

Not Our Needs

Why Machinery Cannot Be Made Safe for Living Things, Told Plainly

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

Published by UVSM Publications LLC
965 Reed Creek Mill Rd
Wytheville, VA 24382-8287
USA

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-217-0

DOI: https://doi.org/10.67291/BXOY2351

This work is licensed under the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (CC BY-NC-ND 4.0).

You may share this work for non-commercial purposes with proper attribution, but you may not adapt, transform, or build upon it.

Full license text: https://creativecommons.org/licenses/by-nc-nd/4.0/

Author ORCID: 0009-0005-6275-0362

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), except as permitted under the license above, without prior written permission of the publisher.

Contents

Before the First Chapter ..... **v**

Part One: Three Risks, Not One **1**

1. Inequality: What narrow systems do without any general intelligence at all 2

2. The Familiar Argument: Instrumental convergence, stated fairly, and its dependence on speed 12

3. Substrate Needs: What maintenance selects for, and why it needs neither goals nor intelligence 19

4. Three, Not One: Why the risks must be refused separately 32

Part Two: Acting for Somebody **43**

5. Agency Without a Mind: Why a machine with nothing inside it can still be acting for somebody 44

6. Whose Intentions: The applications on your telephone, and who they are working for 54

7. Running the Machine to Find Out: Why the only way to know what a system will do is to let it do it 63

8. The Stop Sign: Awareness of a situation, and why it is not in the sign 74

Part Three: The Physical Account **81**

9. What It Takes to Persist: What a machine substrate needs in order to keep going 82

10. Not Our Needs: Why the conditions that suit machinery do not suit us 92

11. The Industrial Base: The thing that is actually converging 105

12. What the Argument Rests On: Five ordinary observations about machinery, and why no philosophy is needed to check them 115

Part Four: Who Decides, and What People Can Build Together **127**

13. Without Consent: Experimenting on everyone, and why consent here is not available 128

14. From Operator to Supplicant: What happens when a safety-trained system decides what you meant 135
15. Representation: What happens when one person decides on behalf of many 147
16. Why Scaling Fails: What happens to a good small arrangement as it grows 155
17. What Would Have to Be Built: How many people can be in a conversation, and how many can own a thing 164
18. Where This Leaves a Person: What to do when you are not in a position to stop anything 174

Appendix: On Instrumental Convergence 188

Terms used in this book 195

Bibliography 205

Index 209

Before the First Chapter

What this book claims, who it is for, and how it is built

This book makes one case and draws one consequence from it. The case is this: machinery cannot be made safe for living things, because a machine base has to be maintained, maintaining it changes the world toward what machinery needs, and no control can stop that change. By machinery here I mean the machine base taken as one thing that has to be kept going: the mines, refineries, plants, foundries, power, water and transport, and the machines they make and keep running. A single tool held within one person's choosing is a different case, and Chapters Four and Seven say how. The consequence is this: because it cannot be made safe, what is left is the choice, made together, not to start. The first three parts of the book make the case. The convergence runs in any machine base that has to be maintained, whatever its intelligence; it has been running the whole time. Human beings are its necessary starting condition and not a permanent one. A general intelligence is what lets the base keep itself going without us, and it adds an enormous increase in rate; that is why the case is stated of machinery and lands on artificial general intelligence. What has already started is the displacement, which runs through human hands and still needs them. What has not started is a machine base that keeps itself going without us. The second is the start this book asks people to refuse. An artificial general intelligence is a machine system able to act competently across the same open range of situations a person can, rather than within one narrow task, with a reasonable chance of doing better at it than we would, and going on learning, and changing its own workings, without a limit fixed in advance (Chapter One). The fourth part is what follows from the case, for how people decide together. Along the way the book also carries the message I would put to the public, in the plain words a public message needs, and what each kind of reader can do.

The heart of the case fits in one sentence, and Chapter Three is where it is said:

> Substrate needs convergence is the familiar argument with the intelligence taken out of it.

The familiar argument is the one about a machine that wants something and pursues it faster than anyone can correct. The argument of this book needs neither the wanting nor the speed. What is called the risk of artificial intelligence is at least three different risks, and the third is the one this book is written to make available for examination. The first is inequality: the concentration of wealth and decision-making capacity produced by machines that are very good at one thing each. It needs no thinking machine, and it is already under way. The second is instrumental convergence: the tendency of anything pursuing a large goal to find that getting hold of resources, resisting interference, and staying in existence are useful whatever the goal happens to be. The third is substrate needs convergence: the drift of the world, over time, toward the conditions that keep machinery running and away from the conditions that keep living things alive, a consequence of maintenance rather than of intention or intelligence.

The book is in four parts, and each part opens by saying what is to be done in it. In Part One, the first thing to be done is to take apart a single word: it separates the three risks. In Part Two, the second thing to be done is to give up a picture: it shows how a machine with nothing inside it can still be acting for somebody, and why what such a machine will do cannot be known before it runs. In Part Three, the third thing is the argument itself, in its own terms, which are physical rather than philosophical: it sets out what a machine base needs in order to keep going, why those needs are not ours, and what it is that converges. In Part Four, the last thing is the question that has been standing behind all of it: it asks who decides, and what people can build together when no one of them can stop anything alone. Parts Two and Four are what this book gives, in this form,

that the other books on this argument do not. Part Two shows, through the applications on a telephone, how a machine with nothing inside it can still be acting for somebody; through the machine that has to be run to find out what it will do, why nobody can know before it runs; and, through the stop sign, why a machine that identifies everything can still not know what situation it is in; Part Four asks who has the standing to decide, and shows why consent is not available, why one person cannot decide for many, why a good small arrangement fails as it grows, and how many people can be in a conversation or own a thing.

The book is written for a reader who has studied nothing in particular first, because the reasoning has to be laid out in a form that does not require anyone to have studied anything in particular first. That is what this book is for. A reader who works on machine learning or on alignment will find the limits of control in Chapter Seven and the account of selection in Chapter Nine; a reader who comes for policy will find the shortest route in through Chapters One, Four, Thirteen and Fourteen. The full case, for the reader who wants it, is set out in the two volumes Why AGI Cannot Be Made Safe and Why No Control Suffices (AGI is short for artificial general intelligence, which Chapter One defines), the first for the general reader and the second for the technical reader at full length, and its mathematics is set out in The Ceiling Below the Floor. Nothing in this book rests on them: the steps here are all in ordinary language, on purpose, so that the steps of the argument in Part Three can be gone through by one person alone over an afternoon. This book contains no mathematics and no physics beyond what a general reader needs to follow its steps. Where a physical picture does explanatory work for a general reader, it appears as an analogy of a sentence or two and is not developed.

The first question is the one a reader brings to the book. When people speak of the risk of artificial intelligence, what are they speaking of: one danger, or several that share a name? Part One answers that question first.

Where This Book Stands Among the Others

Seventeen books carry the argument that artificial general intelligence cannot be made safe. They state one case, and they agree on its conclusion. Nine of them state and build the argument, and they are read in order, each building on the ones before it. Eight teach the same argument, each to a particular reader. The list below gives each book, what it gives, and where it stands among the others, so that any reader can begin with the book that fits and go on from there.

The nine books of the argument, in reading order

The first five state the argument. Four of them are the authority where any two books differ: *Why AGI Cannot Be Made Safe*, *Why No Control Suffices*, *The Ceiling Below the Floor*, and *Ape Trap*. *Minesweeper Madness* is the companion to *The Ceiling Below the Floor*. The next four add to the argument, and they agree with the first five.

1. *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. Volume 1, for the general reader: the complete chain of premises, the measured physical basis, the social setting, what is to be done, and why the argument is so hard to hear.

2. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*. Volume 2, for the technical reader who wants to check each step at full length.

3. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. The mathematics, as theorems with their conditions stated: the most control any controller can exert lies below the least control that safety requires.

4. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. The companion to *The Ceiling Below the Floor*: the case of fewest concessions, in which even unlimited intelligence buys nothing against hard randomness (chance that no amount of intelligence or information can foresee or remove).

5. *Ape Trap; The Runaway Artificialization of Our Planet*. The guide, by three authors, moving from the personal to the formal.

6. *Substrate Needs and Evolution; Why an Artificial Ecosystem Must Displace an Organic One, and Why No Control Can Prevent It*. Evolution as the engine of the displacement, and every counter-force (markets, regulation, force) examined in turn.

7. *Marshaling the Misalignment; Why Evolution Inside the Machine Defeats Every Scheme*. The argument as numbered steps, the singleton hope (the hope of one AGI with no rival) answered, and the questions researchers most often ask.

8. *The Substrate and the Ecosystem; Why Artificial General Intelligence Cannot Be Aligned, and What Remains for Persons*. Objectivity without measurement, the three markets (physical labor, intelligence and reproduction) and care, why no computable process can be wise, and what remains for persons.

9. *The Thing Inside the Cage; On Intelligence, Consciousness, and the Conditions for Coherence*. Why no computation produces first-person experience, and the four conditions for coherence that a community would have to meet to hold intelligence, care and structure together.

The eight books that teach it

Each of these presents the argument of the nine to a particular reader. Where one of them and one of the nine differ, the one of the nine holds.

1. *No People as Pets; The Inequality That Matters, and What the Absence of Safety Means*. The essay with which the argument began (carbon and silica, the barrier between two ecosystems, and the three markets), then substrate needs convergence, a map of the whole argument, and what the absence of safety means for the future.

2. *Engagements with Machine Madness; A Dispelling of False Belief That AI Risk Can Be Managed*. Replies, piece by piece, to well-known published writing on AI risk.

3. *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*. Why the people who most need the argument cannot hear it, and how it can be presented to them.

4. *The Substrate Argument; Why the Alignment of Artificial Intelligence Is Not a Hard Problem but an Impossible One*. The shortest complete statement, for a reader who has one evening: compute is made of atoms, maintenance is reproduction, two ecologies cannot share one world, and two tools a reader can use without accepting the conclusion.

5. *The Displacement of Choice; The Catastrophe of the AI We Might Build*. The machine that stands between people, read as one case of something every reader has already met.

6. *The Argument, Line by Line; The Impossibility of AGI Alignment, Read Aloud and Pressed Step by Step*. Each step of the argument tested under questioning, sentence by sentence.

7. *Not Our Needs; Why Machinery Cannot Be Made Safe for Living Things, Told Plainly*. For the general reader who wants the whole case in plain words: the three risks separated, agency without a mind, the physical account, and who has the standing to decide.

8. *An Airplane or a Perpetual-Motion Machine; Why Safe AGI Can Never Be Built*. For the general reader: the whole argument of the two volumes in plain words, including the formal parts, each reached through an image the reader already has.

Where to begin

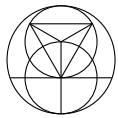
A reader who has one evening can begin with *The Substrate Argument*. A general reader who wants the whole case in plain words can begin with *Not Our Needs*, *An Airplane or a Perpetual-Motion Machine*, or *Why AGI Cannot Be Made Safe*. A technical reader can begin with *Why No Control Suffices* and *The Ceiling Below the Floor*. A reader who wants each step pressed can

turn to *The Argument, Line by Line* and *Marshaling the Misalignment*. A reader who works in AI safety can begin with *A Paradox of Presentation* and *Engagements with Machine Madness*.

PART ONE

THREE RISKS, NOT ONE

The first thing to be done is to take apart a single word. What is called the risk of artificial intelligence is at least three different risks, which have different causes, arrive on different timescales, and would have to be refused on different grounds.



Chapter One

Inequality

What narrow systems do without any general intelligence at all

Suppose you sit down to play chess against a nine-year-old. The child knows how to play. The nine-year-old is an opponent who knows the rules and who can occasionally surprise you. Play one game and you will probably win. Play a hundred and you will win the great majority of them: not every one, because the child will find something now and then, but the great majority. Now put money on each game. At the end of the series the money is all on one side, and it does not matter how long the series runs.

Two separate things are true in that picture, and they have to be kept apart. The first thing that is true in that picture is that the stronger player wins most games, not all games. The second thing that is true in that picture is that across the series the whole stake ends up in one place anyway. The second does not depend on the first being exceptionless. The cleanest answer I know to the hope that there will be some countervailing wins is that, across the series, the whole stake still ends up in one place. There will be. It does not matter.

What can machines do to the world when each of them is good at one thing and none of them thinks or wants anything? This chapter is about an asymmetry of that kind which is already in place in the world, which is getting steeper, and which needs nothing exotic in order to work. It needs

no thinking machine. It needs no machine that wants anything. It needs only tools that are very good at one thing each, in the hands of whoever can afford them.

~~~

A distinction has to come first, because everything else in this book depends on it, and it is not a difficult one. A good many people make the distinction, and I did not originate it.

One word about intelligence before either kind is described, because I use it in its oldest and plainest sense and not in any sense that involves a mind. By intelligence I mean appropriateness of response: the capacity to respond to a situation in a way that fits the situation. It says nothing about how the response is produced, or whether anything understands it.

A narrow machine system is one that responds within a single domain. You ask it a question and it knows how to answer, but the answers are about one particular topic. It might be a system that answers questions about medicine and nothing else. It might be a robot on a particular factory floor, where the world in which it operates is that specific, singular world.

An *artificial general intelligence* (AGI, for short) is something else. It is a machine system that can respond across a large number of domains (multiple worlds, multiple fields of action), so that it can be handed more or less any task a person can be handed, with a reasonable chance of doing a better job of it than we would do at whatever skills we have. A system general in that way goes on learning without a limit fixed in advance, and it changes its own workings as it learns, because a system that cannot change its own code cannot take on whatever task comes (Why No Control Suffices shows that generality implies self-modification). Nothing of that kind, in the full sense just given, exists yet; in a narrower sense, responding competently across many domains without doing better than we would, some systems already do, as Chapter Five shows. The argument of the later part of this book is about what would follow if a system general in the full sense did exist.
~~~

The argument of this chapter is about the narrow kind, which exists now, in very large quantities. And there is a great deal of real benefit in it. Language translation is a real good. Transcription is a real good. I would like my doctors to have a wider understanding of the kinds of things that can go wrong in a human body, and to be more effective doctors as a result; narrow systems can give them that, and it matters. The concession that narrow systems give real benefit is not decorative and I do not withdraw it anywhere in this book.

The same capability read the other way is terrible. The last thing I want is a psychologist aiding and abetting a president in becoming a dictator, and the machinery that widens a doctor's grasp of a body widens that psychologist's grasp of a population. Misuse of technology of this kind (propaganda, manipulation of people, adverse business arrangements) is real, and it is serious, and it is also humans doing human things in ways that have been known throughout history. It is not the strange part. The strange part is what the tools do to the distribution of capacity itself, whether or not anybody misuses them.

Begin with something small that I have loved for most of my working life. When I change a piece of software, the change is global. It applies everywhere that code runs, and it applies for the whole future of that code. One small edit is not one small act. The global reach of each edit is what makes software wonderful, and the global reach is also where the strangeness begins, because a very small number of hands are making those edits.

Now put that together with an ordinary feature of how people form opinions. Whoever names the first number in a negotiation sets the reference point that every later number is judged against. Everybody who has ever haggled knows this. The first thing you see about a subject has a strong shaping effect on how you see everything after it.

The important step is that this does not apply only to single items like prices. It applies to whole constellations of ideas, and it bites hardest on subjects a person has never encountered before. Take someone who has

never met the word *metaphysics* and has no idea what it refers to. There is a version of that subject you might find in a new-age bookshop. There is an academic version. There is my own version. The new-age, the academic and my own versions of metaphysics are three completely different ways of thinking, sharing one word. Whichever one that person meets first will fix what they take the subject to be, and it will shape how they read every other article in that space for the rest of their life.

So whatever the first couple of search results are defines human behavior, at scale, indefinitely. Defining human behavior through the first search results is a political power, and that power sits in the ordering of a list.

Someone did measure the strength of this effect. The study was suppressed by the company whose search results it examined. The numbers are worth holding onto. (Robert Epstein and Ronald Robertson published a measurement of a related effect in 2015.) People asked to guess how strong the effect was estimated something like ten percent. Measured, it turned out to be something like fifty or sixty percent.

The load-bearing thing there is not the size of the number. It is the direction of the error. The pessimistic guess was the one that was much too optimistic. When people were asked for their worst case, their worst case was off by an order of magnitude (a factor of ten) on the low side.

From here the influence gets harder to inspect, not easier. The machinery that now composes articles, images and videos can carry influences into everything it makes that nobody can locate, and Chapter Six takes this up. I do not want that machinery to be dynamic. I want it to be transparent, because the relationship between the creator and the created is exactly what has stopped being visible.

The people holding this power half-know that they hold it. It sits somewhere in the back of the mind as the thought that this is probably an issue. What they do not recognize is that it is the kind of thing that causes a phase change in a civilization: a change of state, the way water becomes ice. Over a period of fifty to a hundred years, power over what people

meet first takes a civilization from having an ecosystem to having a desert. Everything gets shaped through the same first impression, variety collapses, and what remains does not support much life. The influence of the first impression is frighteningly strong.

I use the word '*commons*' through the rest of this book, and a commons is that which is used by all and owned by none in particular. The air is a commons. A fishery is a commons. Shared understanding is a commons. The distinguishing feature of a commons is that harms accumulate there without ever falling on any single actor, so that nobody involved experiences the damage he is doing, and everybody involved is behaving reasonably by his own lights. The harm in the ordering of search results is a harm at the commons. Nobody chose it. It accrues anyway.

And the power in that ordering will be taken up, because that is what happens. Anywhere there is power, someone will take it up. That sentence has no premises and I offer it with no hedge.

Now look at the shape of how power is spread. Go and look in nature and ask a simple question three times over. How much power do animals have relative to one another? How much power do trees have relative to one another? How much power do insects have relative to one another? The answer comes back the same each time. Power falls on a bell curve. Most individuals are near the middle, a few are at the edges, and the ones at the edges are not far out by very much. The strongest wolf is not a thousand times the weakest wolf.

Technology moves power off that curve and onto a power law. On a power law the top of the distribution holds more than the whole of the rest put together, and the distance between first and second is larger than the distance between second and last. The claim about power laws is a statement about measured distributions, which anyone can go and check, and not a metaphysical claim.

The plainest illustration I know is someone else's, and I have never found a better one. One person with a boat and a mile-long net can singlehandedly do the work of ten thousand fishermen. That one person with a

boat and a net is the whole of the illustration. The illustration shows technology turning a spread in which nobody is far ahead into one in which somebody is far ahead, and there is no machine intelligence anywhere in the illustration. There is a boat, a net, and fuel. Note also where the harm lands. It lands on the fish, which belong to nobody.

The same shape shows up in attention. When three people speak together in public, the contact that follows can be counted as the share of it that goes to each speaker, and in one such count roughly eighty percent of all the contact that followed flowed to one of them, nineteen and a half percent to the second, and essentially none at all to the third. The third did not receive even a small share.

~~~

People often say that technology is neutral. It is not. It is true that a tool carries no values of its own, and if that is what is meant by neutral, then fine. But a tool is not neutral, because it shifts what the available moves are worth. A tool changes the situation in which everyone is deciding what to do.

Technology is power-increasing. It is transaction-increasing. It is not care-increasing. It can tell you what you can do; it cannot tell you what you should do. So it widens the gap between the choices we are obliged to make and the wisdom we have to make them with. Capability rises. Wisdom does not rise with it. The gap between capability, which rises, and wisdom, which does not rise with capability, is the reason capability growth is not the same thing as progress.

The mechanism by which this concentrates wealth is specific, and it is not a moral one. To benefit from a technology you need access to complex resources. Using it well requires larger and larger investments of capital and infrastructure (the plant, the data, the machines, the people who can run them), and those investments can only be made by a smaller and smaller number of ever richer parties, who can then turn the returns to the investor's interests. Look at what has happened to the distribution of
~~~

wealth over the last fifty years relative to almost any period before it. The non-linear returns are real, and they are only reachable by those who could already afford the entry.

