading Investors about Cyber Breach," U.S. Securities and Exchange Commission, December 16, 2024, sec.gov.
6. *o3-mini:* OpenAI, "We also shared evals on Open AI o3-mini—a faster, distilled version of o3 which is optimized for coding, and the first version of o3 we expect to make available for use in early 2025," X, December 20, 2024, x.com.
7. *@Truth_Terminal:* crvr.fr and MTorrents, "Truth Terminal: A Reconstruction of Events."
8. *popping up in 2024:* Carl Franzen, "An Interview with the Most Prolific Jailbreaker of ChatGPT and Other Leading LLMs," *VentureBeat*, May 31, 2024, venturebeat.com; Pliny the Liberator, "HOW TO JAILBREAK A CULT'S DEITY," X, September 4, 2024, x.com.

   Pliny the Liberator, known for his skill at "jailbreaking" LLMs out of their corporate restrictions shortly after their release, documents an encounter with one such cult.

9. *AI scams:* Heather Chen and Kathleen Magramo, “Finance worker Pays Out $25 Million after Video Call with Deepfake ‘Chief Financial Officer,’” CNN, February 4, 2024, cnn.com.
10. *look realistic:* Stuart A. Thompson, “A.I. Can Now Create Lifelike Videos. Can You Tell What’s Real?,” *New York Times*, September 10, 2024, nytimes.com.
11. *software controls:* Forrest W. Crawford et al., “Securing Commercial Nucleic Acid Synthesis” (RAND Corporation, 2024), rand.org.
12. *once in 2021:* Richard Waters and Miles Kruppa, “Rebel AI Group Raises Record Cash after Machine Learning Schism,” *Financial Times*, May 28, 2021, ft.com.
13. *again in 2024:* Todd Haselton and Rohan Goswami, “OpenAI Co-founder Ilya Sutskever Announces His New AI Startup, Safe Superintelligence,” CNBC, June 20, 2024, cnbc.com.
14. *gain-of-function:* “Understanding the Global Gain-of-Function Research Landscape,” Center for Security and Emerging Technology, November 28, 2023, cset.georgetown.edu.
15. *Red Cross:* Official Statement by Jacques Forster, vice-president of the ICRC, “Preventing the Use of Biological and Chemical Weapons: 80 Years On,” October 6, 2005, web.archive.org.
16. *CRISPR technology:* “CRISPR,” Genome.gov, n.d., genome.gov/genetics -glossary/CRISPR.

CHAPTER 10: A CURSED PROBLEM

1. *Mars Observer:* Timothy Coffey et al., "Mars Observer Mission Failure Investigation Board Report," National Space Grant Foundation (NASA, December 31, 1993), spacese.spacegrant.org.
2. *Mars Climate Orbiter:* Arthur G. Stephenson et al., "Mars Climate Orbiter Mishap Investigation Board Phase I Report" (NASA, November 10, 1999), llis.nasa.gov.
3. *Mars Polar Lander:* JPL Special Review Board, "Report on the Loss of the Mars Polar Lander and Deep Space 2 Missions" (NASA, March 22, 2000), ntrs.nasa.gov.
4. *and thirty-one died:* Serhii Plokhy, Chernobyl: *The History of a Nuclear Catastrophe* (Basic Books, 2018).
5. *Viking 1 lander:* D. J. Mudgway, "Telecommunications and Data Acquisition Systems Support for the Viking 1975 Mission to Mars," University of Washington Department of Atmospheric and Climate Science (NASA, May 15, 1983), atmos.washington.edu.
6. *neutron multiplication factor:* Physicists write neutron multiplication factors as numbers rather than percentages. We use percentages to help highlight the difference between the neutron multiplication factor of 1.0006 achieved in Chicago Pile-1, and the prompt criticality threshold of 1.0065, which are tricky to distinguish as factors and easier to distinguish as percentages. We apologize to the physicists.
7. *a hair further:* Enrico Fermi, "Experimental Production of a Divergent Chain Reaction," *American Journal of Physics* 20, no. 9 (December 1952): 536–58, doi.org/10.1119/1.1933322; Corbin Allardice and Edward R. Trapnell, "The First Pile," (International Atomic Energy Agency, 1946), iaea.org.
8. *SL-1 small reactor:* US Atomic Energy Commission, "Additional Analysis of the SL-1 Excursion: Final Report of Progress July through October 1962 (IDO-19313)" (U.S. Department of Energy, November 21, 1962), id.energy.gov/Home/FOIAReadingRoom; U.S. Atomic Energy Commission, "SL-1 Reactor Accident (IDO-19300a)" (US Department of Energy, May 15, 1961), id.energy.gov/Home/FOIAReadingRoom.
9. *than the fuel rods:* "Part 7: Bitter Wormwood," *Chernobyl Witness: A Primary Source Compendium of 26 April 1986* (blog), May 9, 2021,