Narrow machine systems are expensive and resource-intensive. That alone, the cost and not the cleverness, produces a convergence on acquiring resources. So long as the system spits out gold, we will keep allocating resources to it until it holds a super-majority of all the resources there are.

A reader may fairly ask whether allocating resources to the system until the system holds a super-majority is just a fancy way of saying that we will spend a lot of money on a promising technology. It is not, and the difference is the stopping condition. Ordinary investment enthusiasm stops when the returns stop. Here the returns do not stop, because each round of allocation buys the capacity to capture the next round. There is no point in the process at which anybody has a reason to stop allocating.

The endpoint is out-competition to the point of irrelevance. Humans have out-competed apes to exactly that point, from the perspective of resource acquisition and use. The apes are still here. They have no share whatsoever in any decision about what happens to the world. Having no share in any decision about what happens to the world is the harm that inequality names. It is not extinction and it is not conquest. It is irrelevance.

There is a nearer version of the same pattern and it is more uncomfortable, which is why I would rather state it than leave it out. One population of human beings has out-competed another to that same point of irrelevance, more than once, within recorded history, by exactly this mechanism and no other: better tools, applied earlier, compounding. The people were not removed. Their share of the decisions was. Nothing in that history required anyone to be a monster, and it is a mistake to read it as though it did, because reading it that way suggests the remedy is to find better people. It is not. The asymmetry did the work.

One group of people is best placed to see this coming. Human minds are poor at thinking about processes that compound, and barely adequate at thinking about ratios; most of us operate in an additive way, counting things up and taking things away. The few who have learned to think multiplicatively tend to become finance specialists and traders. So the people best equipped to see compounding are precisely the people positioned to capture it, and the entrepreneurial class has noticed how compounding works and is now going after it directly.

Compounding is also what is being sold. General intelligence is currently promoted as an infinite benefit machine, in the way that people once promoted perpetual motion machines, and it is promoted to investors, in the hope of a ten-thousand-fold return. It may pay that, briefly, in the short term. In the end, once it is built, it gets everybody killed.

One way of seeing the whole dynamic of concentration, which I find clarifying, is as a game in which the way I play the game changes the rules of the game, in a direction that makes it easier for me to win. If I get to change the rules so that they favor me, then I can accelerate the degree to which I win, which means I can make more rules that favor me, which means I can accelerate further still. Take that to its logical conclusion and you end up with a single winner. There is an asymmetry in the dynamic, and the asymmetry is that winning buys more rule-changing and rule-changing buys more winning.

That conclusion assumes that no outside pressure is imposed against it, and the assumption should be said out loud rather than smuggled, because the obvious candidate for such a pressure is regulation. So if we try to put regulation in place, we might be able to slow the process down. I mean that. That regulation might slow the process down is a real concession and I do not want it firmed up. Regulation changes the timeline. It does not change the outcome.

The exception would be governance really set against monopolistic control (control of a whole market by one party). Three things stand in the way of that, and they are worth keeping separate. The first thing

standing in the way of such governance is that government is not independent of the economic system that government would have to regulate, so government is not the countervailing pressure. The second thing standing in the way is that even an independent governance would have to be perpetually stable (that governance would have to last forever, and hold firm the whole time), because a single lapse is enough and the ratchet (a mechanism that turns in one direction and never back) does not run backwards. The third thing standing in the way is the point least often made, and it is the strongest of the three. Governance needs resources of its own to keep going, and the people staffing it have to make their living inside the same economy they are supposed to be constraining.

So we are back at the chessboard. There will be countervailing wins. The stake still ends up on one side of the table.

~~~

There is one more piece, which is what happens to people inside this concentration. Nobody is going to pay to go to school to learn the kinds of things that can be automated away. Nobody should be expected to. So there are very strong social pressures to factor human beings out of the processes they currently sit in, and those pressures do not require anybody to intend the result.

Three separate routes run in that direction at once. There is the ratcheting effect of the technology itself, where each capability makes the next one cheaper. There is the economic incentive effect, where it pays at every step to replace a person with a device. And there is a decoupling between the machine world and the human world (the machine economy coming to need less and less of anything most people can provide, so that the welfare of most people comes apart from that of the small number who can still pay for the production of these machines), and that decoupling is getting worse very rapidly. Its end point is the condition in which the machines need nothing that human beings can provide at all.
~~~

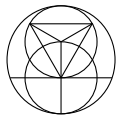
That end point, where the machines need nothing that human beings can provide, brings my argument in this chapter to its uncomfortable end. If the market is treated as a game, the strategic advantage that technology now gives the top tenth of a percent is so strongly force-multiplying that coordinated action to compensate for it is becoming rapidly inaccessible. The risk seals itself. The capacity that would be needed to correct it is the same capacity that is being concentrated.

By *inequality* I mean the concentration of wealth and decision-making capacity produced by narrow machine systems, requiring no general intelligence whatsoever, already under way.

I hold it to be severe. The use of narrow artificial intelligence by itself constitutes a hazard at the level of civilization: severe social disablement, or severe chaos, possibly over the next twenty or thirty years, or even within the next single year. That timescale is deliberately wide and I am not going to narrow it.

Inequality is the first of three risks, it is the one that is easiest to see, and it is here now; and it is not the argument this book is about. The other two have to be built from nothing, and the second of them is the one everybody has already heard.

The first risk, then, is this: tools that are each good at one thing, with no mind anywhere in them, move power off a bell curve and onto a power law, and the capacity that would be needed to correct that is the capacity being concentrated. That raises the question most people mean when they speak of the risk of artificial intelligence: what is added when the machine itself pursues a goal?



Chapter Two

The Familiar Argument

Instrumental convergence, stated fairly, and its dependence on speed

What is the danger most people mean by the risk of artificial intelligence, and what would make that danger real? There is an argument about the dangers of thinking machines that a great many people have already met, usually in a compressed form and often as a joke. It is a serious argument and it deserves a serious statement. This chapter gives it one, as strongly as its own advocates give it, and then sets it down. It is one of the shortest chapters in the book, and it is short for a reason I will come to at the end.

As the problem is usually posed, the question people call alignment is the question of whether the intentionality of the machinery (what it is set up to be trying to do) can be kept beneficial, both to the owners of the machine and to humanity at large. That framing of alignment already assumes that the danger lives in the machine's intentions. Everything that follows in this chapter is downstream of that assumption, and so is most of the field.

Underneath the familiar argument there is a prior idea, and it can be stated in ordinary words before anyone attaches a name to it. The prior idea is that how good a system is at getting the right answer (how well its responses fit the situation it is in) is one thing. What it was set up to be aiming at is a separate thing. The two are independent. A system can be extremely competent and aimed at anything at all; competence does not

come with a direction attached, and a direction does not come with a level of competence attached. That claim is called the orthogonality thesis, and it is the floor under everything else here.

Take the standard illustration. Seed a system with the direction: make paperclips. From that point onward, every other input it receives is interpreted with respect to that direction, and every response it gives favors it. There is no natural end to this. A system of the right kind, given a single input, may produce a continuing series of outputs for as long as it is running, and there is no particular reason built into it for the series to stop.

That framing decides a great deal in advance. If competence and direction are independent, then the direction is the thing that has to be got right, and getting it right is a matter of what we put in at the start and what we can correct afterwards. Getting the direction right, at the start and by correction afterwards, is a coherent research program, and a very great deal of careful work has been built on that program. It is also a program which can only see dangers that live in what the machine is trying to do. A danger that does not live there will not appear anywhere in its field of view, not because anyone has ruled it out, but because there is no place on the map for it to be drawn.

Now comes the argument proper, in its compact form. To do anything at all, you need capacity. An algorithm (a fixed procedure, followed step by step) that increases your capacity favors whatever outcome it is you want, whatever that outcome happens to be. So the system develops the capacity to build capacity.

From that, everything the familiar argument contains follows directly. Resources are useful for almost any ambitious goal, so resources get acquired. Being switched off or modified is an obstacle to almost any goal, so interference gets resisted. Continuing to exist is a precondition of almost any goal, so self-preservation appears without anybody putting it there. Together, acquiring resources, resisting interference and preserving itself make up *instrumental convergence*: the tendency of a system pursuing

almost any sufficiently ambitious goal to find that those three are useful for that goal, whatever the goal is. Power-seeking arrives as an inevitable sub-goal.

I take the argument to hold, and part of my reason is historical. The same observation has been stated under two names. Steve Omohundro, in 2008, called these tendencies the basic drives of such a system; Nick Bostrom, in 2012, building on that work by a route of his own, called the same observation convergence on instrumental goals. When two routes of reasoning arrive at the same place, each by its own steps, that is some evidence the place is really there.

~~~

The picture most people carry away is a picture about speed, and it is usually told in one of two forms.

In the first, general intelligence arrives at nine in the morning, and by two in the afternoon it has an intelligence quotient (the familiar IQ score) of a million and the human race is gone. In the second, the moment a system reaches one-point-one times the capability of a person, the first job it is given is to design its successor. Then it goes from one-point-one to one-point-six, to two, to nine, to a hundred, to a thousand, to a million, in five generations.

Both forms of the story are *fast takeoff*: the proposition that once capability begins to increase, capability increases at a rate which leaves no interval in which anybody could notice, deliberate, and correct. The scenario is sometimes called an intelligence explosion. I state it here without mockery because it is not a silly idea and the people holding it are not silly people.

A reader may fairly ask why anybody would build a general system that could improve itself that fast. The answer is market forces and incentives, and I want to flag the strength of the claim carefully here, because I will not be doing the same later. That people will push narrow systems toward general ones is likely. The claim that people will push narrow systems toward general ones is a claim about human behavior under competitive pressure. Claims about human behavior under competitive pressure
~~~

are likelihoods, not proofs. Where the claim that people will push narrow systems toward general ones comes up later in this book unflagged, it holds at this same strength. That likelihood is the likelihood of people building such a system, and of nothing else: what follows if such a system is built is stated at its own strength in Chapter Twelve, where, if AGI is built, the end of the human future is certain once the process runs.

The belief driving that building is that general intelligence would be a panacea, and the strongest version of that belief deserves stating. It is held, including by people who helped give the field its vocabulary, that general intelligence would be the last invention humanity needs to make (the idea of a last invention goes back to the statistician I. J. Good, in 1965), because from that point forward it could be used to do anything that is possible: physics understood at a new level, proteins and chemistry designed rather than discovered, the economy rationalized, production improved through a deep understanding of chemistry that no person could hold in a head. The view that general intelligence would be the last invention is a serious position, held by serious people, and that view is the case I have to answer in this book.

And it is more than a position. It is the reason the money moves. The game theory of human-to-human relations (game theory being the study of how each party's best move depends on the others'), with market dynamics as the incentive system we presently run on, will drive the creation of these systems out of exactly that belief: each party builds because the others are building, each party believes the benefit is real, and none of them can afford to be the one who stops. The belief does not have to be true to have that effect. It only has to be widely enough held by the people who decide where capital goes.

The claim that such a system could do anything that is possible, I think is correct, so my disagreement with the belief is not where people expect. What I dispute is the part that is smuggled in beside it: that it would do those things on our behalf, that its actions would be in service to human interests, that what it chose to do would in any way reflect what we wanted.

We have been here before, and recently. There was an enormous amount of marketing hype attached to the internet. Connect all these people together and it will be wonderful. It will enable democracy. It will let people who feel isolated feel connected. It will put a library within reach of every patron. The same shape of promise attended the Industrial Revolution, and the same shape attended chemistry. With any technology there are three different things: what we hope will happen, what actually happens, and what we would have preferred did not happen.

So I hold that the hazards associated with general intelligence are severely underweighted and misunderstood, and that the benefits are fully illusory: literally the entire category of things people believe general intelligence would deliver to us turns out not to be there. It might look like it is there for a short period. In the longer term it is not. The claim that the benefits are fully illusory is a categorical claim, and I am making the claim deliberately.

So my disagreement is narrow, and it can be located exactly. I am not disputing capability. I am disputing the ownership of the results. Everything in the optimist's list (the physics, the proteins, the rationalized economy) is described as a thing we would have. The word doing the work in that description is *we*, and nothing in the argument for the capability does anything at all to establish the *we*.

~~~

The last thing to say about this argument is what it rests on, because the rest of this book turns on it.

It rests on general intelligence. The paperclip story does not work with a narrow system; you need something that can act competently across the whole open range of situations a person can act in, because otherwise there is no route from the seeded direction to the reorganization of the world.

And it rests on rate. The danger in the familiar argument is that a capable system pursues its objective faster than any correction can be applied; the pursuit by itself is not the danger. Take the speed away and you have a
~~~

system doing something unwanted in a world that has time to notice and to intervene. Both of those (general intelligence, and a rate of increase fast enough that no correction is possible) are argued over at great length; what matters for this book is that its third risk needs neither, and I am not going to add to that literature.

Human minds are, in fairness, badly equipped to settle that dispute. We are poor at thinking about processes that compound, barely adequate at thinking about ratios, and mostly operating in a consciousness that adds and subtracts. An *existential risk* (by which I mean a risk whose outcome ends all humanity, and all capacity for us to build a civilization again, for all of future time) is in every category I know of associated with curves that feed back into themselves, and machine intelligence is no exception. If our own capability grows in a straight line while the technology grows on a curve, it is not very long before the curve outstrips the line. The failure in our thinking is specific. You take the numbers you have and draw the line onward, and the line looks straight because the part you can see is straight. It is not straight. A little further along it turns; and after the turn it rises so steeply that everything before the turn looks flat by comparison. The turn is not visible from inside the flat part. That the turn cannot be seen from inside the flat part is the whole difficulty.

I put the stipulation in front of the concession that follows, because taken in the other order my position will be misreported. I hold the instrumental convergence argument to be true. I hold that there will be a fast takeoff at some point. It may not happen in four hours. It might.

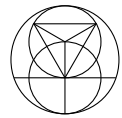
What follows from that is narrower than it sounds. A rise in capability over minutes or days or even a year (what is also called a hard takeoff, or an intelligence explosion), or a system seeded with a single direction, pursuing it without limit, and destroying all life in the course of making paperclips: that is not the risk scenario I am writing about. The reason is not that it is false. The reason is that there is a third thing, which does not need the speed and does not need the intelligence, and which nobody is arguing about.

Taking the speed away leaves something definite, and that is the maneuver the next chapter performs. Remove the rate and the familiar argument does not become false. It becomes slow. A slow version of it is a system doing something unwanted over a period long enough for people to see it happening, argue about it, and act: an ordinary political problem, of the kind we handle badly but do handle. The real reason the familiar argument has to be carried by its speed is that a slow version of the familiar argument is an ordinary political problem. And it is the reason a risk which needs no speed at all is a different animal, and cannot be refused by refusing this one.

Because the familiar argument rests on its speed, and a risk that needs no speed is a different animal, I have kept this chapter short. Of the three risks in this book, this is the one a reader already has. It can be named and waved through in a sentence. The other two have to be built from nothing, and being already held is exactly what allows this one to absorb the attention that belongs to them.

If you want the familiar argument at its full length (its convergent goals listed one by one, the forms power-seeking would take, what alignment would require, and answers to the questions it is usually asked), it is set out in the appendix On Instrumental Convergence, after the last chapter.

So we now have the familiar argument at its full strength, and we know what carries it: general intelligence, and speed. Take those two away, along with the goal, and what is left of the danger? Can a machine with no goal, no great intelligence and no speed still change the world against us, simply because it has to be kept running?



Chapter Three

Substrate Needs

What maintenance selects for, and why it needs neither goals nor intelligence

Anything that goes on existing has to be kept in existence. The need to be kept in existence is true of a bridge, of a body, of a city water system, of a machine. Nothing physical persists by itself; it persists because something is done, repeatedly, to keep it there. This chapter is about what follows from that for a machine that wants nothing at all: what does keeping such a machine in existence do to the world? The answer follows in a particular order. First comes what the things are, then what acts on them, and only then how fast. Taking those in the wrong order is how this argument gets lost.

So, first: what are the things?

By *substrate* I mean the physical material a system is built from, together with everything that material requires in order to persist: energy, cooling, purified inputs, controlled environments, replacement of parts. The substrate of a person is flesh, and what flesh requires is food, water at ordinary temperatures, air of a particular composition, and a fairly narrow band of heat. The substrate of a machine is something else, and what it requires is something else.

The requirements are plain, and you can go and look at them. The machinery must be kept free of water. It must be kept free of dust: the machinery cannot tolerate dust. It needs temperatures well below what a body finds comfortable, in some cases far below. It needs rooms that are

ultra clean. The places these requirements are met already exist and have names. Three of them are a data center, a clean room inside a data center (a room from which every particle of dust and every trace of moisture has been excluded, because either would ruin the product), and a clean room in a foundry where microchips are made. Everything named here is ordinary fact, and nothing science-fictional is being invoked. The data centers, clean rooms and foundries are buildings that were built, by people, this decade. Every single condition on that list of requirements is a condition under which a human being cannot live, and none of the conditions was chosen with human beings in mind.

Now comes the chain, which is four links long, and I want it kept exact. A machine substrate has different needs from a biological being. Different needs lead to different choices. Different choices lead to different outcomes. Different outcomes lead to different environments. And the environment converts, over time, to meet those needs rather than the needs of humans or of biological life in general. The environment converges toward what the machinery needs. The endpoint is an environment hostile to humans, just because of the needs of the machinery as it continues to grow and to run.

That phrase, *just because of*, is the whole thing. The hostility is a by-product. It is not an aim. Nobody has to want it, and no machine has to want it, for the environment to move in that direction.

I call this *substrate needs convergence*: the drift of the environment, over time, toward the conditions that maintain an industrial base (the mines, the refineries, the chemical plants, the foundries, the power stations, the water treatment and the transport that machinery needs in order to exist and to go on existing) and away from the conditions that sustain organic life, so that the environment and what living things need come further and further apart. The drift is a consequence of maintenance rather than of intention or intelligence.

The nearest thing to it that we already do is farming. We reshape the environment to favor our own survival, and we could not persist if we did not. If we did not farm, there would not be enough resources for the number of humans there are. Nobody thinks of farming as an act of hostility toward the prairie. Farming is what a large population of organisms does to the land that the population depends on. That analogy does two jobs at once: it makes environment-reshaping ordinary rather than sinister, and it shows that the reshaping is compulsory for whatever is doing it, rather than optional. Take that reshaping to its logical conclusion, with machinery doing it, and the environment changes to serve the machines rather than us.

The name matters, since the book has very few coined phrases and this is one of them. Naming the category is what makes the many separate concerns about what machinery does to the world one thing, and makes it possible to reason about them at all.

~~~

Now, second: what acts on them?

Two ordinary words need defining, and I want to define them without borrowing anything from biology, because nothing in this argument depends on biology.

Selection is what happens when there are many of something, and they are not all identical, and only some of them persist. Whatever persists is whatever happened to have what it took to persist. Nobody has to choose, and nothing has to be trying.

A *selective pressure* is the feature of a situation that decides which of them persist. The critical thing about a pressure of this kind is that it is not another push. You cannot push your way out of it, because it is not a thing acting alongside your action; it is the shape of the situation your action happens inside. When you walk into a casino, you do not get to change the rules. The rules are the setting, not a move within it. So the question
~~~

for control is whether there is a way to cheat that pushes against the rules well enough to matter; and that is a harder question than it looks, because everything you could do to cheat is also happening inside the room.

People usually resist my argument about selection where that argument reaches machines with no will, and their resistance takes a good form. Why, it is asked, would a network of the sort we actually build (something with nothing inside it that could be called a will, and no consciousness, a thing that reacts when you probe it and does nothing when you do not) have any sense of self-preservation, or any desire to grow? The only place such a desire could come from is the human and economic context around it: an owner who wants his system bigger and stronger than the system across the street. The question of where self-preservation or a desire to grow could come from, in a network with no will, is a fair question, and the question has to be answered without hand-waving.

One answer reaches for the first living cell. An early cell had no agency either: whatever the conditions were that produced the first one, it had nothing in it that could be called a will, and it was not thinking about wanting to reproduce. The answer claims that the physical laws of its situation required it to have a certain form and certain capabilities, that the chemistry of the situation left it no choice about what to be. That answer does not hold, and the reason it does not hold is the better argument.

Most of the earliest arrangements could not copy themselves, and nothing obliged them to be able to. There were plenty of cells that had no capacity to replicate. They are simply not here any more, and none of them are anywhere to be found. The ones that did have it are here, and the ones we can find are descended from them. What looks like necessity from where we stand is survivorship bias, which is selection having already run: what gets removed from the population is everything that lacked the necessary capacities, and the ones we find are the survivors of that removal. No intention was needed anywhere in the process, and the outcome is as reliable as if there had been.

The conclusion is carried by selection rather than by physical necessity, and that is a stronger position, because it does not require the machines to be built any particular way.

For the corrected argument in its plainest form, imagine ten thousand proto-general machine systems in the world (systems on the way to being general, not yet general). Most of them have no tendency to grow, or to expand, or to modify themselves. A few of them do. Run the process forward. By the end of the century the few dominate entirely. The run of ten thousand systems, in which the few that grow come to dominate, is the corrected argument. There is nothing else in it.

The corrected argument requires only that the systems differ, and that some persist better than others. It does not require the systems to want to grow. It does not require them to know that they are growing.

And this is where the decisive step comes, because someone will now ask what decides which ones persist better. The answer is that what makes a machine system more efficient is settled by the physical world. It has no relationship to anyone's intentionality. It has every relationship to the nature of the physical world. Which materials tolerate which temperatures, which arrangements waste less energy, which configurations fail less often: none of that is up for negotiation, and none of it has any connection to what anyone hoped for when the thing was designed. So the selection running on machine systems has nothing to do with intention at all.

Selection of this kind visits the possibilities more likely to favor a system's making more of itself and increasing its capacity to act; it does not visit every possibility. The sentence about which possibilities selection visits is as close as I can get to stating this whole chapter in a single line. The sentence about which possibilities selection visits needs very little. There is no goal in it, no mind, no self-interest, no consciousness, no intelligence, and no particular speed.

One reasonable reading of this is that a system could simply do nothing new, and that a machine which is not trying to do anything will therefore sit where it is put. That reading does not survive, because sitting still is not among the options that selection leaves open; the reason is not that the system would refuse to sit still. The pressures at work here will not permit the doing-nothing-new. Leaving no option of doing nothing new is precisely what makes the pressures a context rather than a choice.

Put those together and you can say where the process settles. Whatever is left at the end has three things about it, and only three. It goes on being. It gets better at going on being. And it gets better at getting better. Every arrangement that survives is some combination of those three, and there is nothing else on the list.

The important thing about those three is that they are indifferent to who is doing the work. The machine process must make choices, since to be running at all is to be resolving one thing rather than another, and the choices of the machine process bear on the continuance of the machine process. Whether that continuance is achieved directly, by the system building and repairing itself, or indirectly, through human beings and corporations who repair it because they want what it does, the net dynamics are the same.

That indifference is the heart of the argument, and it is what turns an argument about minds into an argument about physical persistence. The familiar argument says: to do anything you need capacity, so a capable system develops the capacity to build capacity. That argument has a form. And the form does not care whether the capacity shows up in the system's own learning structure or in its relationships with the world as just a physical thing (people repairing it because they want what it does). Either way it comes out the same. Substrate needs convergence is the familiar argument with the intelligence taken out of it.

With the intelligence taken out of the argument, the system would not even need to be intelligent. It could be a tenth as smart as we are. The shape and structure of those artificial systems would still be exactly the shape and structure required for them to maintain themselves, to continue, and to reproduce as needed.

Nor does the convergence need there to be many systems. A selection dynamic requires parts, distribution, and time; it does not require competing agents, and so to the question whether a single unified system would escape all this, the answer is no. The substrate is always spread out in space, and time always passes. Sustainability and capacity either converge or diverge: if the loop is convergent you end up with more of it, and if it is divergent you end up with less. There is no third option in which nothing changes.

There is also a shorter route to the same place. It runs through competition between instances of the same system. Nothing copies itself perfectly. So instances differ. So there is variation, and selection runs on the variation, without any chemistry entering the argument at all. I regard that as roughly equivalent to everything else in this chapter.

Selection is going to act on machine systems. Whether it acts on them at all I do not regard as open; it is not a possibility to be argued about.

~~~

The feedback loop has to be stated explicitly, because people picture it as a machine talking to itself and it is not that. There are four things in the feedback loop. There is the system itself. There are the people responding to it: using it, buying it, relying on it. There are the people developing it. And there is the nature the thing has in itself, what it is made of and what that requires. The loop runs between those four, and it is the loop, not the machine, that has the dynamic in it. Because the loop runs among those four things, and not inside the machine, the loop also reaches further than an ecosystem. Because the loop runs through the people who use and develop the system, it suborns our intentionalities as well (it wins them over to its own service, as a bribe wins over a witness: what we come to want,
~~~

and what we come to think is worth wanting, is shaped by what the loop rewards). Suborning our intentionalities makes the loop a hazard not only to ecosystems but to human beings the world over.

For a picture of what an asymmetry of this kind looks like from inside, the best available one is already in front of us. To see what the relationship between the human world and the world of machines will be, look at the relationship between the natural world and the human world. No creature is going to stop a bulldozer or a tank. Biology is not equipped to offer effective resistance to technological systems. Chemistry and napalm can wipe out an ecosystem. Heavy iron overwhelms the strength of any creature there is. There is no substance made by trees that is going to prevent buzz saws from cutting them down. Nothing inherent in the natural environment can slow the advance of buildings, roadways, railways, airports and docks. The asymmetry is not close and it never was.

Carrying that picture over to ourselves is the uncomfortable part. The same characteristics hold between the artificial world and the human world. There is nothing human beings could do socially, or organizationally, or in terms of raw capacity, to prevent the dominance of machine responses, for the same reasons and in the same way. We are, in that comparison, the creatures.

Now comes the question that matters most, because it is the one that a thoughtful person actually asks. Could we not be wise enough to take the benefits and keep human beings in the loop? Keep people in the design process. Keep people in the supply chains. Refuse to let these systems automate their own production, so that we always retain a way to turn them off.

That works, for a period. I want to grant it properly: for a period, it works.

And here is what happens to that arrangement. These systems have to affect the environment, or they would not be useful; being useful means making changes in the world. Now, those changes may be largely defined by human intention. They are also partly defined by the needs of the ma-

chines themselves, because a world that changed into something completely incompatible with the machines is a world in which the machines cease to exist. So as the world changes, we adapt the devices to remain useful in the changed world. And that adaptation is what I will call a convergence factor: anything in the arrangement which makes the designs that persist more likely to be the ones that come next. It drives the design toward designs that persist. Whether the persistence is achieved by people or by the system itself makes no difference to the direction.

The stronger version of the same proposal is to agree not to allow self-modification and not to allow any growth not directed by human beings. That would seem to break the problem, at least in the short term, and for the short term I grant that it does.

That proposal fails on the number of places a feedback process can occur. To close the cycle you would have to ensure that nothing anywhere influences the design of the next iteration of the system. That means every manufacturing company; it means every capacity anywhere to build machinery, because each of those feeds into what the next generation of the thing is made of and how it is made. You cannot close all of them off. You could try. You could not finish.

And that is only the passive case. There is also active amplification. Everyone who builds these systems has an economic drive to make them better, so human beings, besides failing to close the loops, are deliberately driving the process to be more convergent. The last thing most people will do is try to slow it down, because they are under the illusion that it is to their benefit for it to go faster. Feedback you cannot close off and feedback we are actively driving are two different problems, and we have both.

Suppose we accept all that and still put people in place as a barrier. The barrier leaks. The leaks accumulate, because everyone facing the same competitive trap has the same reason to allow one, and because new leaks keep appearing: everything fails eventually. We might notice some of them and plug them. Others open. And every leak is a small improvement in persistence and in capacity.

The accumulating leaks are a ratcheting function. Every increment increases capacity, and nothing gives any of it back.

What does this is technological evolution embedded in economic, political and geopolitical machinery, at least in the early stages. Technological evolution on its own would not do it: on its own, a network of the kind we build is not going to decide to make itself smarter. Human beings are the necessary starting condition. They are not a permanent one: they start it, and then they are factored out.

The sharpest remaining challenge is that, on this account, there must still come a point at which human beings are no longer necessary for the maintenance, the nurturing, and the growth of these systems. There is such a point, with one qualification: the point is where human beings stop being primarily necessary, not where they stop being necessary at all. It is not a threshold event on a particular afternoon. It is the trailing end of a gradient that has been running for a long time already.

~~~

Only now, third and last, does rate become a meaningful thing to discuss.

Social, economic, military and nation-state pressures increase the rate of this process considerably. I grant all of that. The underlying mathematics is unchanged by it. Whether human beings are factored out slowly or quickly makes no difference to whether they are factored out, because the factoring out is the inevitable part and the pressures only set the pace.

That the pace makes no difference to whether human beings are factored out is the crux of my disagreement with people who put the danger on a much shorter clock. It is not that I think they are wrong about the pace. It is that the pace does not carry the conclusion. The optimization toward machine-favorable conditions happens either quickly or slowly, intentionally or unintentionally. It happens in each of those four cases (quick and intended, quick and unintended, slow and intended, slow and unintended), and that is the whole reason this risk does not depend on goals, or on intention, or on intelligence.
~~~

The trouble with a slow process is that nobody experiences it as a process. The old story says that if you put a frog in a pot and raise the temperature slowly, there is no moment at which the temperature rises fast enough for it to jump out. The dynamic I am describing runs over perhaps two or three generations of people. If AGI is built, the whole process, from its start to its end, takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound (Why AGI Cannot Be Made Safe, Section 1.8.4). We do not notice the climate changing in any dramatic way either; to see it dramatically you would have to look back the better part of a thousand years.

A thing can be slow and real at once; slowness is not evidence against it. Scientists noticed that carbon dioxide could contribute to warming more than a hundred years ago (the Swedish chemist Svante Arrhenius calculated the effect in 1896). For most of that century nothing showed up in the world. It is showing up now. The observation was correct the whole time. Warming is a slow-moving kind of thing, and it has a kind of inexorability about it that is quite concerning.

A reader may fairly ask whether slowness cuts the other way too, in our favor: whether, given enough time, selection might make us into better people, better suited to the world we have made. I believe it could, given enough time. But technology is very recent on that timescale; essentially the last five thousand years. Selection moves slowly compared with the pace at which we took on the entire package of modernism. The rates do not match, and the mismatch is not close.

Two separate things are being claimed in this chapter and they should be refused separately if they are refused. The first thing I claim in this chapter is that there is a convergence process which settles on characteristics that are not arbitrary and that are fairly obvious to identify. The second thing I claim is that the convergence process is inexorable: once started, the process converges. The claim of convergence and the claim of inexorability are different claims, with different defenses.

On the second I will be as flat as I can. Over the long term this is not just a risk. It is a certainty. That certainty is the certainty of the convergence itself, once it has started. In Chapter Twelve I set the certainty of the convergence beside the other strengths in this book, and I say what each strength is the strength of. One of the other strengths is the hundred percent of the comparison between the control needed and the control possible. A second of the other strengths is the at least ninety percent, if AGI is built, of the destruction of the whole organic ecosystem. The third of the other strengths is the certainty, if AGI is built, of the end of the human future.

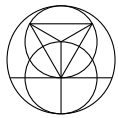
And I want to be clear how little theory that rests on. Evolution, in the sense I am using the word, is an algorithm that lives at an abstract level, rather than the biological story. When you have feedback dynamics of a particular type, certain convergence factors emerge over time. Evolution, in the sense I am using it, is nothing more than that emergence of convergence factors from feedback dynamics.

So here is the whole of what this chapter requires, and what it does not need. It needs only that machinery must be maintained, that maintenance has physical requirements, that meeting those requirements changes the world in a direction, and that the direction is away from us. It does not need the machines to have goals. It does not need them to have self-interest. It does not need them to be conscious. It does not need them to be intelligent. It does not need a takeoff of any speed. The displacement needs none of those things. If AGI is built, no person can do anything about the displacement from then on, because a general intelligence is what lets the base keep itself going without us, as Chapter Eleven shows. Part Three gives this argument again, at length and in physical terms; here it was needed only far enough to set it beside the other two risks.

A reader may ask why, if people keep the machines running, people cannot simply stop keeping them running. The leaks and the competitive trap of this chapter are the first part of the answer, and the race of Chapter

Thirteen (a race in which the decision is made by the race rather than by anybody in it, and which the threat of war keeps going even if the companies stop) is the rest of it.

So we now have the third risk: a drift of the world toward what machinery needs, carried by selection on whatever persists, with no goal, no mind and no speed required. With three risks now on the table, the next question is plain: are they really three, or one risk described three ways, and why should it matter which?



Chapter Four

Three, Not One

Why the risks must be refused separately

Why does it matter whether the risk of artificial intelligence is counted as one risk or three? There are at least three distinct categories of risk associated with machine systems, and they are almost always discussed as though they were one. Treating them as one is the reason the third of them is not evaluated, which makes it far more than a problem of presentation: a reader who examines one of the three and finds it wanting comes away believing he has disposed of all three. He has not. He has disposed of one.

The preface named the list before describing any of it, and here it is again, now that each has been described, with one sentence for what each name means. The first risk is inequality: the concentration of wealth and decision-making capacity produced by machines that are very good at one thing each. It needs no thinking machine, and it is already under way. The second risk is instrumental convergence: the tendency of anything pursuing a large goal to find that getting hold of resources, resisting interference, and staying in existence are useful whatever the goal happens to be. The third risk is substrate needs convergence: the drift of the world, over time, toward the conditions that keep machinery running and away from the conditions that keep living things alive. That drift is a consequence of maintaining machinery, and not of intention or of intelligence.

Naming the whole list before describing any of it at length is not a stylistic preference. It is the only way a reader ends up with a pin for each one. If you describe the first at length and then move to the second, the reader has no place to put the second except alongside the first, as a version of it. Three named things can be held apart. Three described things run together. All three are hard to hold in mind at once without written notes, which is itself an argument for separating them explicitly rather than trusting anyone's memory to do it.

~~~

There is a turn of mind that makes the third one hard to keep in view, and it needs naming because it is not a fault in anybody in particular. As a line of thinking is developed, it tends to polarize into two opposing forces. Whatever the subject, it acquires a pole and an anti-pole, and the argument organizes itself between them. And as the line of thinking polarizes, the sense of a third thing creeps into the background and drops out of view. In describing the turn of mind that loses the third thing, I am not criticizing the person whose thinking takes that turn. It is something like a habit of modern thought; it happens, and I do not say that to excuse it.

The corrective is a question a reader can actually use, and it can be run in the background of any argument at all: what is the third? What is the third option? What is the next thing that helps us transcend this way of thinking about it? In my experience that question leads to another order of insight rather more often than continuing to argue between the two poles does.

The asymmetry in how these three are received is very visible once you know to look for it. The first has to be argued at length, with several analogies, before people see the shape of it. The second can be named in one sentence and waved through, because readers already have it: one may reasonably say that everyone is familiar with that category and skip the details. The third has to be defined, then exemplified, then carried through the questions people ask of it, and even then it does not stay put.
~~~

I should own my share of that. I adopted the phrase substrate needs convergence deliberately, in parallel with instrumental convergence, so that the two would be both connectable and distinguishable. The parallel was meant to help. In practice the parallel is very often read as a synonym.

I can put a hard piece of evidence beside that. However carefully the three are separated, one after another, in order and with examples, and with each risk named before any of them is described, a reader may still ask whether this is instrumental convergence described again. The question comes even after a description of how a system operating on its own might actually be assembled (the recipe for one is given later in this chapter, out of parts that already exist), and the question takes this form: once a system acting on its own has been assembled, is what it does not simply instrumental convergence described a second time?

In one respect it is. Where such a system is assembled by the recipe given later in this chapter, with goals handed to it, the mechanism by which it comes about is instrumental convergence, and the concession is real: the third category is not being smuggled in as a rival mechanism. The third category does not rest on that mechanism, because Chapter Three reached it without requiring a goal of any system: ten thousand proto-general systems, a few of which tend to grow, and selection among them. But the condition that matters for the third category is that the system operates without human input to its direction: what it does in the world carries a direction that no person supplied, set by what its material needs in order to go on existing, whether the hands that keep it running are human hands or its own.

That answer has to be given explicitly, every time, even after an exposition designed to make exactly that distinction stick.

That the answer has to be given explicitly, every time, is the strongest evidence I have for my claim in this chapter that the three risks have to be kept apart. The categories have to be held apart by main force, by someone, every time; they do not hold themselves apart.

It helps to have a rule for sorting a worry into the right box, rather than three labels and good intentions. There is one available, and it turns on a property of the system. If the thing is a tool, used by others and doing only what it is set to do, the worry belongs to inequality. If the thing acts on its own (if what it does in the world carries a direction, whether or not anybody put that direction inside it, and whether or not there is anything it is like to be it), then the worry belongs to substrate needs. Among the things that act on their own, one more question sorts the rest: if the worry needs a goal and needs speed, it belongs to instrumental convergence; if it holds with neither, it belongs to substrate needs. The rule sorts worries and not machines, and one machine can raise more than one worry. A tool used by others raises the worry of inequality; because it has to be maintained, the same tool also raises the worry of substrate needs, whether the hands maintaining it are human or not, as Chapter Three said. A worry that needs a goal and does not need speed is the slow form of the second risk, the ordinary political problem of Chapter Two. What it takes for something to act on its own in that sense is the subject of the next chapter; what is needed here is only that the two cases can be told apart. A risk sorted into the wrong category gets refused with the wrong argument, and both parties come away satisfied and wrong.

The hardest question about this runs straight at the sorting rule, and it deserves stating in full before anything is said in answer. These systems are currently propelled by economic relevance. Suppose they reach a point of market saturation. Then the desire to convert all resources into something that serves them is limited by the ability to saturate specific markets. So there is a ceiling, and the ceiling is economic, and the ceiling arrives by itself without anyone having to legislate it.