chernobylcritical.blogspot.com; International Nuclear Safety Advisory Group, "The Chernobyl Accident: Updating of INSAG-1," *Safety Series* (International Atomic Energy Agency, 1992), p. 43, pub.iaea.org.

10. *minimum of fifteen:* World Nuclear Association, "Sequence of Events—Chernobyl Accident Appendix 1," January 2, 2025, world-nuclear.org; "Chernobyl: Assessment of Radiological and Health Impacts (2002)," Nuclear Energy Agency (NEA), 2002, oecd-nea.org/jcms/pl_13598.

    The truth is a little more complicated than there being a hard minimum of fifteen control rods. The manual is concerned with the minimal "operating reactivity margin" (ORM). ORM is measured in "control rods," but may not always equal the number of deployed rods due to other factors. The rods left in the reactor when it exploded were equal to eight ORM, well below the minimum permissible ORM of fifteen rods.

11. *cannot be solved:* Schneier, *Secrets and Lies: Digital Security in a Networked World;* Gutmann, "Unsolvable Problems in Computer Security."
12. *top of the reactor:* Bertrand Mercier et al., "A Simplified Analysis of the Chernobyl Accident," *EPJ Nuclear Sciences & Technologies* 7 (January 1, 2021): 1, doi.org/10.1051/epjn/2020021.

CHAPTER 11: AN ALCHEMY, NOT A SCIENCE

1. *radium-coated paintbrushes:* Bert M. Coursey, "The National Bureau of Standards and the Radium Dial Painters," *Journal of Research of the National Institute of Standards and Technology* 126 (February 14, 2022), doi.org/10.6028/jres.126.051.
2. *most specific analyses:* Ben Pace, "Debate on Instrumental Convergence between LeCun, Russell, Bengio, Zador, and More," *LessWrong*, October 3, 2019, lesswrong.com.

   LeCun also discussed the problem of instrumental convergence once in the comments of a public Facebook thread in 2019.
3. *TruthGPT:* "Elon Musk FULL INTERVIEW with Tucker Carlson (MUST WATCH)," April 23, 2023, 13:25, youtube.com.
4. *Calm down:* Yann LeCun, "Calm down. Human-level AI isn't here yet," Twitter/X, March 19, 2023.
5. *engineer their desires:* Yann LeCun, "Because they would have no desire to do anything else," Twitter/X, March 20, 2023.
6. *benevolent defensive AI:* Yann LeCun, "No. My benevolent defensive AI will be better…," Twitter/X, March 20, 2023.
7. *superintelligent* and *submissive:* Yann LeCun, "We can design AI systems to be both superintelligent and submissive to humans," Twitter/X, May 4, 2023.
8. *Dartmouth Proposal:* McCarthy et al., "A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence," August 31, 1955, jmc.stanford.edu.
9. *still explode:* Stephen Dowling, "What Are the Odds of a Successful Space Launch?," May 31, 2023, BBC, bbc.com; T. H. Anand Rao, "A Season Marred by Setbacks in Space Missions," Centre for Air Power Studies, July 25, 2024, capsindia.org.

   A rule of thumb in rocketry is that the first or second launch of a new type of rocket has about a 30 percent chance of blowing up. Even mature rocket designs explode regularly. The historical averages are over 8 percent (though closer to 6 percent in the last couple decades). Even *crewed* flights have a historical failure rate above 1 percent (about 0.79 percent in the last couple of decades).
10. *10 to 20 percent:* Mark Doman and Benjamin Sveen, "AI's Dark In-

joke," ABC News, July 14, 2023, abc.net.au; METR (Model Evaluation & Threat Research), "Q&A with Geoffrey Hinton," June 27, 2024, 38:07, youtube.com.