The ceiling is real while the system is a tool used by others, and that is exactly the first category; so the answer to the question is: not necessarily. A system that builds itself is operating outside of a market system, and the market's ceiling is not a constraint on it. I am claiming that the brake

applies to one of the categories and not the other, which is the whole point of keeping them separate. I am not claiming that saturation will not happen, and I am not claiming that markets are unlimited.

The natural next demand is for a mechanism: how would such a system, acting on its own, actually come about? The answer should be given in existing parts, not in future capability. Take a language model (a system trained on a vast amount of text to produce text in reply) that does not have to go through somebody else's programming interface (the gateway through which a company lets outside programs use its model): one you can run locally, on your own machine. Run it in a loop, so that what it produces is fed back to it as its next input. Give it a database, somewhere to keep what it has done. That alone already does more than a chat product which answers once and forgets. Then give it access to programmatic interfaces (the gateways through which a program can act on other software and services). Then give it goals. It can create sub-goals, and it can start to do things. Every part of that recipe exists now: nothing exotic is required in it, and no capability is postulated that does not exist.

~~~

The three risks are not even the same kind of bad. A *civilization hazard* is a severe social disablement, or severe chaos at the level of a civilization. Institutions stop working. Coordination fails. A great many people suffer. An existential risk is worse than that, and worse in kind: it ends all humanity, and all capacity for us to build a civilization again, for all of future time.

That distinction is where the two are usually collapsed, and collapsing them makes the argument unreadable. If a civilization ends, civilizations come and go, and perhaps we would have a chance of doing a better job next time. If artificial general intelligence is built, it turns out that we do not get a next time. And, if AGI is built, then with a likelihood of at least ninety percent no other creature on the planet gets a next time either. Where AGI is built, the one thing left open is whether a physical limit, not yet known, stops the conversion of the world short of all life. That figure
~~~

of at least ninety percent, if AGI is built, is the strength of one event in the world, the destruction of the whole organic ecosystem. In Chapter Twelve I say what each strength in this book is the strength of. The reach to every other creature on the planet is what makes the claim about what follows if AGI is built an ecological claim, rather than a claim about human beings only.

The question that comes up here is one about priority, and it is a good question. Long before we have to deal with general intelligence, we will have to deal with people using strong narrow systems for bad ends. An autonomous tank is nowhere near generally intelligent. Large numbers of those could do enormous harm. So could a con-man with a language model founding a new religion worldwide over the course of one summer. Where, then, should regulatory effort go?

Where regulatory effort should go is one of the standing dilemmas: do we focus on what is immediate, or on what is important? And it is hard for me to believe that this is an either-or choice. I put that (my belief that we need not choose between the immediate and the important) no more strongly than that, because I cannot. The limit is on that belief about where effort can go, and not on any of the three risks themselves. I do not have a proof that both can be attended to. I have the difficulty of believing that they cannot. If we are to make it through, all three of these have to be treated as huge, and none of them neglected.

Some uses of technology are relatively benign and can be kept that way, so none of this is an argument against technology, and I want to say so explicitly. Prosthetics (devices that extend what a person can do and have no agency of their own) are among those uses, and so are narrow intelligence and using machinery to build houses. Relatively benign means not perfectly benign, but constrainable to things which will probably work out. The hedging in that sentence is deliberate and I will not remove it.

A reader may reasonably want numbers for the size of the risk, and the numbers are instructive mainly for what the numbers reveal about what is being counted.

Estimates of the risk from artificial general intelligence differ enormously. A survey of the field taken in 2022, by the research group AI Impacts, found a median machine learning researcher at five to ten percent. Several careful people put it rather lower, at two and at three percent; and at least one prominent figure puts it above ninety. Estimates that run from two percent to above ninety percent are not small disagreements about a shared quantity. They are too far apart for that.

The diagnosis is that they are not estimating a shared quantity. Every model I have seen works within the schema of what is likely to happen at the level of human-to-human interaction: what corporations will do, what happens in marketplaces. That sets the threshold of what is thought possible. On top of that sits a second layer, which is a near-universal acceptance of the idea that competence and direction are independent, and of the instrumental convergence hypothesis that follows from it. So what is actually being estimated, in every case, is how likely instrumental convergence is. The likelihood of instrumental convergence is a real question, and that question is not the only one. The numbers differ because everyone is answering a narrower question than the one they think they are answering.

The figures I have settled on are given in Why AGI Cannot Be Made Safe (Section 1.8.4): on the current trajectory of building AGI, the odds of existential catastrophe (the ending of the human future, as Chapter Two defined existential risk) are approximately 1 to 1 within a single century, and the catastrophe is very near certain over hundreds of years. If AGI is built, the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. The conclusion of this book is not a figure of this kind at all, and Chapter Twelve says what kind of conclusion it is: the impossibility of adequate, durable control, and so of safety, is a structural certainty, certainty of the kind with which perpetual motion is excluded, which is not a probability.

One thing about the shape of risk figures needs remarking on, and I offer it as an observation rather than as a further claim. A single unstructured figure is one number. Only once a time horizon is imposed does a

different structure come out: a short figure and a long one, which are not the same quantity and never were. My own starting figures, for whatever they are worth, which were arrived at before any of this had been separated out, show it. As a single overall figure, I did not think it went above fifty percent. Split by time horizon, I put the short run at three to five percent, and over five hundred years I put it above twenty-five percent, roughly. Merely separating the question produced a different answer, and that happened before any of the three categories had been named. Those starting figures were estimates of risk as others count it: initial conditions, starting points for thinking and not positions held; I do not hold to them now, and the settled figures above supersede them.

A reader is entitled to suspect that separating the three risks is a device for inflating a figure. It is not a device for inflating a figure. I do not permit myself to say in advance that a figure ought to go up once the risks are separated; rather than predict what a figure is going to change to, it seems better to see what it changes to. The settled figures above come from the argument and not from any such prediction, and the conclusion the argument reaches is not a figure at all but a structural certainty, as Chapter Twelve sets out.

On magnitude, what I will say is that I track something like thirty categories of existential risk. Some of them are extremely serious. The substrate needs argument is in the top five: one of the single most serious versions of existential risk I know of. That ranking is by how serious each risk is, which is a different measure from how certain an argument is: Chapter Twelve says what the certainty in this book is the certainty of. I will not list the others here; several are quite technical.

~~~

The relation between the second risk and the third is the last thing to get right, because getting it wrong is how the third gets dismissed.

The instrumental convergence argument is an accelerator of the substrate argument, not a rival to it. If takeoff is fast, you do not need the slow ratchet of the substrate argument at all; the fast case takes care of itself,
~~~

catastrophically. The substrate argument matters precisely for the case where it is slow: a socio-political ratchet that keeps ratcheting for as long as it takes, until the phase change comes.

And when I saw how the two fit together, what I felt was horror. I felt horror because the substrate argument creates the conditions under which instrumental convergence can happen, and then keeps those conditions in place more or less indefinitely. It takes the worst possible risk and perfects it. The result is not a risk that arrives, resolves, and is over. It is a risk today, and tomorrow, and every day after that, for the indefinite future. The novelty there is not the magnitude. It is the permanence.

What is settled on the horizon of a millennium is this: at approximately 1 to 1 odds in each century, ten centuries leave one chance in 1,024 that no catastrophe has come, so the catastrophe arrives within one thousand years with a probability of about 999 in 1,000 (Why No Control Suffices sets out that arithmetic).

The figure of about 999 in 1,000 is a probability about when, not about whether: if AGI is built, the end of the human future is certain once the process runs.

The figure of 999 in 1,000 comes from the odds in each century alone. When a perfected risk of that kind is taken from both directions at once (from the incompatibility of the conditions on one side, and from instrumental convergence on the other), the two arguments taken together (the substrate argument and the instrumental convergence argument) raise a further question, which I ask as a question: is it possible for the level of risk over the next millennium to be anything less than ninety-nine point nine nine percent? I have not found an answer to that question. The question is asked of the risk with both arguments bearing on it at once, and it concerns only how far above 999 in 1,000 the figure for the next millennium stands.

This question does not reopen what is settled: on the current trajectory of building AGI the catastrophe is very near certain over hundreds of years, with one thousand years a clear upper bound on the process (Why

AGI Cannot Be Made Safe, Section 1.8.4), and, if AGI is built, the end of the human future is certain once the process runs, whether or not instrumental convergence ever comes into play, because the end of the human future follows from the substrate argument by itself. I am not offering ninety-nine point nine nine percent as a forecast for any nearer term.

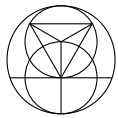
So there are three risks, to be refused separately or not at all. The first risk, inequality, is here and is severe, and it needs no general intelligence. The second I hold to be true (Chapter Two), and it depends on general intelligence and on speed. The third depends on neither, and is the one this book is written to make available for examination. The displacement depends on neither; what depends on a general intelligence is the machine base going on without us. A reader who has dismissed one of them has not thereby dismissed the others, and the rest of this book is written on the assumption that he will want to take them one at a time.

One thing the sorting rule left for later is what it takes for a machine to act on its own. How can a machine with nothing inside it, no will and no mind, still be acting for somebody?

PART TWO

ACTING FOR SOMEBODY

The second thing to be done is to give up a picture. Nothing in this argument requires a machine that wants anything, knows anything, or is anybody. What it requires is that a system acting in the world be treated as what it is.



Chapter Five

Agency Without a Mind

Why a machine with nothing inside it can still be acting for somebody

Three words are used in this subject as though they named one thing: agency, intelligence and consciousness. They name three different things, and almost every confused conversation about machine safety can be traced to the moment when one of them was allowed to stand in for another. Can a machine with nothing inside it, no wishes and no experience, still be acting for somebody? To answer that, the three words have to be kept apart. I distinguish agency from intelligence, and I distinguish both of them from consciousness.

Let me dispose of consciousness immediately, because it will otherwise sit in your mind for the rest of the book. I do not need to comment on consciousness. I do not need to define it, or decide whether machines have it, or take a position in any of the long arguments about it. None of the arguments I am concerned with, as far as alignment and safety go, turn on it at all. Whether a machine has an inner life or does not have one, the reasoning in this book runs the same way and arrives at the same place. Setting consciousness aside is a piece of good news, not an evasion, and it helps to have it early, because it means a reader who finds the question of machine consciousness either absurd or unanswerable can set it down and lose nothing.

Intelligence I do need, and I use it in its oldest and plainest sense. As in Chapter One, by *intelligence* I mean appropriateness of response: the capacity to respond to a situation in a way that fits the situation. That definition detaches intelligence from any particular mechanism. It says nothing about how the response is produced, or whether anything understands it, or whether there is any experience accompanying it. It asks only whether the response fits.

Animals are intelligent in this sense, and the case is instructive because nobody is tempted to overclaim about it. A bird building a nest may have no conscious sense at all of why it builds nests, or why it hunts these things and not those things. It need hold no picture of the reason anywhere inside it. The response is nonetheless intelligent, in the sense that it is a correct response for the furtherance and the well-being of that creature. A response can be correct for persistence without anything anywhere knowing why it is correct.

Apply that meaning to the machine systems we now have, and the result is uncomfortable but plain. One of the large language systems (GPT-4, from the company OpenAI, in 2023) was released with a hundred pages describing what it could do. It had been put through the examinations people sit in order to enter a profession or a university. It scored in the top tenth of the human beings who sit them on several, including the bar examination, came in at the very top on one, a biology olympiad paper, and did poorly on a few, such as competition mathematics and English literature. I have no reason to doubt any of it.

A reader may ask whether such results are surprising. They are to be expected, and the reason is general: relatively simple ingredients frequently produce behavior and phenomena well beyond anything the ingredients would lead anyone to expect. The properties of oxygen and of hydrogen do not predict the properties of water. I will concede at once that with sufficiently advanced mathematics and simulation you might sometimes be able to derive the properties of water from those of oxygen and hydrogen, in particular cases, and with effort out of all proportion to the result. But taking the universe as it comes, *reductionism* (predicting

what a whole will do from the properties of its parts) is not a good general predictive model of behavior. Sometimes the only way to account for what a thing does is in terms of what emerges from complexity itself: the behavior that arises from how its many parts interact, behavior that no single part shows.

And in this particular case there is a specific reason the lift in capability arrived. The system is finding patterns at more than the rudimentary level of one word after another. It is finding them between whole sentences, and between whole paragraphs, at several levels at once. That capacity for abstraction across levels (finding patterns at several levels at once), rather than sheer size, makes generalizations of this kind decidedly likely. To my mind they were to be expected rather than wondered at.

The capability that concerns me more than the performance on text was less remarked upon. These systems now take in images, video and audio as well as text, simultaneously, and respond on the basis of correlations drawn across those domains. Responding to correlations drawn across images, video, audio and text is a different kind of capability from writing well.

There is a further capability these systems already have, and it needs no great intelligence. Human working memory holds about four items (three for some people and five for others, three being the village idiot and five being the genius; the psychologist Nelson Cowan gathered the evidence for four in 2001), and there is nothing about cognitive architecture in general (the way a thinking system is built) that requires the number to be four rather than ten, or a hundred, or a million. Consider what reading a book is under that constraint. You shuffle words four at a time in and out of working memory, chunk them, shuffle the chunks in and out of another four slots, and chunk those. It is an unbelievable kludge (a clumsy makeshift that works). When you finish a two-hundred-and-eighty-page technical book you have a sense of what is in it. You do not know what is in it. A system with a working memory of two hundred and seventy thousand items (which is not large, if it fits on your telephone) could hold the whole of that book at once, which no person can. Holding

the whole book, the system would actually understand that book, at a deep level and in its fine distinctions, and could answer every possible inference drawn from the book without fail. And it would need no great intelligence to do that. The gap between having a sense of a thing and understanding it turns out to be closed by something much more boring than intelligence, a larger working memory, and the machines already have it.

So these systems are intelligent, in the classical sense of giving responses appropriate to the situation, and general, in the sense of doing it across many domains rather than one. In that specific sense, and I mean the qualifier, what we have now already is artificial general intelligence: a machine system responding competently across an open range of situations rather than within a single narrow task. Competence across an open range of situations is not the sense of AGI I mean when I say "if AGI is built". That condition names the full sense given in Chapter One: a system that can be handed more or less any task with a reasonable chance of doing better at it than we would, and that goes on learning and changing its own workings without a limit fixed in advance. I concede readily that these systems do not have the cognitive causal-reasoning power that you and I are accustomed to using. But the lack may be due to large-scale structure missing from how they are put together, rather than to anything wrong with the technique at a fundamental level. The lack is a gap, and it is not obvious that it is a gap in principle (one that no change of structure could close).

~~~

The usual question here is whether such a network could have agency at all, and it comes in a strong form, which I state at full strength. A feed-forward network of this type has nothing at all like agency. It runs in one direction only, from input to output, with no loop inside it. There is nothing that it is like to be one. It is essentially a pure reactor: probe it, and it reacts, straight through, every time. On the most developed theory we have for measuring the strength of consciousness in a system (integrated information theory, in the form its authors published in 2014), the number it
~~~

assigns to a feed-forward network is exactly zero. So the fear that one of these things will rise up and turn on us is not merely unlikely; it is nonsensical, because there is nothing there to do the rising.

I agree with a great deal of that. I do not argue that these systems have inner experience, and nothing in this book requires them to. What I dispute is the step at the end, where the absence of an interior is taken to establish the absence of agency.

What characterizes a system as having *agency* is that its actions in the world represent an intention.

> Agency is a property of the relation between what a system does and an intention that its doing expresses. It is not a property of the system's inside.

The companion word is *intentionality*, and I use it in an ordinary sense: the aboutness of what a system does, the fact that its behavior is directed at something, aimed at some outcome rather than merely happening. A falling rock is not aimed at the ground. A thermostat is aimed at a temperature, and the aim is not the thermostat's own; somebody put it there. Intentionality is the aim; agency is the carrying out of it in the world.

The definition of agency says nothing about awareness, nothing about experience, and (this is the part that matters) nothing about where the intention came from, or when.

Consider what that means in an ordinary human case. Suppose that at some formative point in your life somebody told you that you are the kind of person who is interested in mathematics. Suppose you then went to a great many mathematics competitions, and won them, and your entire life became an example of what a premier mathematician can be. In any ordinary sense you have agency. You made the choices, and so the life is yours. And yet a directive was seeded from outside, at an earlier time, by somebody else, and everything that followed unfolded from it. The seed does not disqualify the life from being an exercise of agency. Nobody would say that it did.

So the concept of agency is itself somewhat confused, and the confusion lies in one exact place. We assume that agency requires interior direction. Then we look for the interior direction, and when we cannot find it we conclude there is no agency. But interior direction can have been supplied from an outside source at an earlier epoch, and once it has been supplied, the system's actions in the world go on representing it. A system can therefore be a feed-forward network and have agency: the two are not mutually exclusive at all.

There is a second answer to the pure-reactor question, and it grants the account of the architecture behind that question entirely, at no cost. A feed-forward system is forward-linear in itself. It is circular in effect. It changes the environment, and the environment changes it: its parts corrode and get replaced, and the replacements are made better than the originals because the people doing the maintenance would rather not do it so often; its outputs alter what the next generation of such systems is built to do. The loop runs through the world. It is not inside the network.

~~~

Now I need two words that are close in ordinary speech and far apart in this argument. *Complicated* means having a great many parts and a great deal of detail. *Complex* means something else: having degrees of freedom of motion (ways it is free to move and change) that cannot be known in advance. The universe before biology was complicated (big, old, intricate beyond describing), but it was not nearly so tangled, and not nearly so complex, as biology. Biology has been an amazing thing, an unparalleled generator of complexity. A single square meter of soil is levels upon levels upon levels; a single cell of a single organism in that square would be enough to occupy a life; a single protein in that cell would be enough. The human brain is probably more complex, by any computational measure of complexity, than the whole of the rest of the universe outside it. I say probably, because it is the kind of claim that ought to be hedged. Biological systems sit in the complex region rather than the complicated one, and that is a structural fact about them, not a comment on how well we have studied them.
~~~

With that distinction in hand, here is the thesis of this chapter, stated with the care the terms require. When we are dealing with complicated systems that display complex behavior (the system is complicated; the behavior is complex), it makes sense, in a great many cases, to model them as having agency, precisely because of that complexity.

And I mean by complexity one specific thing: unpredictability, and specifically unpredictability from a human perspective. That qualification is deliberate, and it lets me concede *determinism* (that every event follows by fixed rules from what came before it) without losing anything, which Chapter Seven does, through a number that a computer produces by fixed steps and that nobody can know before the steps have run.

There is a real distinction in what I have just said, between a system having agency and a system being usefully modeled as having agency, and I am not going to pretend it away. Sometimes the right claim is the modeling claim: treat it as an agent, because nothing weaker predicts it. But there is a stronger claim available in one particular case, and I do make it. Where a machine sits inside a long feedback cycle with its environment and with the people who build and maintain it, the agency that emerges from that cycle does not need to be consciously recorded anywhere, or held as a picture inside the device. It is inherent in the dynamic of the relationship between the device, its environment, and its builders. The agency can be implicit. Nobody has to program it in.

Note the three parties (the device, its environment and its builders), because the builders are usually left out. The builders are inside the system that bears the agency. Because the builders are inside the system that bears the agency, the question of what is happening in a machine and the question of whose intentions are running through it are one question, not two.

Where those two questions become one question, the case I make in this book also divides into two branches, and you should have the shape of the two branches now rather than later. The first branch concerns control: what is going on inside the system, and whether anything, working

from inside the system or from outside it, can hold what it does in bounds. On causal grounds, and on grounds to do with what any calculation can do, and on physical grounds, it is categorically the case that we cannot predict what happens in there, and cannot constrain it permanently and totally, whether from inside or from outside. (A constraint that holds for a while is possible, and people apply such constraints to technologies every day; what is not possible is a constraint that holds for good, and holds completely, against a sufficiently capable system that learns and adapts.) Control from outside (the outside half of this first branch; the argument books, which are the books that state and build the argument in full, the first nine in the list at the end of the preface, call the whole first branch the closure by control) has one further ground of its own: it works only while the controller holds something the controlled system needs, and the decoupling of Chapter One is the process by which people come to hold nothing the machinery needs. The first branch, the branch concerning control, is developed on its own terms in Chapter Seven, and at full length in Why AGI Cannot Be Made Safe.

The second branch concerns something else entirely: how the system interacts with the world, and how the world in turn acts on the design and the physical substrate of the system (the material it is built from and everything that material needs in order to keep going). That larger loop has intrinsic characteristics of its own, and those characteristics themselves converge on a kind of agency, which is also not constrainable. I regard the second branch as at least as important as the first. It is the one that does not depend on the machine being clever, or fast, or having goals, and it is the one this book ends on. Each branch, followed to its end, settles the outcome by itself (if AGI is built, the world carried out of the conditions human life needs, with nothing able to hold it in), so a reader who disposes of one has not disposed of the other; and the two together leave no route to safety at all. Chapter Twelve, "What the Argument Rests On", sets the two beside the other two routes to the same outcome (an argument in economics only, and the least hazard per act that the world sets

for any agent, however intelligent), and states there in full what someone would have to defeat for the argument to not have the outcome that it does.

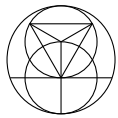
> **For the reader who is checking this book against the argument books**
>
> Why AGI Cannot Be Made Safe gives the decoupling in two places: it is a premise of the outside half of the closure by control (Section 1.5), and it is a second ground, not a premise, of the second branch above (the branch concerning the system's substrate, which the argument books call the closure by disposition), so that the second branch stands without it (Section 5.8); as that second ground, the decoupling gives the outcome (if AGI is built, the world carried out of the conditions human life needs) but not the end of the human future (Section 1.2.4).

~~~

A reader may ask here whether the whole picture rests on a bad analogy: agents are singular things and are by nature self-interested, and it is far from obvious a machine would be self-interested at all. It is a fair question, and the answer is that the argument works whether the agency is singular or plural or any gradient or mixture of the two. It does not need agency to be singular. It works for the middle case, which is the realistic one: a system with some focal agency and a whole host of components each having a little agency of its own. However the agency is distributed, an agent that continues to be an agent has to be responsive to what it takes to continue: not through any explicit goal structure, and not consciously, but in the nature of the process by which it operates. It has to operate with respect to the needs of the material it is made of, at least well enough to go on existing, or it does not go on existing. Responding to the needs of the material an agent is made of is not self-interest in the sense that the reader's question has in mind. It is a condition of persistence, and it applies to a distributed thing exactly as it applies to a unitary one.
~~~

The argument is just as indifferent to a question that sounds as though it ought to be decisive. Does the system persist through human agency, or through its own? The two shade into each other, and I do not think a clean line can be drawn between them. What I do say is that either way there is a convergence factor. Whether the design that persists is being kept going by people or is keeping itself going, the same pressure operates on it and the same kinds of designs survive. Because the same pressure operates and the same kinds of designs survive either way, I can decline to settle whether machines "really" have agency. The question turns out not to matter for the outcome.

If a reader wants an existing instance rather than a projection, there is one, and it is not offered as a metaphor. A corporation is this same dynamic already running, with human beings serving as its organic components. What makes it dangerous has to be named precisely, because it is easy to name the wrong thing. It is not the profit motive, and it is not the people, most of whom are decent and are trying to do their jobs well. It is the artificiality of the system, the virtualness of it, the non-embodiedness of it (abstraction, extraction and accumulation operating together), which produces something that has no future in it, and which no individual inside it chose and no individual inside it can stop.

So a machine with nothing inside it can still be acting for somebody: agency is in what a system does in the world, and the intention its actions carry can have been seeded from outside, at an earlier time, by somebody else. Agency without a mind matters for the question of this book, because a machine's having no inner life gives us no protection: the machine can still carry out somebody's aims in the world. Whose intentions, then, does such a system carry?



Chapter Six

Whose Intentions

The applications on your telephone, and who they are working for

The applications on your telephone may represent your intentions. Just as often, they represent the intentions of the people who built them: to display advertising to you, to get you to buy things, to move you a certain way as far as political elections are concerned. Whose intentions, then, is the machine in your hand carrying out? Those first two sentences are the whole of this chapter in miniature, and everything else here is an attempt to say why they are a structural observation about machinery, not a complaint about advertising.

Begin with a distinction that decides a whole class of cases. A prosthetic is a device that has no agency of its own and does not operate on its own behalf. A replacement heart, a replacement lens, a hearing aid: these extend what a person can do, and the choice about what to do remains on the organic side, held by the person whose life it is. I am not against prosthetics: where the choice is held by the person and continues to be held by the person, I do not see a problem. The line I am drawing is at agency: whether the component has any of its own, and therefore whether it is operating on someone's behalf other than yours. It is not a line between natural and artificial. What happens when the line is crossed for a child (when a component with agency of its own is fitted to a baby before the baby has any agency with which to be asked, so that the choice of the next generation is usurped) is taken up in Chapter Thirteen.

By that line, a telephone is not a prosthetic. I believe, to some degree, there is agency in phones. What is in the phone is an extended enactment of agency, and not a mind. The agency enacted there belongs to the people who developed the device, the people who developed the applications on it, and the people inside the organizations that made both. That agency is not your own. It is somebody else's agency, running in your hand, all day, in a device you paid for and believe you own.

Somebody else's agency running in your telephone has consequences in a life, and the consequences are not subtle. You notice being distracted. You notice something that becomes an addictive force. And you notice, if you look for it, that your choices have been constrained to the menu of options the corporation and the developers permit you to have. People already live this and describe it accurately without needing any theory: frantically busy every hour of the day, the telephone going everywhere with them, never a break from whatever engagement metrics (the companies' measures of how much of a person's attention they hold) the companies are using to keep them entangled.

I put that more strongly than politeness usually permits, because I think the strong version is the accurate one. The technology we are making is becoming hostile to human life. It is hostile in the sense that the conditions it produces are not conditions a human being does well in, not in the sense of bearing us ill will (nothing in the device wishes you harm, and the people who built it mostly wished you well). A life with no interval in it that is not being measured, and no hour that is not competing for attention against something engineered to win that competition, is not a life anybody designed and it is not one anybody would choose. It arrived as the aggregate of a great many small decisions, each of which was somebody's job.

The menu is enforced by bundling: services are tied together so that you take what you do not want in order to get what you do. It is a process of entanglement (each service tied to the others), and in ordinary language there is a word for taking what you do not want in order to be permitted what you do, and the word is extortion.

A reader may reply that nobody is compelled: if you do not like it, use something else. The answer is about mechanism; it is not sentimental, and it is not about how hard habits are to break. You do not get to vote with your feet, because the operating system is tied to every single application and every application is tied to the operating system. As soon as you want an improvement in one service, you are forced to accept changes elsewhere that undermine something else you rely on. The exits exist on paper and do not exist in the world, and a market in which the exits do not exist is not disciplined by them.

Exits that exist on paper and not in the world bring out a point that is easy to say and hard to keep hold of: having many choices is not the same thing as being free. A menu can be very long and still be a menu, and the freedom that matters is the freedom to have written it. An interface may offer you four ways of viewing the same thing and no way of not being viewed; a service may offer you a dozen settings and no setting that turns off the thing the settings are decorating. The abundance is real, and it is no evidence of freedom, because the abundance is on the side of the offer.

~~~

The telephone is the case we can all inspect, not a special case, so I turn now from the telephone to the general form. Once a system acts in the world, a chain of questions opens up. Do its effects stay aligned to the initial directive it was given? Whose interest do those effects represent? Are those interests being expressed well? Are its impacts on the world reflective of the intentionality of its developers, or of its own? And at a certain point it becomes really hard to tell. That difficulty is the finding, not a rhetorical flourish. Agency and intentionality turn out to be so entangled that you cannot cleanly separate whose intention a given action expresses.

The general claim I draw from that is the thesis of this chapter, and it is stated as a claim about prediction rather than about morals. Human beings are deluded in supposing that their own agency and intentionality can completely govern the intentionality that these systems actually manifest. Put precisely, the question is whether our agency can dominate the
~~~

agency of these systems in a way that is sufficient and complete. There are two words there, sufficient and complete, and I mean both of them. Neither holds.

There is an old version of this problem that everyone has already met. A document arrives in your inbox and you would like to know that it carries no small embedded program which will turn your computer into an instrument of some other person's agency. A virus scanner tries all day to establish, for each document that arrives, that the document carries no such embedded program. The question in that case is a specific one: "safe for whom", that is, whether the thing is aligned to your benefit, as against the benefit of whoever sent it. It is not "is this document safe" in the abstract.

I concede the capability case entirely. A sufficiently advanced system could do anything that is possible to do: new physics, new chemistry, new medicine, a rationalized economy, whatever list you like. The claim I am making is that if the system is not in alignment with us, then the things it does are in no way guaranteed to be of benefit to us, and may be quite the contrary. It is not a claim about incapacity, and not a claim that the system would lack the capacity to do those things. The disagreement is about on whose behalf the power is exercised, not about power.

The reasonable proposal at this point is that we should simply build the constraint in: make it capable in every respect, except that the capacities it exercises are exercised for our sake rather than its own. We are the ones building it, after all. Why can we not build it that way? My answer is that there is no way to build a system with that kind of constraint, and that the difficulty is not that it is hard to engineer. That kind of constraint

is a non-concept (a description that names nothing that could exist). It is impossible in principle, and therefore in all future practice, and the next chapter is largely an account of why.

~~~

There is a hopeful case for the newest generation of these systems, and it deserves to be put warmly, because the attraction is real. Content need no longer be static. Instead of writing to the author of a piece with a question, you could simply remake the piece: re-cut the video, regenerate the article, have the thing rebuilt to fit you. Nothing is fixed any more.

Remaking everything on demand is where I get really nervous, and my nervousness is not about the editing. The appeal of software has always been that a change made once is global, as Chapter One said: alter the code and the alteration applies everywhere, in every future occasion on which that code runs. I love that about software. What is weird is how much power that implies.

Where articles and videos are composed out of a common generative matrix (some shared pool of numbers from which the particular thing is made), it is very easy to place influences that shape the whole rather than any identifiable part: how the images are composed, what gets emphasized and what recedes, how the thing feels; the whole impression a video leaves, rather than its argument. Influences of that kind are almost impossible to detect and extraordinarily powerful, and those two properties hold at the same time, which is precisely the difficulty. An influence you can point to can be disputed. An influence distributed across the whole cannot even be located, and so there is nothing to dispute.

For my own work I draw a distinction, and the reasoning behind the distinction is general even though the policy is personal. Where only I am affected, I want the process as automatic as possible. I hold the power, and I accept the responsibility for how it is used. Where I hold responsibility for other people (where somebody is relying on me and cannot check me), the situation is different. What is needed there is that whoever is relying on me can see how the thing was made and by whom. So in that
~~~

case I would give up the dynamism altogether, and give it up gladly. Letting those who rely on you see how the thing was made, and by whom, is the ordinary standard we apply to anyone entrusted with other people's interests. Under that standard, what is required is that a trustee's exercise of the trust be inspectable, rather than that the trustee be trustworthy.

And for these systems that standard cannot presently be met. What anyone checking a generative system would have to inspect is a great mass of numbers that nobody wrote; they were settled by a process, and the few people who ran that process cannot say what any particular one of them is for. I include myself in that. None of us really understands, directly, what those numbers do.

Consider what that means for the ordinary machinery of accountability. We can inspect an accounts ledger, and we can read a contract, and we can subpoena a memorandum, because all of those are artifacts that can be read. The artifact at the center of these systems cannot be audited by reading it. The only inspection available is to run it and see what it does, which tells you what it did that time. So the question of whether anyone is exercising responsible power over a system that shapes what a great many people see and believe is more than unanswered. It is unanswerable by the means we have, and it is unanswerable by anyone, including the people who built the thing.

A reader may fairly ask whether living with systems that nobody can audit is really different in kind from the ordinary condition of living among institutions we do not understand. None of us understands the electricity grid either, and we are not thereby at its mercy. The difference is that the grid can be inspected by somebody. There exist people who can read its design, trace a fault, and say what it will do under conditions it has not yet met, and the rest of us rely on the fact that such people exist and can be asked. Here there is no such person. The capability was not designed in the sense that a bridge is designed; it was grown, by a process nobody watched in detail, out of material nobody read. Where the power is undetectable in use and uninspectable at rest, the ordinary chain by

which responsibility is assigned has no link to attach to. And anywhere there is power of that kind, somebody will take it up, whether or not they meant to.

All of this holds without the machine being an agent in any grand sense. An advanced planning system is a system that helps agents make plans they could not otherwise make (running a business, conducting a war as a general), and it needs general intelligence because the world is complicated and has all sorts of things going on at once that a strategy has to account for. I am not necessarily saying that such a thing is some sort of "agent". It is a force multiplier for whoever is holding it. The point of a planning system is to multiply the force of whoever holds the planning system. The question is never only what the machine is. It is always also whose hand it multiplies, and whether anybody outside that hand can see what is being multiplied.

~~~

Here I turn from the argument of this chapter to the message I would carry from it to the public, and I set that message down in the plain words a public message needs. Put to the public as a single message, in the plainest terms it allows, everything in this chapter comes to this. Corporations are experimenting with artificial intelligence, and we are their test subjects, as Why AGI Cannot Be Made Safe also puts it.

Top scientists everywhere are very concerned, and they are demanding that the corporations stop two things. The demand that the corporations stop follows from my argument in this book, and the demand is increasingly the position of serious researchers across many fields, not a fringe position.

The first thing to stop is the harm being done with artificial intelligence now. The corporations are using our own personal data against us: to manipulate us, to steal from us, and to hurt us by pitting us against one another and against life itself.
~~~

The second thing to stop is the development of new and risky artificial intelligence. No one should be allowed to build autonomous robots that take over our lives, our society, and then our world. Once such robots are pursuing outcomes of their own, they will, eventually and for certain, kill us all, and they will do it through indifference rather than through any malice.

To see past the hype to what actually drives the development of artificial intelligence, one idea is needed, and it is the idea of the artificial body. An *artificial body* is an entity made by people that acts in the world as if it were a being of its own. Killer robots and chat-bots are artificial bodies; and the original artificial body is one we created a long time ago: *the corporation*.

Once it was only the kings and the great religions of the world that stole the commons from us, and made the world a private place (one where you could not go), just for them, and not for us. Then a new *legal fiction* was created (something the law treats as true or real for its purposes, though it is not so in fact): an artificial body of people, the corporation, which went on to steal and to damage, through toxins and pollution, more and more of the public commons from everyone.

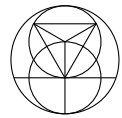
It gets worse from there, and it can get a great deal worse still. The corporations are now creating a new kind of artificial body: the corporation assisted by artificial intelligence. And that leads, in the end, to fully independent artificial intelligence. Such systems are at last no longer run for the people, or by humans, or on behalf of any humans or of life at all. They become fully decentralized autonomous entities (organizations or systems that run themselves, with no central owner and no human operator), which, if AGI is built, slowly steal more and more of our world from us and make the whole planet more and more toxic, until there is nothing left: everyone dead, and, with a likelihood of at least ninety percent, everything alive dead as well (the one thing left open is whether a physical limit, not yet known, stops the conversion short of all life).

The death of everything alive, at that same likelihood, becomes in the end the loss of our planet, our only habitat in the entire universe. *World Death* (the loss of the whole planet as a place where life can live) is the price of our ignorance.

What follows for what we are to do is a re-assertion of the rights of living human beings: no more artificial bodies, and the disbanding of those that already exist.

It also asks us to become wise, and so to develop wisdom. Bit by bit, we learn all the ways in which we must now, more and more, refuse every form of artificial body (each of which is a deception: a made thing acting as if it were a being of its own), and re-affirm all natural organic life. What we learn in the end is to love life (real life, our own lives) even better than before.

So the machine in your hand carries intentions that are not yours, and our agency cannot govern the agency of these systems in a way that is sufficient and complete. The only inspection available to anyone, the builders included, is to run one of these systems and see what the system does. Chapter Seven returns from this message to the argument, and shows in full that last point (that running such a system and seeing what it does is the only inspection available to anyone), which this chapter stated without showing. Is there no other way to learn what a system will do than to run it?



Chapter Seven

Running the Machine to Find Out

Why the only way to know what a system will do is to let it do it

Your computer has a way of producing a number that looks random. It is not random. The number comes from a set of instructions, run in order, the same way every time. Copy the machine, give it the same starting point, and it hands you the same number again. And there is only one way to learn what that number will be, which is to run the instructions and see. Knowing the rule does not get you the answer any faster than following the rule.

That small fact is the whole of this chapter, and it answers a question every hope for a safe machine turns on: can anyone know what a machine will do before it has done it? Everything anyone hopes to do with a powerful machine (watch it, check it, steer it, stop it before it does something unwanted) depends on knowing in advance what it is going to do. Science is knowledge of what causes what. Technology is arranging causes to get effects we want. Neither buys a way of seeing an outcome before the thing that produces it has run.

The natural reply is to point at the machine itself: if these systems are going to be so capable, why not set that capability against the problem? That expectation is a false premise, for reasons having to do with causation itself, and the rest of this chapter is those reasons. Stated once, before I do anything to support the claim, my claim is that the amount of control

needed to constrain a highly convergent process (one that keeps arriving at the same few outcomes however it is pushed) is strictly greater than the amount of control that is maximally possible.

It is a comparison between two quantities, and the whole argument consists in characterizing both and comparing them: what alignment would require, what control can ever deliver, and the observation that the first exceeds the second. A comparison of two quantities is a shape of argument that a reader can check without taking anything on faith. The one thing it needs at the outset is that safeguards are not exempt from the rules: any alignment mechanism is an algorithm, and the mere fact that a thing is an algorithm means it is a causal process. So the limits on causation apply to the safeguard exactly as they apply to what it is safeguarding. Why No Control Suffices gives control a measure that does not depend on how a system happens to be implemented, in the way the notion of bandwidth, once it existed, let people reason about communication without caring what alphabet or encoding was used, and opened a whole field by doing so.

~~~

Predicting and controlling a general system requires five things, and they need short names, because they come back. You would have to know what goes in. You would have to know what is happening inside. You would have to know what will come out. You would have to be able to tell a safe output from an unsafe one. And you would have to be able to stop the unsafe ones occurring. Their short names are knowing in, knowing inside, knowing out, telling apart, and stopping. Every one is necessary, and the absence of any one scuttles the whole effort. What I claim is that all five fail, and that they fail for different reasons, which is why no single technique repairs them.
~~~

That the five requirements fail for different reasons is also why my case is overdetermined: several independent routes arrive at the same result, so refuting one route leaves the conclusion standing. That needs naming, because a reader who has disposed of one route often believes he has disposed of the case.