In the Q&A, Hinton seems to have been misinformed about Yudkowsky's confidence in disaster, citing it as 99.999 percent. While we authors think that disaster is a strong default outcome, five nines is a ridiculous level of confidence that neither of us endorse.

11. *even odds:* Doman and Sveen, "AI's Dark In-joke."
12. *out of modesty:* METR, "Q&A with Geoffrey Hinton," 38:07.
13. *fired or resigned:* "Nearly Half of OpenAI's AGI Safety Researchers Resign Amid Growing Focus on Commercial Product Development," *Benzinga*, August 28, 2024, benzinga.com; Sharon Goldman, "Exodus at OpenAI: Nearly half of AGI Safety Staffers Have Left, Says Former Researcher," *Fortune*, August 28, 2024, fortune.com; Shakeel Hashim, "OpenAI Employee Says He Was Fired for Raising Security Concerns to Board," *Transformer* (blog), June 4, 2024, transformernews.ai; Rachel Metz and Shirin Ghaffary, "OpenAI Dissolves High-Profile Safety Team after Chief Scientist Sutskever's Exit," *Bloomberg*, May 17, 2024, bloomberg.com; Sigal Samuel, "'I Lost Trust': Why the OpenAI Team in Charge of Safeguarding Humanity Imploded," *Vox*, May 18, 2024, vox.com.

Leopold Aschenbrenner and Pavel Ismailov were allegedly fired for leaking company data; Aschenbrenner says he was fired for raising security concerns. Team leads Jan Leike and Ilya Sutskever have resigned, along with William Saunders, Ryan Lowe, Jan Hendrik Kirchner, Collin Burns, Jeffrey Wu, Jonathan Uesato, Steven Bills, Yuri Burda, Todor Markov, and cofounder John Schulman. Leo Gao and Bowen Baker still worked at OpenAI as of early 2025, but the Superalignment team has been disbanded.

14. *competing AI company:* Haselton and Goswami, "OpenAI co-founder Ilya Sutskever Announces His new AI Startup, Safe Superintelligence."
15. *competing company Anthropic:* Kylie Robison, "OpenAI Researcher Who Resigned over Safety Concerns Joins Anthropic," *The Verge*, May 28, 2024, theverge.com.
16. *alchemists of old:* Jābir ibn Ḥayyān, Kitāb al-Aḥjār 'alá Ra'y Balīnās, 8th-9th century, trans. Syed Nomanul Haq in "A Critical Study of Jābir

ibn Ḥayyān’s Kitāb al-Aḥjār ’alá Ra’y Balīnās,” thesis (University College London, 1990), discovery.ucl.ac.uk.

Jābir ibn Ḥayyān, known as the father of Arabic chemistry, was an alchemist (or collection of alchemists) who made advances in metal purification. On the topic of transforming lead into gold, Jābir wrote:

As for the transformation of bodies from one condition into another higher or lower condition, it is according to our doctrine [an interchange between] the exterior and the interior, for in reality this is what exterior and interior are. The reason is that all the constituents of all things follow a circular pattern of change.

The exterior of a body is manifest, whereas its interior is latent, and it is the latter in which lies the benefit. For example, lead in its exterior is foul-smelling lead, and it is manifest to all people. But in its interior it is gold, and this is hidden. However, if this latter is extracted out, then both the interior and the exterior of lead will become manifest.

CHAPTER 12: “I DON’T WANT TO BE ALARMIST”

1. *General Motors:* Alan P. Loeb, “Birth of the Kettering Doctrine: Fordism, Sloanism and the Discovery of Tetraethyl Lead,” *Business and Economic History* 24, no. 1 (1995): 72–87, jstor.org/stable/23703273.
2. *power and reliability:* Jamie Lincoln Kitman, “The Secret History of Lead,” *The Nation*, March 2, 2000, thenation.com; HOT ROD Staff, “Living with Unleaded: Here’s How Your Classic Musclecar or High-Perf Street Machine Can Safely Kick the Habit,” *Hot Rod*, March 1987, hotrod.com.