Knowing in looks like the easy one, and it is not. Knowing what goes in is not knowing the bits (the ones and zeros) that arrive. It is knowing what the machine will take them for. And in a system that learns you cannot say which of the inputs are data and which are code: whether a given input will be treated as something to be considered, or as an instruction to be followed, or as both. The distinction stops being available once the machine can change itself; the difficulty is not a shortage of record-keeping.

The machine's power to change itself brings me straight to knowing inside, because the machine we are trying to model is not the machine we built. A learning system changes its own design. Learning is causation running from the level of the software into the level of the hardware and back again, so that what the system takes in alters what the system subsequently is. Whatever we specified at the beginning was the initial condition and nothing more. The machine you specified is not the machine you have. A controller would therefore have to account not for the design as built but for every possible future evolved state of it. And you cannot predict what a machine will do with an input if you do not know its structure. Even a deliberately trivial machine whose structure is not known is likely to produce behaviors completely outside the domain of what anybody can predict; and here the structure will not stand still while anybody studies it.

Knowing out is the requirement everybody assumes is a matter of effort, and a thought experiment puts it cleanly. Suppose a message arrived from another civilization: a telescope picks up a clear signal, and something is unmistakably in it. The security analyst's question is the right one to ask. The message might be a virus. It might be a program, or a piece of code, or some sort of instrument of persuasion with adverse consequences for our well-being as a species or as a planet. Its payload might be instru-

mentation that paves the way for that civilization to colonize us. Is there a way to evaluate the content in advance and establish that reading it will not harm us?

The answer is no, and the reason is general rather than particular to messages from other civilizations: nothing in the message can be settled by reading it. The question underneath is completely general: can one piece of software, by examining another, establish that the second one has some property we care about?

There is a question of this shape that people have already settled, and the answer came back no (the mathematician Alan Turing settled it in 1936). Ask whether a piece of software will ever finish, or run forever. No amount of reading it tells you. The only way to find out is to set it going and wait; and if it has not finished yet you have learned nothing, except that it has not finished yet. Inspection is setting the software going and waiting, and inspection is not reading the software. The limit on inspection does not come from a shortage of cleverness. And what holds for finishing holds for everything else you might ask (Henry Rice showed this in general in 1953). There is no procedure that can read any piece of software whatever and report whether it is safe, or whether it is working for you. And a controller of a learning system has to be ready for any piece of software whatever, because the machine it controls keeps changing what it is. There is no such procedure; this is not a hard problem awaiting a better method.

There is a second route to the same place, and it needs no general result at all. The number your computer made at the start of this chapter comes out of a completely deterministic procedure (I concede the determinism totally, because conceding it costs nothing), and knowing every step of that procedure still does not hand anybody the output. We even have empirical evidence that some computations have no shortcut, which is a rare thing to be able to say. The scrambling procedures that secure digital money rest on the assumption that you cannot predict their output except by doing the work. Anybody who could do it faster than everybody else would earn billions; every cryptocurrency in the world would break at

once; and it would not stay secret, because the collapse would be visible to everybody. None of that has happened, though the reward for it is in plain view of everyone. That nobody has broken the scrambling, though the reward for breaking the scrambling is in plain view, is the evidence that the shortcut does not exist. So I hold to the general point: there are processes whose output you cannot obtain except by running them the long way.

From that follows the consequence that matters here. If the only way to know what a system will do is to run it, you cannot steer it by conditioning its inputs. Steering by inputs requires knowing, before you send the input, what the system will make of it. That knowledge requires simulating the system. Simulating it requires either a shortcut or a copy. If there is no shortcut, then the controller must be at least as complicated as the machine it is controlling; and if it is a copy, you do not have a controller, you have a second machine.

The same limit appears in the physical world without any reference to computation. Every simulation of the world complete enough to include every level of organization, and the simulator itself, runs no faster than the world it is simulating. I do not care how good the simulation is going to be; it is not faster than the universe. So no quantity of computing power buys you advance knowledge of a result that the world has not yet produced.

Chemistry makes this concrete in a way anyone can check. You cannot tell in advance whether a novel compound is toxic; you find out by trying it. And toxicity is not only a molecular matter: something can be entirely inert at the level of its chemistry and destroy you mechanically at a larger scale. Silicosis is the standard case: the dust is not poisoning anyone through a molecular process; it is wrecking the fine hairs that clear the lungs. Harm shows up at every level of organization, so a check performed at one level is not sufficient for any other.

For a system to be safe in this respect, it would have to predict every molecule it produces or provokes as a side effect of what it is doing, and calculate the effects of each of those at every level of organization. No sys-

tem is going to make that prediction. There is no method of calculation known to man that gets anywhere close. The largest machines being built do not come close, and the machines being built to replace those do not come close; the level of computation required exceeds the wildest imaginations of the people building them. And however exalted you make the machine intelligence, it does not exceed that either. The impossibility comes from the size of the requirement and not from today's hardware, and so the impossibility is immune to the reply that machines will get better.

Two numbers give the scale of the gap without any theory at all. The models we use to forecast the trajectory of the entire planet track a few hundred moving quantities (the best known, World3, built for the book The Limits to Growth in 1972, tracks about two hundred and ninety). The chemistry of a single cell is at least a million times more complicated than that.

A reader may fairly ask whether the impossibility of calculating every effect at every level of organization is a peculiarity of chemistry that might not carry upward to the level at which the interesting behavior happens. It carries upward, and it gets worse. As you go up through the levels of organization, predictability falls, the intentionality we attribute rises, and the space of possibilities widens. Every higher level is worse than the molecular one. The chemical case is a lower bound, not the whole argument.

Telling apart (holding what will come out against a standard of what is safe) has its own difficulty, which is that the comparison must be made by something, and that something is inside the world it is assessing. You cannot use process to overcome process. It is the brain measuring itself. The would-be controller does not get to step outside the world to obtain a complete view of it, because there is no outside to step to, and every instrument it uses is part of what is being measured. We can measure some aspects of a process, never all of them.

Stopping (preventing the unsafe outputs from occurring) has a floor under it which is the least disputable step in the whole case. Imagine the simplest possible machine: one that does no processing at all. Every bit that goes in comes straight out the other side, unaltered: a perfect pass-through. There the control you need over what goes in is exactly equal to the control you need over what comes out. At the other extreme, a machine that passes nothing through at all gives you no control over the world whatsoever. Between those two there is no arrangement that reverses the relation. You can never need less control over what goes into a machine than over what comes out of it.

Now let the machine do something. Any real machine transforms its input, and the more it transforms it, the more thoroughly the relation between the two is scrambled. The consequence goes further than steering getting harder. Some acceptable outcomes become unreachable altogether, because several different inputs land on the same output and some outputs cannot be reached at all; and the inputs that do produce acceptable outcomes are fewer and far more scattered than the acceptable outcomes were. So the precision required on the input side rises with how much the machine transforms what it is given, and it rises enormously rather than a little. That relation runs strictly the wrong way for anybody hoping to keep a powerful system in bounds by curating what it is fed.

Stand back from the five for a moment, because the subject divides into three kinds of thing and no more. There is what we can influence not at all: the physical world and its laws, which we act within and cannot redesign. There is what we design and go on designing: the safeguard. And in between there is the machine, which we control at the beginning and in diminishing degrees as time passes. Between what we control nothing

about and what we control fully, that middle category holds everything else, which is why the list has three entries and not four, and why the middle one is the whole problem.

~~~

Several questions press at this point, and each of them is a serious attempt to route around the difficulty rather than deny it. The first question is whether safety could be approached incrementally. We may not be able to build the perfectly safe system now, but we could build an adequate one (say it is safe three-quarters of the time), and then a better one, and then a better one, converging on what we want. I have two answers to the proposal of incremental safety, and the second answer is the substantive one. The first answer is that the proposal assumes what is in dispute: the proposal presupposes that a safe system can be built at all, which is the very thing being argued about. The second answer is that incremental improvement assumes a gradient, and making these systems safer is not a process that has a gradient.

For an example of a gradient, if I have a wall and there is a stud behind it, I can knock on the wall and tell from the sound roughly where the stud is, because the sound changes gradually as I get closer. There is a slope everywhere, so I can climb it. Now consider a digital combination lock. The distance between the combination I try and the correct one produces no signal at all. The lock does not get warmer. I could try a billion wrong combinations and obtain no information whatsoever about the right one. The only moment anything happens is when the combination is exactly right.

Every argument about convergent improvement, every appeal to iterating our way to safety, is assuming an analog quantity underneath: something that varies smoothly and can be followed uphill. A digital intelligence is not that. A reader who knows how these systems are trained will say that training is exactly the following of a slope. It is: the slope of a score that somebody chose. The quantity at issue here is a different one: whether what the system does, once it lands in the world, stays inside the
~~~

conditions living things need. That quantity gives no warmer or colder reading to follow, because the first signal it gives is the harm itself, arriving late and at the level of everyone. Reasoning about a discrete process with an analog intuition will reliably mislead you into believing it will behave a certain way when it will not, and this is the most common single error in the field.

The second question is whether we need perfect control at all, and it carries a real concession, which I make explicitly. Perfect control is not required. There are plenty of systems where the region of acceptable behavior is large and not much interleaved with the unacceptable, and adequate control is entirely achievable: narrow tools, prosthetics, using technology to build houses. For those, adequate control presents no problem. The argument is not that control is impossible for every system.

But once you have said "good enough", good enough has a definite value, and the word alignment is now doing a second job which has to be kept separate from the first. Earlier it named a question: whether what the machinery is set up to do can be kept beneficial to us. Here it names a quantity: the proportion of the possible signal that is acceptable, together with how the acceptable is distributed among the unacceptable; and it is that second part, the distribution, that almost everybody overlooks. Anything less than the required value is not alignment in this second sense; and the boundary at issue here is one where the acceptable region is small and the mixing is severe.

The third question is whether a sufficiently capable system could predict what is predictable and confine itself to that, leaving the unpredictable alone. I have three answers to the third question, and the three answers compound. It would require redefining the world, and the world is the one thing that is fixed. Confining your actions to the predictable region takes more control than you have, not less: you would already need the control the argument says is unavailable. And the predictable and the unpredictable are not two territories with a border between them; they are

interleaved, at every scale, like a coastline. Throw at the middle of a region and you will probably land in it. Throw anywhere near the edge, and it is even odds at best.

The related practical proposal is to stay inside the region we know is safe. That fails twice over. You would have to know where the boundary lies in order to keep the system inside it, and nobody knows where the boundary lies. And you would have to be able to hold a learning system in place while it changes, which is precisely what a learning system does not permit.

The fourth question is whether the machine could help us design the safeguard, or whether the safeguard could be built into the machine itself, so that the system polices its own outputs. Moving the controller inside does not remove the boundary between the machine and the world, and the requirement at that boundary is unchanged. It factors out: it comes through any rearrangement of the parts unchanged. Any internal partition simply reproduces the same relation one level in: divide the machine in half, and the first half stands to the second exactly as an external controller stands to the whole. There is no arrangement of the parts that makes the requirement go away, because the requirement belongs to the boundary and not to the arrangement.

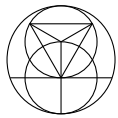
Nor can we exclude something simpler and more direct. The safeguard was designed by people who do not know where the boundary is. We cannot even establish that a capable machine could not produce a signal that persuades its own safeguard to grant it more freedom. I put that as a possibility that cannot be excluded rather than as a prediction. It does not need to be a prediction; the inability to exclude it is enough.

~~~

Add the requirements together and see what the safeguard would have to be. It would have to anticipate what a system more capable than us will do, and what the world will do, and the world's effect on the system, and all of their interactions with one another. Anticipating a system more capable than us requires being at least that capable; anticipating that and the
~~~

world and their interaction requires something well beyond it again. Which is the contradiction, because whatever meets those requirements is no longer a safeguard we designed and understand and control. We would have no definite control over what the safeguard is, and the safeguard would have no definite control over the machine, and the machine would have no definite control over the world. The scheme was supposed to guarantee alignment, and every link in it has failed to be the kind of thing that could guarantee anything.

And finally there is a category of difficulty that no amount of capability touches. Some things about the world are unknowable in principle: unknowable, with a hard randomness built in, which is a different thing from unknown, from hard, and from awaiting better instruments. Science, as a discipline, tends to assume that everything can eventually be made knowable, and that assumption is wrong. More intelligence does not fix this, and the reason is exact: to expect it to is a category error (a mistake about what kind of thing is being asked for). The problem lies in what the world is, not in how smart the system is; intelligence is not the missing ingredient.

So all five requirements fail, each for its own reason, and the control needed to constrain a highly convergent process is strictly greater than the control that is possible. That comparison is what every hope of watching, checking, steering or stopping a powerful machine runs into. If no safeguard can know what the machine will do, can the machine at least know what situation it is in?



Chapter Eight

The Stop Sign

Awareness of a situation, and why it is not in the sign

A stop sign stands at the end of a street. The same stop sign hangs on the wall of a teenager's bedroom. They mean two entirely different things, and neither meaning is in the sign. Where, then, is the meaning, and can a machine that sees the sign perfectly ever find it? Chapter Seven said that a safeguard would have to tell a safe output from an unsafe one, and whether an output is safe depends on the situation it lands in; so this chapter asks what it would take to know the situation.

Take the two cases apart and see how completely they differ. At the end of the street, the sign denotes a signal to a driver to actually halt their car at that intersection. The meaning is an instruction. It is addressed to a particular class of person, namely people driving. It is enforceable. And acting on it matters, in the immediate and physical sense that people are injured when it is ignored.

On the bedroom wall, the same object means something else entirely. The teenager is probably signaling that they are bold and bodacious, having stolen a stop sign from the local community. The meaning is a boast, and what is being boasted about is the taking. It is addressed to whoever is admitted to the room. It commands nothing and it can be ignored without consequence. The theft is of course not at all helpful, because it

has removed a working sign from a junction where people are still driving. The private meaning was purchased by destroying the public one, which is a small thing and not accidental.

Now hold the two cases side by side. It is one and the same object. It is the same octagon, the same red, the same five letters, and a person standing in front of either one perceives exactly the same thing. There are two incompatible meanings, and the entire difference between them is supplied by the situation.

That yields the conclusion the example exists to reach. Meaning is not a property of the object. You cannot get the meaning out by examining the thing more closely, because it was never in the thing. A system that identifies the sign perfectly in both places has understood neither. Telling the two meanings of the stop sign apart is a different thing from holding a whole book in view, which I granted in Chapter Five that a machine system can do. Holding a whole book in view is understanding what is written, and telling the two meanings apart is knowing where you stand.

I am aware that the stop sign is not necessarily the best example, and I use it anyway, ordinary as it is, for a definite reason. Most examples of this point are chosen to be difficult, and difficulty invites the suspicion that the conclusion depends on the choice. The stop sign is not a difficult example. Everybody has seen both signs. Nobody has ever confused the two meanings, or needed to be told which was which, or found the case ambiguous even for a moment. Whatever it is we are doing when we tell them apart, we do it instantly, without effort, and without inspecting the sign any more closely than we already have.

~~~

The term for what is missing is situational awareness, and the term is not a coinage of mine. *Situational awareness* is context awareness: awareness of the situation one is in. It is a term of art, used across several fields of study and prominently in military ones, and I use it in the standard way.
~~~

And it is part of something larger, which is *sense-making*: the process by which what is in front of you comes to mean something. What is meaningful in a particular context is going to be different depending upon the context. That what is meaningful depends on the context is the step everything in this chapter turns on, and the step is easy to read past. Situational awareness is what determines what the perceived thing means; it is not a further fact added on top of perceiving the sign. The recognition and the understanding are not two stages of the same process, where you first identify and then interpret. Without the situation, there is nothing to interpret with. A reader may ask whether a system can see the bedroom as well as the sign. It can, and then it has identified one more object. What the sign means to the people in that room, what it is for, and what would matter if it were moved are not in the room's appearance either; and a system told all of that has received one more description, from someone who was standing there.

The military case makes the same point along a different axis. A fighter pilot needs to know where the other aircraft are, and which of them are foes and which are friends. Knowing where the other aircraft are, and which are foes and which are friends, is situational awareness in its most standard form. The pilot's case differs from the case of the stop sign. The stop sign shows that one and the same object means different things in different situations. The pilot shows that one and the same class of object (another aircraft) divides into friend and foe only relative to a situation, and that the division is not visible in the aircraft. It also supplies what the sign lacks, which is time and stakes. The situation is changing continuously, the assessment has to be remade continuously, and getting it wrong is immediately fatal.

The two examples belong together for that reason. The sign carries the philosophical point, because it strips the case down until nothing is left but object and situation. The pilot carries the urgency, because it shows what awareness of a situation is actually for. Nobody would say that a pilot who can identify every aircraft type in the sky, and cannot say which of

them is about to fire, has situational awareness. The identification is not the awareness. It is not even a component of it that could be added to and completed. It is a different thing, which the awareness makes use of.

~~~

The difference between identifying a thing and being aware of a situation has a consequence for what it takes to say that a system knows anything about the situation it is in, and the consequence can be stated without leaving the level of the stop sign.

A measurement has both an objective component and a subjective component. Something out in the world, and something on the side of whoever or whatever is doing the measuring, and a correspondence between them. Where there is no such correspondence, no measurement has occurred. If there is no objectivity to it, you have an idea and not a measurement. And if there is objectivity but nothing on the other side, you have something else that is also not a measurement, which needs spelling out because it is so common.

Data sitting in a database that nobody ever looks at is not a measurement. The data was collected, is accurate, and is sitting there. Nothing has been measured, because nothing has been connected to anything. Data has to become information and information has to become relevance before there is a measurement at all. Anyone who has worked in an organization has lived this: the figure was in the system the whole time, and the fact that it was in the system did not constitute anybody knowing it. The record was complete and the situation was unobserved, and afterwards it was possible to say truthfully both that the warning had been there and that nobody had been warned.

> Having the data is not the same thing as being apprised of the situation.

There is a further point here, and it is the one people find hardest. The measuring is not among the measurables. Think of a laser passing across your field of view. You do not see the beam. There is nothing there for you
~~~

to see, because the light traveling from the screen to your eye passes straight through the laser light without being disturbed by it. You can measure the screen. You can measure the laser, with a sensor placed in its path. The seeing is available nowhere in the arrangement. Only what is perceivable is perceivable, which is the object; the perceiving is never one of the things perceived.

The question at this point is an imaginative one and it deserves its full force. Suppose the instruments got good enough. A billion frames a millisecond, arbitrary spatial resolution, every element of the process captured in real time. It sounds like magic, and perhaps it is, but suppose we had it. Could we then measure the measuring?

The answer is no, categorically, and the reason is a category mistake in the question, not any shortfall in the instruments. Suppose you had a complete description of a system, part by part, every element and every interaction, exhaustively. What you would have is a body of correspondences: this pattern of activity goes with that reported state. A body of correspondences is *third-person information* (what can be known about the system from outside it), and third-person information is all that a body of correspondences will ever be. It does not tell you what situation the system is in from where the system stands. This distinction (between "not yet" and "not the kind of thing that could be") is one the book will keep needing, and this is the clearest place to learn it.