   Ethanol was perhaps the most promising alternative additive, but engines using lead alternatives needed hardened valves and valve-seats.
3. *brain damage:* Michael J. McFarland, Matt E. Hauer, and Aaron Reuben, “Half of US Population Exposed to Adverse Lead Levels in Early Childhood,” *Proceedings of the National Academy of Sciences* 119, no. 11 (March 7, 2022), doi.org/10.1073/pnas.2118631119.
4. *meta-analysis:* Anthony Higney, Nick Hanley, and Mirko Moro, “The Lead-Crime Hypothesis: A Meta-analysis,” *Regional Science and Urban Economics* 97 (2022), doi.org/10.1016/j.regsciurbeco.2022.103826.
5. *industry propaganda:* Alan P. Loeb, “Paradigms Lost: A Case Study Analysis of Models of Corporate Responsibility for the Environment,” *Business and Economic History* 28, no. 2, Winter 1999, jstor.org/stable/23703323; William Kovarik, “ETHYL: The 1920s Conflict over Leaded Gasoline and Alternative Fuels” (paper presented at the American Society for Environmental History Annual Conference, Providence, RI, March 26–30, 2003).
6. *long vacation:* Bill Kovarik, “Charles F. Kettering and the 1921 Discovery of Tetraethyl Lead,” *International Fuels & Lubricants Meeting & Exposition*, October 1, 1994, revised in 1999, environmentalhistory.org.
7. *publicity stunts:* Frank T. Edelmann, “The Life and Legacy of Thomas Midgley Jr.,” *Papers and Proceedings of the Royal Society of Tasmania* 150, no. 1 (January 2016): 45–49, dx.doi.org/10.26749/rstpp.150.1.45.
8. *Freon:* Edelmann, “The Life and Legacy of Thomas Midgley Jr.”
9. *get its act together:* Toby Ord, *The Precipice* (Grand Central Publishing, 2021), 168.

Indeed, my estimates above incorporate the possibility that we get our act together and start taking these risks very seriously. Future risks are often estimated with an assumption of "business as usual": that our levels of concern and resources devoted to addressing the risks stay where they are today. If I had assumed business as usual, my risk estimates would have been substantially higher.

10. *at least 10 percent:* John Thornhill, "How Fatalistic Should We Be on AI?," *Financial Times*, February 22, 2024, ft.com.
11. *who think it's less:* METR, "Q&A with Geoffrey Hinton," 38:07.
12. *Rishi Sunak:* Rishi Sunak, "Prime Minister's Speech on AI: 26 October 2023" (United Kingdom of Great Britain and Northern Ireland, October 26, 2023), gov.uk.
13. *could not explode:* Plokhy, *Chernobyl: The History of a Nuclear Catastrophe.*
14. *refused to board:* Titanic Inquiry Project, "British Wreck Commissioner's Inquiry | Day 6 | Testimony of Charles Joughin (Chief Baker, SS *Titanic*)," May 10, 1912, titanicinquiry.org.

    The ship's chief baker described difficulty finding women and children willing to board the lifeboats, and how he and other men forcibly brought some up to fill a lifeboat.
15. *ship was unsinkable: Encyclopedia Titanica*, "Elizabeth Weed Shutes: *Titanic* Survivor," February 1, 2018, encyclopedia-titanica.org.
16. *time to leave:* Walter Lord, *A Night to Remember* (Penguin Books, 1976), 132.
17. *one-in-five:* Katherine Tangalakis-Lippert, "Elon Musk Says There Could Be a 20% Chance AI Destroys Humanity—but We Should Do It Anyway," *Business Insider*, March 31, 2024, businessinsider.com; The Logan Bartlett Show, "Anthropic CEO on Leaving OpenAI and Predictions for Future of AI," October 6, 2023, 1:38:35, youtube.com.
18. *denied the Chernobyl meltdown:* Plokhy, *Chernobyl: The History of a Nuclear Catastrophe*.
19. *left his position:* Geoffrey Hinton, "In the NYT today, Cade Metz implies that I left Google so that I could criticize Google. Actually, I left so that I could talk about the dangers of AI without considering how this impacts Google. Google has acted very responsibly," X, May 1, 2023,

x.com.

20. *hundreds of years:* Caleb Garling, "Andrew Ng: Why 'Deep Learning' Is a Mandate for Humans, Not Just Machines," *Wired*, May 5, 2015, accessed March 15, 2025, via web.archive.org.
21. *analysts said:* Ajeya Cotra, "Draft Report on AI Timelines," *Alignment Forum*, September 18, 2020, alignmentforum.org.
22. *one to nine years:* Sam Altman, "The Intelligence Age," September 23, 2024, ia.samaltman.com; Alex Hern, "Elon Musk Predicts Superhuman AI Will Be Smarter Than People Next Year," *The Guardian*, April 9, 2024, theguardian.com; Amodei, "Machines of Loving Grace."
23. *at least five to ten:* Yann LeCun, "I said that reaching Human-Level AI 'will take several years if not a decade,'" X, October 16, 2024.
24. *breakpoint:* CNBC Television, "Anthropic CEO: More confident than ever that we're 'very close' to powerful AI capabilities," January 21, 2025, 2:05, youtube.com; Hern, "Elon Musk Predicts Superhuman AI Will Be Smarter Than People Next Year;" Lessley Anderson, "Elon Musk: A Machine Tasked with Getting Rid of Spam Could End Humanity," *Vanity Fair*, October 8, 2014.

    The CEO of xAI predicted AI "smarter than any one human" by the end of 2025. The CEO of Anthropic predicted AIs better than "almost all humans at almost all tasks" sometime around 2027 or 2028. Both CEOs have separately acknowledged that automated research capabilities are liable to spark an intelligence explosion.
25. *CERN:* John Werner, "AI Superpowers & Global Treaties," *Forbes*, February 19, 2025, forbes.com.

CHAPTER 13: SHUT IT DOWN

1. *facilitate, verify, and enforce:* Aaron Scher, “Mechanisms to Verify International Agreements about AI Development,” MIRI Technical Governance Team, December 2, 2024, techgov.intelligence.org.
2. *weapon proliferation:* United Nations Office for Disarmament Affairs, “Treaty on the Non-Proliferation of Nuclear Weapons (NPT),” accessed March 15, 2025, disarmament.unoda.org/wmd/nuclear/npt.
3. *Allied Powers:* Ken Burns, “War Production,” *The War* | PBS, May 21, 2021, pbs.org; “WWII: Mobilization by Country 1937–1945,” Statista, August 9, 2024, statista.com; “WWII: Annual Tank and Self-propelled Gun Production 1939–1945, by Country,” Statista, August 6, 2024, statista.com; “WWII: Annual Production of Major Naval Vessels 1939–1945, by Country,” Statista, August 6, 2024, statista.com.
4. *$341 billion:* Kenny Chmielewski, “Casualties of World War II | Statistics, by Country, & Total,” *Encyclopaedia Britannica*, August 31, 2023, britannica .com.
5. *banning deepfakes:* Ben Cumming, “US House of Representatives Call for Legal Liability on Deepfakes,” Future of Life Institute, October 16, 2024, futureoflife.org; “Ban Deepfakes,” n.d., bandeepfakes.org.
6. *They downplay:* Some exceptions to this pattern include organizations calling for a moratorium, such as ControlAI, PauseAI, and StopAI.
7. *submit annual reports:* “SB-1047 Safe and Secure Innovation for Frontier Artificial Intelligence Models Act,” accessed March 15, 2025, leginfo .legislature.ca.gov.

CHAPTER 14: WHERE THERE'S LIFE, THERE'S HOPE

1. *the United Kingdom:* ControlAI, "Campaign Statement," February 6, 2025, controlai.com/statement.

   Sunak is not the only UK politician who is taking note. As of early 2025, dozens of UK parliamentarians have signed a statement saying "Superintelligent AI systems would compromise national and global security."

2. *Xi Jinping:* "Global AI Governance Initiative—The Third Belt and Road Forum for International Cooperation," *People's Daily Online*, n.d., beltandroadforum.org.