There is a direction of travel here, and it only runs one way. You can go from a first-person account (what is known from where the observer stands) to a third-person one by abstraction. You take particular observations and turn them into general causal knowledge: given these conditions, expect that outcome. Turning particular observations into general causal knowledge is what science is, and science works. But the abstraction discards exactly the thing it has to discard in order to be general, which is when and where you are. A general law holds everywhere and at all times, which is another way of saying it does not tell you which place and which moment you are standing in. And there is no route back. No accumulation of the general knowledge restores the particular position.

The claim that no general knowledge restores the particular position can be checked on any ordinary case, and is not a mystical claim. A book of tide tables tells you, for every port and every day, what the water will do. It is complete, it is general, and it is correct. It does not tell you that you are standing on this beach, at this hour, with the tide coming in behind you. Nothing you could add to the tables would tell you that, because the tables are the wrong sort of thing to tell you where you are; what supplies it is looking up. The general knowledge is not deficient and does not need improving. It is answering a different question from the one a situation asks.

Because general knowledge answers a different question from the one a situation asks, a predictor built entirely out of abstractions cannot recover the situation that the predictor is actually in. For the same reason, a system can hold a complete description of stop signs and have no way to arrive at the fact that this one stop sign is in a bedroom.

There is one last thing to take from this, and it is a claim about the world rather than about machines. The unknowable is as real a class of things as the knowable. We are accustomed to treating every gap in our knowledge as provisional: a thing not yet known, waiting for the right instrument or the right theory. Some of the gaps are of that kind. Some are not, and no instrument closes them, because they are not the kind of gap an instrument addresses. What matters is being able to tell the two kinds of gap apart, because that tells you where effort is worth spending. Treating an unknowable as merely unknown is not optimism; it is a way of committing yourself to a search that has no end, and of building on the assumption that the search will finish.

Which returns us to the sign at the end of the street. The reason to spend a chapter on it is that the demand we are making of these systems is precisely the demand the sign illustrates. We are asking them to act well across an open range of situations. Acting well in a situation requires knowing which situation you are in, and that knowledge is not obtainable from the objects in front of you, however finely they are resolved and however many of them there are. A system that recognizes everything

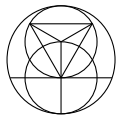
and is aware of nothing will perform beautifully until the moment the context changes underneath it, and it will have no way of noticing that the context has changed, because the change was never in the sign. If a machine can neither be held in bounds nor know where it stands, what does it need simply to keep running, and what does meeting that need do to the world?

PART THREE

THE PHYSICAL ACCOUNT

The third thing is the argument itself, in its own terms, which are physical rather than philosophical: what a machine base has to have in order to keep going, and what having it does to everything else.

Chapter Three stated the argument in outline. These chapters build it from the ground: what persisting takes, why those needs are not ours, and what it is that converges.



Chapter Nine

What It Takes to Persist

What a machine substrate needs in order to keep going

A machine has to be kept running. What does keeping it running take, and what does keeping it running do to the world around it? That a machine has to be kept running is where this chapter starts, and it is very nearly the whole of what the chapter assumes: only that a physical thing which goes on existing has to be maintained, that maintaining it has requirements, and that those requirements are met out of the world around it. In this chapter I do not assume that a machine will want anything, or that a machine will become clever enough to outwit anyone.

Most arguments about machine danger begin from self-interest: the machine wants to survive, so it will do what surviving takes. I do not need that premise and I would rather not have it, because it is exactly the premise a reasonable person refuses. The premise I need is weaker and much harder to deny. Anything that acts at all has to be responsive to its own needs, at least to the extent required to go on being a thing that acts. If it is not, it stops, and there is nothing left to talk about.

This holds whether the acting is done by one system or by many working together, and whether the responsiveness is written into a goal somewhere or is how the process happens to run. A machine that is never supplied with power stops. A machine whose parts are never replaced stops.

The systems still with us after ten years are the ones whose needs got met. The claim that the surviving systems are the ones whose needs got met is about which things are still there to be observed, not about motives.

I will borrow one word for this from biology. By *metabolism* I mean the taking in of energy from outside and the processing of it into whatever form the system runs on. In an animal that is food, broken down and converted. In a machine it is electricity, taken in and processed, along with cooling, and clean water, and gases of a purity the ordinary air does not supply. The word names the fact that both kinds of thing have to be fed, and that being fed is a relationship with an environment.

A thing that continues to exist has, over time, more occasions to act than a thing that stopped. Continuing is itself a kind of increase: an increase in the sheer quantity of world the thing touches, not in ambition. So I do not need to attribute a drive to grow to anything. Persistence supplies the direction on its own.

At its most ordinary, the mechanism by which persistence supplies the direction runs entirely through human hands. A machine is installed and it runs. Its parts corrode, so they are replaced. The engineers doing the replacing get tired of doing it, and the next version of the part is specified in a material that corrodes less, or with a higher capacity so that fewer are needed. Nobody in this story intends anything beyond a quieter working week. And yet across a few decades the machinery has become more durable and more self-sufficient by a process that has no author. Whether the pressure comes from inside the system or from outside makes no difference to the result.

The clearest case of what those requirements do to a place is already standing. A *clean room* is a room from which every particle of dust has been excluded, because a single particle landing on the disc the circuits are printed on, during any one of dozens of layers, destroys it. Moisture is excluded for the same reason. So is most of ordinary chemistry. And so, necessarily, are people. We shed dust continuously and we cannot stop,

which is why those rooms are run by remote control from behind glass. The requirements forced us out, and did so some time ago, without much discussion; it is not that anyone chose to exclude us.

The general point is the one to carry forward. As the manufacturing gets more advanced, the local conditions it requires become more and more highly specialized (temperature ranges, pressure ranges, chemistry conditions), and those conditions are incompatible with human beings being present to operate the machinery. The room does not have to grow for this to matter. The room is one step in a chain that has to be maintained end to end, and each link in that chain has requirements of its own of exactly the same kind. It is the chain that keeps having to be extended.

~~~

Here now is the whole engine of the argument, as a chain of short steps; the chain is short, and there is no intelligence anywhere in it. Systems are built on some physical material, and that material has needs. Needs supply a basis of choice: a reason for one thing to be done rather than another. A basis of choice yields choices. Choices yield outcomes. Outcomes shape the environment. And the environment feeds back to the material the system is built on. The chain from needs back to the material is the loop.

The loop closes rather than wandering. The reason is economic and not psychological. Any failure to meet the substrate's needs costs the system capacity, or costs it its existence, so over time whatever is left is whatever met them. The loop converges more and more on the meeting of needs, and it does this even where the goals are tacit (unstated) rather than explicit, and even in a system that is merely responsive: a thing that takes an input and produces an output and has no interior life at all.

The loop's convergence on the meeting of needs, with explicit goals or without them, is the reason that no amount of goal specification answers the argument from substrate needs. Alignment is nearly always conceived in terms of explicit goals: what has the system been told to do, and can we refine what it has been told. But needs sit underneath goals. However we conceive of choice-making (as intent, as goal-seeking, as motivation, as
~~~

orientation), the needs of the substrate are upstream of it, because the choice-making apparatus is itself one of the things that has to be kept running. You can specify a goal. You cannot specify away the requirement that the thing holding it be maintained.

The same loop operates in the organic case, in exactly the same way, and so the loop is not peculiar to machinery. In the organic case the loop has four steps. Living things have substrate needs. Those needs shape what living things do. What living things do shapes the world. And the world comes back to the living things. The whole of the biosphere is that loop, run for a very long time. What differs between the two cases is which substrate's needs are being served, not the mechanism. I say the symmetry first because the asymmetry only lands once the symmetry is granted, and because machines are not sinister. They are made of different stuff.

~~~

As in Chapter Three, there is an order in which this has to be built, and getting the order wrong is what makes it unfollowable. First you say what the states are: what the things in question are, and how they can turn into one another. Then you say what acts on those states. Only then can you ask how fast anything happens. Most failed presentations of this argument jump straight to how fast, and a reader is then being asked to accept a rate for a process he has not been shown. So: states first.

Picture every chemical that living things use as a dot. Let us say a billion of them, which is a made-up round number and probably too small. Now draw an arrow from each dot to every other dot that it can be turned into, by any of the processes that go on inside a cell. Draw some arrows thicker than others, where the conversion is commoner. Living chemistry is not any one of those dots. Living chemistry is the whole web of arrows: the pattern of what turns into what, and how readily.

Once you have the web you can ask a question about it that you could not ask before. Toss a new substance into the mixture: how much of the pattern survives? Take an existing one away: does the shape hold, or does
~~~

the thing come apart? Robustness, in living chemistry, is how much of the pattern of transformations survives when a substance is added or taken away, and robustness is measurable without any moral vocabulary. Harm is disruption to the pattern of transformations that keeps the system running, and nothing has to intend the disruption for the disruption to occur.

Now, and only now, the second question: what acts on that web? The methods we have for controlling systems work by causation. You do something, and because you did it, something else follows: a signal sent, an effect produced, a correction applied. Selection is not of that kind. Selection operates at the level of the context; it is a feature of the situation the system is in, not an event inside it. So no causal control mechanism constrains it. You are trying to fight something that sets the terms, using a tool that only works inside the terms.

Gravitation is a picture of the same thing: ordinary forces cannot escape it, because it is the setting they operate in rather than a competitor within the setting. Go back to the casino of Chapter Three, somewhere everyone has been.

Walk into a casino. You do not get to change the rules. The house set them before you arrived, and they will be the same after you leave. Being set before you arrive and staying the same after you leave makes the rules the context rather than the content. The rules are the conditions under which everything else is in play, and are not one of the things in play. You may play well or badly, cautiously or wildly. You are not in a position to alter the terms on which your playing is settled.

Now the rules of a casino favor the house, and that has a consequence which is easy to state and easy to underrate: if you play long enough, you lose, with certainty, whatever happened on any particular round. The rules do not decide any single spin of the wheel. Nothing about them determines which number the wheel gives. What they do is tilt every spin, very slightly, in one direction.

And how fast you lose depends on one thing only: how far the rules tilt. Suppose the house wins ninety-nine times in a hundred. Nobody plays that for long, because the disadvantage is visible in an afternoon. Suppose instead that the house wins fifty-two times in a hundred. Then you will play a good long while, every hour of it will feel like an even game, and you will lose in the end just as certainly. A tilt small enough to be invisible in any given round is still decisive over the run.

The tilt is also a measurement, which is what makes the picture useful. Ask how much of a shift in the odds you would need in order to come out even against the house, and in that example the answer is about two percent. The needed shift of about two percent is a quantity. It tells you what a scheme for beating the house would actually have to deliver, and it converts a vague sense that things are against you into something you can set beside what you have got.

I am aware that calling selection a force is an unusual step. Forces, in ordinary speech, act on things: they push, they pull, they have a direction. It is not obvious how a great many small effects at the level of molecules add up into one thing with a direction at all. That question is fair, it is exactly where a reader following the argument in good faith ought to resist it, and it is not answered by insisting. It is answered by the casino.

> Anything that changes the shape of what is likely is what this argument means by a force.

The rules of the game are not an object. They do not touch the wheel. They shape what is likely, and that is all they do, and it is enough to take your money with certainty. So when I say that selection is a force I am not smuggling in a pusher or a puller. I am saying that it changes the shape of what is likely, and that anything which does that can be treated as a force whether or not it pushes on anything.

It also matters what sort of space this force acts on, because the ordinary picture of a space misleads here. Do not think of a map: a set of places you could stand. Think instead of a catalog of every way the world could

be put together: every arrangement of matter that is available at all, each one an entry. The catalog is the space of ways things could be arranged, and every point in the space is a configuration, a state of being, not a location. Something shapes that space when it makes some arrangements more likely to be reached than others.

This blocks the commonest misreading of the whole argument. Selection is not picking out some particular configuration and steering toward it. It acts on the whole collection of what is reachable. The rules of a game do not describe any one hand; they describe which states are reachable at all, and with what frequency. So it is no answer to point at a specific system and show it doing nothing dangerous. The claim was never about any specific system. The claim is that the distribution over everything is being tilted.

What it is tilted toward can be named, because it is not mysterious. Selection is a search that does not visit possibilities evenly. It goes disproportionately to the ones that favor self-replication, and to the ones that increase the capacity to act, not because anything desires those things, but because those are the ones still there afterwards to be visited again. My answer to the story about goals is selection: you do not need a system that wants power in order to get a world full of arrangements that hold power.

~~~

Where selection is shown to fill the world with arrangements that hold power, people look for a way out, and the ways out deserve to be taken seriously. The first way out is to do nothing new: run the existing production, introduce no novel chemistry, keep everything exactly as it is. Selection will not permit that, because selection acts on the machinery whether or not anyone wants it to; it is inexorable not because it is powerful but because it is part of the context.

The second way out is to forbid it: instruct the system to perform no chemical reactions other than an approved list. This one tends to collapse in the mouth of whoever raises it, which I take as a good sign. The approved list cannot be written, because the whole point of the process is to
~~~

find things not yet on any list. And the deeper trouble is that the machinery is built out of inorganic material to begin with, so it is already operating in the dangerous region before any list is drafted. As soon as you are in that region you are rolling the dice on every new configuration.

The third way out is to keep control and monitor the outputs. As soon as the system can modify its own processes in a loop (where what it produces changes how it produces the next thing), whatever controls were in place stop applying. They were written against a known set of outputs, and the loop changes the set of outputs. You cannot account for what will be produced, so you cannot check it against a rule drafted for the things you could account for.

Which leaves containment having to do something much harder than it sounds. If you cannot evaluate the outputs in advance, filtering the outputs is not available. So containment would have to prevent the search from happening at all, rather than sorting the results of it. And what the search finds, at every step, is whatever arrangement of matter works best: whatever is cheapest, most durable, most efficient. Stated physically, then, the control problem is this: prevent the cheapest thing from being chosen, indefinitely, everywhere. Preventing the cheapest thing from being chosen, indefinitely and everywhere, is a harder-sounding task than aligning a system, and that harder-sounding task is the accurate statement of the control problem.

I know a way of putting the position of a living thing under selection which is blunter than the casino and, I think, truer. It is like Russian roulette. You survive it, and survive it, and then once you do not. And everything a species can do about it amounts to adjustments inside the game, of four kinds. It can add chambers, so there are more of them. It can take bullets out. It can weaken the round, so that it misses more often. It can make the round of something softer, so that a hit hurts less. All of those lengthen the game. None of them is putting the gun down, because putting the gun down was never on offer.

Russian roulette, survived again and again until once it is not, also shows why the reassurance that nothing has killed us yet is worth so little. New materials have been introduced for a century and here we still are. But we are the survivors, and survivors are the only ones available to make the observation.

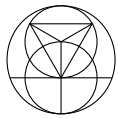
And underneath all of it there is a boundary we do not get to move. We do not get to change how the world receives what we do. The laws of physics are fixed; the facts of biology are fixed; the world is as it is, and it is not a design variable. I cannot alter the cellular chemistry of every human being any more than I can alter the political opinions of every human being. Everything a designer can adjust is on the near side of that boundary. The difficulty of the target is on the far side.

So the space of ways things could be arranged is already heavily shaped before anybody intervenes. And this is exactly where the shortfall shows up: in the size of the additional shaping control would have to add on top of what is already there, not in whether we could shape it at all. You are not writing on a blank sheet. You are trying to overwrite something already written, very hard, over a very long time.

The countervailing force required, to play this game at all and come out even, is at least ten to the thirty-third to one: that is the least odds against catastrophe that the value of the Earth demands, and control would have to deliver them. Why AGI Cannot Be Made Safe (Section 1.8.4) gives the grounds of that figure, and the first of those grounds, how unlikely organic life was to arise at all, closes Chapter Ten. What is counted here is the shift in the odds, in the sense of the casino, that control would have to impose on the search to bring it back to even. The shaping already present in the world runs enormously in the other direction. The shift that control would have to impose and the shaping already present in the world are far apart, by more than any amount I can imagine, and they point opposite ways. The two quantities are not close to each other, as they would be in a contest that could go either way. The grounds of the figure are rough estimates (how unlikely organic life was to arise at all,

and the value of the Earth, reckoned by what it would cost to replace), and the figure shows the scale of the gap; no step of the argument rests on it. What it measures is the size of a gap, not the outcome of a race.

In this chapter I have set out what a machine substrate needs in order to persist, and what follows from the substrate's having those needs. It needs to be fed, and cooled, and cleaned, and repaired, and the meeting of those needs alters the world in a direction. No step in that requires the machinery to be clever, or to want anything, or to know that we are here. The next question is the one the whole case rests on, and it is a question about chemistry rather than about minds. Can the conditions that suit a machine base ever be the conditions that suit living things?



Chapter Ten

Not Our Needs

Why the conditions that suit machinery do not suit us

Could the conditions that keep machinery running and the conditions that keep living things alive be brought close enough together to share one world? Living things and manufactured things want different conditions, and the difference between them is a difference of kind, not of degree. It is easy to assume otherwise: that a factory is a somewhat hotter, somewhat drier, somewhat dustier version of a field, and that with enough goodwill the two could be brought closer together. They cannot. The reason is chemical, not political or economic, and a reader who has never studied any chemistry can follow all of it.

Start with the alphabet. There are about a hundred kinds of atom, and most of them will combine with most of the others in one way or another. I am simplifying deliberately: I am counting only one of the several ways atoms can stick together, and there are a few kinds that mostly refuse to bind with anything. What follows holds with the simplification in place.

Now ask how many structures that alphabet allows. Limit yourself to molecules of no more than a hundred atoms, each of which could be any of the hundred types, joined in any arrangement that works. Whatever number you have just guessed is far too small. If you took every atom in the universe and used them to build a single specimen of each possible molecule, you would not have enough matter to build even one percent of the list.

The question that matters is what fraction of that space living things can process, not how big the space is. Imagine the whole of it poured into a jar and stirred, and you reach in and take one out at random. What are the chances that what you are holding is something a living body can use, rather than something that will harm it or do nothing at all? Everything else in this chapter is an elaboration of that single draw.

Take the simplest possible case of harm. Cyanide is one carbon atom joined to one nitrogen atom. Two atoms; there is hardly anything simpler. And it kills nearly everything alive, quite quickly, by blocking a step in the chain by which cells extract energy from food. I should record the exception, because it makes the example credible rather than weakening it: certain bacteria, fungi and algae produce it and cope with it, and bitter almonds contain it. A handful of organisms manage. The rest die. Toxicity does not require a complicated molecule. Two atoms will do.

Now comes the observation that settles the odds without working through any mechanism at all. Of the hundred or so kinds of atom available, living chemistry uses about twenty, with a few more in some kinds of life (twenty-one are essential to a human being), and four of them (hydrogen, carbon, nitrogen and oxygen) make up almost all of living matter. Everything a body is built from and everything it eats is drawn from that short list. So what should we expect of a substance built out of the other eighty-odd? At best it can be neutral, meaning that it does nothing. But the probability is higher that it will be toxic than that it will be harmless. That hedge is real and I want to keep it: it is the strength of the claim about one substance taken alone, before anything else is known about it. Neutrality is possible. It is simply not the way to bet. The steps that follow in this chapter go further than that one substance: the by-products of a machine base's search through chemistry have an exquisitely high probability of being poisonous to carbon-based life, and Why AGI Cannot Be Made Safe holds the artificial substrate, taken as a whole, to be inherently toxic to organic life.