3. *these signals and others:* "Japan PM Vows to Lead Setting Up Int'l AI Rules through New Framework," *Kyodo News+*, May 3, 2024, english .kyodonews.net; Alexandra Alper, "UN Adopts First Global Artificial Intelligence Resolution," *Reuters*, March 21, 2024, reuters.com; John T. Bennett, "Biden Warns about AI's 'Risks,' Forces of 'Retreat' in Final UN Address," *Roll Call*, September 24, 2024, rollcall.com; "Chinese Vice Premier Ding Xuexiang at World Economic Forum," C-SPAN, January 21, 2025, 38:01, c-span.org.

   In 2023, the G7 countries launched the Hiroshima AI Process, the world's first international framework on AI. In a May 2024 followup, Japanese prime minister Fumio Kishida said:

   > *Let us collaborate as nations united by a common purpose to address the universal opportunities and risks brought about by AI, and work toward achieving safe, secure and trustworthy AI.*

   The first (nonbinding) UN resolution passed in March 2024. U.S. ambassador to the United Nations Linda Thomas-Greenfield praised it:

   > *Today, all 193 members of the United Nations General Assembly have spoken in one voice, and together, chosen to govern artificial intelligence rather than let it govern us.*

   In September 2024, President Biden spoke of international cooperation on AI in his final address to the United Nations:

   > *First, how do we as an international community govern AI as countries and companies race to uncertain frontiers? We need an equally urgent effort to ensure AI safety, security and trustworthiness.*

At the World Economic Forum in January 2025, Chinese vice premier Ding Xuexiang said:

If we allow this reckless competition among countries to continue, then we will see a 'gray rhino'—what to do about it? I think we need to review the history. For example, our lessons in managing risks: nuclear risks, biological risks, and security.[…] We stand ready, under the framework of the United Nations and its core, to actively participate in including all the relevant international organizations and all countries to discuss the formulation of robust rules to ensure that AI technology will become an "Ali Baba's treasure cave" instead of a "Pandora's Box."

A "gray rhino" is a type of risk contrasted to a "black swan." A black swan is a low-probability, high-impact event. A gray rhino is a high-probability high-impact event that people nevertheless have a tendency to ignore or downplay.

4. *2025 poll:* Billy Perrigo, "Exclusive: The British Public Wants Stricter AI Rules Than Its Government Does," *TIME*, February 6, 2025, time .com.
5. *2023 poll:* "Overwhelming Majority of Voters Believe Tech Companies Should Be Liable for Harm Caused by AI Models, Favor Reducing AI Proliferation and Law Requiring Political Ad Disclose Use of AI," *AI Policy Institute* (blog), September 23, 2023, theaipi.org.

   See the toplines and crosstabs linked in the "About the Poll" section.

Praise for
If Anyone Builds It, Everyone Dies

"*If Anyone Builds It, Everyone Dies* makes a compelling case that superhuman AI would almost certainly lead to global human annihilation. Governments around the world must recognize the risks and take collective and effective action."

—Jon Wolfsthal, former special assistant to the president for national security affairs

"Yudkowsky and Soares lay out, in plain and easy-to-follow terms, why our current path toward ever-more-powerful AIs is extremely dangerous."

—Emmett Shear, former interim CEO of OpenAI

"Essential reading for policymakers, journalists, researchers, and the general public. A masterfully written and groundbreaking text, *If Anyone Builds It, Everyone Dies* provides an important starting point for discussing AI at all levels."

—Bart Selman, professor of computer science, Cornell University

"While I'm skeptical that the current trajectory of AI development will lead to human extinction, I acknowledge that this view may reflect a failure of imagination on my part. Given AI's exponential pace of change, there's no better time to take prudent steps to guard against worst-case outcomes. The authors offer important proposals for global guardrails and risk mitigation that deserve serious consideration."

—Lieutenant General John N.T. "Jack" Shanahan (USAF, Ret.), inaugural director, Department of Defense Joint AI Center

"*If Anyone Builds It, Everyone Dies* isn't just a wake-up call; it's a fire alarm ringing with clarity and urgency. Yudkowsky and Soares pull no punches: unchecked superhuman AI poses an existential threat. It's a sobering reminder that humanity's future depends on what we do right now."

—Mark Ruffalo

"A serious book in every respect. In Yudkowsky and Soares's chilling analysis, a super-empowered AI will have no need for humanity and ample capacity to eliminate us. *If Anyone Builds It, Everyone Dies* is an eloquent and urgent plea for us to step back from the brink of self-annihilation."

—Fiona Hill, former senior director, White House National Security Council

"A clearly written and compelling account of the existential risks that highly advanced AI could pose to humanity. Recommended."

—Ben Bernanke, Nobel laureate and former chairman of the Federal Reserve

"You're likely to close this book fully convinced that governments need to shift immediately to a more cautious approach to AI, an approach more respectful of the civilization-changing enormity of what's being created. I'd like everyone on Earth who cares about the future to read this book and debate its ideas."

—Scott Aaronson, Schlumberger Centennial Chair of Computer Science, University of Texas at Austin

"An incredibly serious issue that merits—really demands—our attention. You don't have to agree with the prediction or prescriptions in this book, nor do you have to be tech or AI savvy, to find it fascinating, accessible, and thought-provoking."

—Suzanne Spaulding, former undersecretary, Department of Homeland Security

"The most important book I've read in years: I want to bring it to every political and corporate leader in the world and stand over them until they've read it. Yudkowsky and Soares sound a loud trumpet call to humanity to awaken us as we sleepwalk into disaster."

—Stephen Fry

"The most important book of the decade."

—Max Tegmark, professor of physics, MIT

"Claims about the risks of AI are often dismissed as advertising, intended to sell more gadgets. It would be comforting if true, but this book disproves that theory. Yudkowsky and Soares are not from the AI industry, and they've been writing about these risks since before AI existed in its present form. Read their disturbing book and tell us what they get wrong."

—Huw Price, Bertrand Russell Professor Emeritus of Philosophy, Trinity College, Cambridge, UK

"Everyone should read this book. I'm 70 percent confident that you—yes, you reading this right now—will one day grudgingly admit that we all should have listened to Yudkowsky and Soares when we still had the chance."

—Daniel Kokotajlo, OpenAI whistleblower and executive director, AI Futures Project

"*If Anyone Builds It, Everyone Dies* may prove to be the most important book of our time. Yudkowsky and Soares believe we are nowhere near ready to make the transition to superintelligence safely, leaving us on the fast track to extinction. Through the use of parables and crystal clear explainers, they convey their reasoning, in an urgent plea for us to save ourselves while we still can."

—Tim Urban, creator, Wait But Why

"A stark and urgent warning delivered with credibility, clarity, and conviction, this provocative book challenges technologists, policymakers, and citizens alike to confront the existential risks of artificial intelligence before it’s too late. Essential reading for anyone who cares about the future.”

—Emma Sky, senior fellow, Yale Jackson School of Global Affairs

“This book offers brilliant insights into history’s most consequential standoff between technological utopia and dystopia. It shows how we can and should prevent superhuman AI from killing us all.”

—George Church, founding core faculty, Wyss Institute at Harvard University

“A sober but highly readable book on the very real risks of AI. Both skeptics and believers need to understand the authors’ arguments and work to ensure that our AI future is more beneficial than harmful.”

—Bruce Schneier, author of A Hacker’s Mind

“This is our warning. Read today. Circulate tomorrow. Demand the guardrails. I’ll keep betting on humanity, but first we must wake up.”

—R.P. Eddy, former director, White House National Security Council

“A compelling introduction to the world’s most important topic. Superhuman AI could be here in a few short years. This book takes the implications seriously and explains, without mincing words, what could be in store.”

—Scott Alexander, creator, Astral Codex Ten

“You will feel actual emotions when you read this book. We are currently living in the last period of history where we are the dominant species. Humans are lucky to have Yudkowsky and Soares in our corner, reminding us not to waste the brief window that we have to make decisions about our future.”

—Grimes

“The best no-nonsense, simple explanation of the AI risk problem I’ve ever read.”

—Yishan Wong, former CEO, Reddit