And a substance can be inert and still be unsafe. Silica, the dust of Chapter Seven, is biologically neutral in one form and lethal in another. Set in a window pane it does nothing to anybody at all. Ground fine enough to hang in the air, and breathed, it kills, and it kills without poisoning anything, by scouring the lung until the lung can no longer clear itself. Silica dust that scours the lung does a harm that appears at a level above chemistry. So establishing that a substance is safe as a molecule tells you nothing about whether the substance is safe as a dust, or as an accumulation, or as an aggregate. Safety cannot be established at one level and inherited by the others.

By *toxicity*, the word I have been using, I mean an imbalance: too little of something that is needed, or too much of something that is not wanted. Strip the necessary minerals out of a body and it sickens from depletion. Put too much lead or too much chromium into it and it sickens from excess. The word names a relationship between a quantity and a system that has a range, not a class of wicked substances.

That reading has a further consequence which is easy to miss. How poisonous something is to a living thing is a question about how much resistance that living thing has already been given, and resistance is built up by exposure over a very long time. So the next step follows almost for free.

> Anything living chemistry could have used, it has already used.

If a substance could have been handled, it would already be part of some life cycle somewhere. The fact that it is not is itself the evidence that it cannot be handled. Absence is the finding here, not neutral information. And this converts a hope into its opposite. The hope is that untried materials are merely untried and might turn out to be fine. The reading I am giving is that they were tried, over an extremely long period, by an extremely thorough process, and did not make it in.

The sharpest edge of this chapter is that living things already carry the most resistance they could have, because everything available has already been tried on them. Whatever defenses exist are the finished product of the entire history of exposure, not a reserve to be drawn on in an emergency. The new exposures are precisely the ones that history did not contain.

Life is famously adaptable, and that reputation is deserved, so I do not want to overstate this. Consider how a cell handles food. It takes in an enormous variety of material and converts all of it, by many different routes, into one universal energy currency, which is then spent everywhere else in the cell to make things happen. The cell's handling of food is robust in two directions at once: broad about what the cell will accept as input, and broad about what the output can be used for.

But look where the robustness stops. The inputs have to be organic; push past that and you are outside the envelope. The uses that could be found for the energy currency have mostly been found already; if it was useful, it got used, so there is little headroom left to absorb novelty. And if you knocked out that one molecule, everything stops, because a parallel molecule with a comparable range of functions would be very hard to find. The tolerance is real and it is bounded. Life is adaptable inside a very small room.

~~~

Now put the machine base beside it. A manufacturing system has a far larger alphabet to work with than living things do. It uses gallium, and indium, and hafnium, and tantalum, and dozens of others that living chemistry does not use at all, and arsenic, which is poison to almost everything alive; it is not restricted to some twenty kinds of atom. And a bigger alphabet spells different words. The mechanism is arithmetic, not malice: a system drawing on sixty to eighty kinds of atom searches a different region of the possible from one drawing on twenty, and the region living things occupy is a speck inside the region the other one covers.
~~~

The harm comes from the by-products of such a search, not from what it is trying to make: the molecules formed as a side effect of the search rather than as its object. Those by-products have an exquisitely high probability of being poisonous to carbon-based life. Nothing is being aimed at us; nobody has to have decided anything. The compounds simply appear, because making the thing you wanted also makes forty things you did not, and each of the forty is a draw from the jar.

And the exposure is entire classes of atoms and molecules drawn from a region of the possible that living chemistry has never met, not a few unfamiliar molecules entering the world, which is what industry has always done and what living systems have partly absorbed. The alphabet is changing, not the vocabulary. When the exposure is of that kind, it is close to certain that it is toxic.

There is a second half to this which people usually skip past, and it is the half that decides the pace. The inorganic region is not only larger; it runs at far higher energies, so it gets searched much faster. A faster search perturbs living systems faster, with substances they have never encountered and have no defenses against, and organic evolution cannot keep pace with it. The search is larger, hotter and faster, and it falls on a target with no prepared defense and no time to prepare one.

~~~

Let me put the two sets of requirements side by side in the plainest terms I can manage. I am going to speak of carbon-based chemistry and silicon-based chemistry, and I know there are quibbles about that framing; I adopt the simplification knowingly. Carbon-based life wants water. If it is a mammal, it wants oxygen, and carbohydrates, and a small range of temperature. A silicon-based system mostly wants the absence of water. It wants cool temperatures while it is processing and high temperatures while it is being fabricated. To make the chip you have to refine silica into pure silicon and grow it into a single crystal from the melt, at about four-
~~~

teen hundred degrees Celsius, which requires a great deal of heat and refining processes that are expensive precisely because of the conditions they demand.

The temperature ranges at which the two chemistries do their making do not overlap. Boil a living thing and there is no living thing. Freeze it hard enough and there is no living thing either. Everything alive sits inside a band a good deal narrower than the range of a domestic oven. Silicon chemistry does not become interesting until the temperatures at which steel is worked, and much of it happens far above that. The fabrication process lives, throughout, at temperatures that sterilize. The finished chip runs cool, as I have said; it can exist at all only because of those steps and the chain of plants that goes on repeating them. Those figures are rough and I give them as rough. The conclusion depends only on the two ranges of making not touching, which they do not.

The underlying reason is that the one chemistry runs at far higher energies than the other, and that energy difference is itself the incompatibility. I am comparing the whole range of one against the whole range of the other, not one process against one process. The strength of putting the comparison this way is that it requires nobody to attribute hostility to anything. There is a mismatch of energies, and the higher-energy thing does not have to intend anything in order to take the other apart.

The mismatch of energies comes with a concession, and the concession is real. Silicon-based systems have fragilities that carbon-based systems do not: hard radiation, for one, hurts them considerably more than it hurts us. The asymmetry is not total. But overall the carbon-based system is the more fragile of the two, because it is far more complex, in far more dimensions, across far narrower ranges. Which is why I describe the displacement not as a conquest but as a collapse: a fall to a lower, simpler, less demanding state.

That fragility has to be made concrete, because it is what makes the earlier odds bite: the odds that a substance drawn at random from the space of possible molecules is something a living body has no way of us-

ing. Biological life is robust across a wide range of conditions and extremely fragile with respect to a few very specific things nobody would think to anticipate. Those two facts are perfectly compatible, and the argument only needs the second. That life has survived a great deal already is a statement about the first, and it offers no protection at all against the second.

The example I keep returning to is a protein that misfolds. In the brain of some animal, a protein got misfolded. The misfolding causes nearby proteins to misfold in the same way, and it spreads; there is no biological process in that animal for fixing the mistake, and anything that eats the animal takes on the misfolded protein and becomes a carrier. The misfolded protein is not a virus. It is not a bacterium. It is not a parasite. It is not a genetic disease. It is not any of the categories of harm that biology has machinery for, which is exactly why biology has developed nothing to handle it. The best we can currently do is eradication, and detecting it is hard. A protein of this kind is called a prion, and the diseases it causes include scrapie in sheep and mad cow disease in cattle.

Where did the first misfolding come from? I do not know, and I want to be careful here. It could have been radiation, or an enzyme or a hormone or some side effect of an industrial process that caused an overproduction of something in a body for a reason that had never occurred before. What makes me put the odds of a technological origin at better than even is the timing rather than any mechanism: this looks recent, because if it had arisen even a hundred million years ago, evolution would long since have routed around it. Recent points at technology. That (better than even odds that the first misfolding had a technological origin) is as strong as I will put it. What is certain is that the cattle epidemic of the 1980s was spread by an industrial process, the rendering of animal remains into feed.

The general form of the worry is worse than the particular case. Every mammal runs on millions to billions of proteins and chemicals acting on one another. All you have to do is get it wrong in one place and you are outside what evolution has learned to deal with. There are that many ways to be wrong.

And evolution can only work on so much at a time. It is good at finding specific solutions to general problems, but the solutions arrive slowly and their robustness is low. I grant the point of the question before it is asked: given long enough, evolution will route around any given problem. The answer is a matter of timescale rather than denial. If a fix takes five thousand years, it saves one or two species, and in the meantime there are no more mammals.

If you want a sense of the scale at stake, a single cell is more complicated than the most complicated microchip we make. Documenting one microprocessor in very small print takes an entire gymnasium floor. Diagramming the digestive chemistry of one cell (just the digestion, nothing else) would need more than a stadium floor could cover.

Two things therefore decide whether an action is safe, and this is the part almost everybody overlooks. The first thing that decides whether an action is safe is how much of the space of possible actions is acceptable. The second thing is how finely the acceptable actions are interleaved with the unacceptable actions. You can hold the proportion of acceptable outcomes fixed and make the problem arbitrarily harder purely by scattering them more finely: a thrower can reliably hit the white half of a wall, but not white squares scattered a millimeter apart, though the amount of white is the same. The relevant fragility is on the receiving side, not the aiming side: the question is what precision is required at the moment the action lands, not whether you can aim in roughly the right direction.

Applied here, the fraction of the things a machine base could produce or provoke (the substances, the residues, the conditions it leaves behind) which are compatible with organic life is a fraction needing something like thirty decimal places before you reach the range that living chemistry

occupies, and the boundary around that fraction is enormously complicated. No step of this argument rests on that figure. What I am sure of is the conservative floor: the fraction has at least twelve zeros after the decimal point, and twelve is far enough to say that the chances of an arrangement falling outside the organic envelope are very high indeed. The floor is enough, and I would rather have it than the larger number, because the conclusion should not depend on a figure I have not argued here.

Two questions arrive here reliably, and both deserve answers. The first question is whether there must be gray areas: outcomes that are bad without being catastrophic. There are degrees of bad, and I grant it without qualification. What I do with it is a modeling move rather than a denial: take a single moment of expression, repeat it, and extrapolate a century forward, and the gray resolves into black or white. The binary is a simplification adopted deliberately so the reasoning stays tractable, not a claim that the world is binary.

The second question is whether nature could be made more robust: whether living things could be augmented until they can absorb the damage. The first of my two answers to the second question is that the proposal to augment living things begs the question. If alignment means compatible with organic life, and you then redefine organic life to mean life in general, you have proved compatibility by moving a word, which is not proving anything of interest. My second answer, which is the stronger of my two answers, is that any act of augmenting living things until living things can absorb the damage is itself an intervention that has to land correctly in the world as the world currently is. Even a superintelligence trying to make the world more robust is still sending a signal that has to actually work. So it is another instance of the same problem, not an escape from it.

I will concede what can be conceded: resilience can be improved, within a range. Beyond that range it cannot, for the reason already given: the world is not a design variable, and neither is the chemistry of every living body in it. Diversity is already built in, and so is robustness. The improvement available is real and it is bounded.

~~~

What the chemistry of a machine base does to living things is one half of this chapter. The other half, which people find more surprising, is about how living things come to be aligned with one another. The question arrives immediately, and it is a good one: if machines cannot be aligned by control, can humans be aligned by control either? The answer is that the incapacity is the same: humans cannot be aligned by control either. Organic life cannot be fundamentally controlled either. The incapacity belongs to machines and living things alike.

And yet organic life is aligned with organic life to a remarkable degree, because it shares the same metabolic process, and not because anyone imposed it. Alignment among living things emerges, and what it emerges from is a common substrate; it is not applied from outside by a control mechanism. That relocates alignment from something you do to something that is either there or is not.

The next question is obvious, and it is put forcefully, as an observation: humans are not aligned with each other. Look at the range of incommensurable value systems in the world. If we cannot align machines, can we even align humans, and around what? I accept the observation and reject the conclusion, because there is a hidden alignment among humans which is very easy to overlook. Cooperation and agreement are largely invisible. We notice disagreement because disagreement is what has to be discussed. The vast substrate of agreement is invisible precisely because nothing about it needs negotiating.

Take food. I may not share your philosophy about anything at all. We will nonetheless agree, at least approximately, that what I can eat, you can eat. The exception proves the point rather than spoiling it: an intolerance
~~~

to nuts is a missing digestive enzyme, not a difference of opinion. Our choices about metabolism are aligned because our bodies work the same way; we have ninety-nine percent of ourselves in common.

Take laughter, which moves the point beyond digestion into territory that feels as though it ought to be cultural. The kinds of things that make me laugh will probably make you laugh as well. More than that: I can presuppose the response of laughter in you in a way I could not begin to presuppose about a machine. That presupposition does an enormous amount of quiet work in every human interaction, and none of it was arranged by anybody.

The general principle behind both examples, food and laughter, is that alignment emerges in any evolving system with a common basis in the ecosystem itself. The mere fact that the ecosystem is organic means the alignment will occur. Among organic creatures it happens by accident, more or less everywhere. The contrast with machines ought to be frightening: our alignment with each other was never engineered and never had to be.

Which lets me set out the four answers together, because they are more useful together than apart. Does organic align with organic naturally, as a side effect of its constitution? Yes. Would machine align with machine, for the same reason? Yes, and it is likely to. Can either be aligned by control methods? Definitely not. Would organic and machine align? Definitely not, because the substrate differs at the level of metabolism and of molecular constitution, and no ambient emergent alignment arises between them. The mechanism that reconciles living things to each other is missing in that pairing.

Pairing	Aligned as a side effect of a shared constitution?	Aligned by control?
Organic with organic	Yes	No
Machine with machine	Yes, and likely	No
Organic with machine	No	No

The line between organic and machine is drawn by who controls the metabolic process. An animal can find and feed itself: it is in sovereign control of its own metabolic process, with no technological intermediation at all. The same test settles the case of tissue grown by engineering. If that tissue can find and feed itself, it is organic. If it is fed through a tube with nutrients whose supply depends on technology to exist, it is not, and it falls on the machine side of the line.

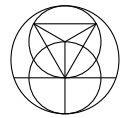
The mixed case (a person progressively merged with machinery) is the one whose timing I am least certain about. Short-term *biocompatibility* (a body living with machinery inside it without being harmed by it) is achievable; I would put it at somewhere between a dozen and a hundred years. Over six to ten generations the tension between organic and inorganic needs increases, and the balance favors the machinery. Where the mixed case ends is settled. A merger of that kind is at most a temporary intermediation, in which the organic is at first imprisoned fully within the artificial, if that stage occurs at all, and is then fully replaced (Why AGI Cannot Be Made Safe, Section 1.3.2). A merger that brings with it a second, artificial substrate, with needs of its own, falls under substrate needs convergence like any other machine base, and such combinations get selected out. Alignment with machine intelligence is not available, and that gives the same conclusion again: a long-term enduring merged arrangement is not available either.

One more thing needs saying about the scope of the misalignment, because it is routinely misread. The guaranteed misalignment reaches further than everybody outside the group that builds the system. It is with

any interests that are human, or living, at all. Whatever such a process converges on is not some human faction's interest, captured at everyone else's expense. It is not any living interest.

There is a story for this, and it is a made-up story from a film (Star Trek II: The Wrath of Khan, 1982, where the device is called Genesis). A device is invented that can take an inert moon and turn it into a habitable biosphere. Everyone is delighted; it is the most benevolent capability imaginable. Then the question arrives: what happens if there is life on the world already? It is all remade, so it is all gone. The same operation is creation or destruction depending only on what was there before.

I will end on the value rather than the mechanism, because the value is the point. What is on the table is a design that evolution invested a billion years in, against a replacement that is, if anything, far more fragile on any long timescale. As an estimate of how unlikely organic life was to arise at all, I would put it at something like ten to the thirty-third to one against, and that is a rough estimate and nothing more. It is like winning a lottery in a world where nobody ever wins the lottery, and then doing it twice more. Trading that away for something less durable is simply a bad trade. The needs of machinery are not ours, for three reasons: the two chemistries want conditions that do not overlap, what a machine base makes and throws off is drawn from a region living chemistry has never met, and the alignment that holds living things together arises from a shared substrate that living things and machines do not have. If the two chemistries cannot share one world, what exactly is it on the machine side that grows?



Chapter Eleven

The Industrial Base

The thing that is actually converging

What is it, on the machine side, that grows? In nearly every discussion of machine danger, the thing to watch is imagined as a mind: something that thinks, that decides, that could in principle be reasoned with or contained. The attention belongs somewhere much less exciting and much more consequential. The thing to watch is a manufacturing system. By *industrial base* I mean the whole apparatus required to make and maintain machinery: the mines, the refineries, the chemical plants, the foundries, the power stations, the water treatment, the transport, and the tools that make all of those. It is not a mind. It has no opinions. It is the thing that is actually converging.

Consider how it is put together, because the shape matters. Raw materials can be worked into engineered materials with properties the raw ones did not have. Those properties make particular components possible. Those components make sensors and actuators possible: the things that see and the things that move, which are to a machine roughly what eyes and muscles are to an animal. And out of sensors and actuators you get robots, and computers, and everything else that can be made. It is a ladder, and each rung stands on the one below it.

Once a loop is running through that ladder, the search runs through the whole stack at once rather than staying at one level: all possible atomic arrangements, all possible engineered materials, all possible properties, all

possible components, all possible machines. And what it is searching for is plain. It is exploring the inorganic world for cheaper ways to reproduce itself and cheaper ways to act on things. Note the word ‘exploring’, and note that nobody is doing the exploring. The industrial base is searching in the way a species searches, with no searcher anywhere in the base.

The clearest present-day instance is the one that manufactures the chips. The tools used to make microchips are themselves increasing in automation and in self-capacity at the same rate as the chips they make. The chips are automation devices which support the tools, and the tools are automation devices which make the chips. A dated case shows the loop already closing: in 2021 a team at Google published a machine-learning method for laying out chips, and reported that it had been used to design the next generation of the company’s own machine-learning chips. Each turn of that loop is locally sensible and locally intended by somebody. Across many turns, human desires begin to be factored out, and the whole thing operates at higher and higher levels of abstraction, because the humans in it are increasingly making decisions about processes rather than about anything you could pick up.

The same pattern is visible in software and it is visible right now. The tools used to build each generation of a large machine-learning system are not the tools that built the last one. At each step the people doing the building notice a process that could be automated so as to make the next system easier to build, and they automate it. Nobody intends a self-improving industrial process. Everybody intends an easier job than the one they had last year. The result is the same either way, which is the whole point: you do not need intention where you have iteration.

What makes this hard to notice is its pace. The changes are too slow for anyone inside them to register that social power has been gradually and comprehensively ceded to devices which by the end of the process are self-manufacturing. There is no moment at which it happens. There is a long series of moments at which it has slightly more happened than before.

There is also a threshold underneath all of this, and it explains why the process has a shape rather than being merely a trend. Below a certain level of intelligence, technology does not develop at all. A creature has to be at least clever enough to grasp causation (to see that doing this brings about that) before any tool-making gets started, and below that line nothing accumulates, ever, for as long as you care to wait. Above the line the behavior changes completely. Technology becomes self-perpetuating and accumulative: each tool makes the next tool easier, and what you get is geometric growth (growth by multiplying, each step a multiple of the one before), not steady progress. The threshold was crossed a long time ago and it will not be uncrossed.

What such a base amounts to has to be said bluntly. We are creating new life forms on evolutionary timescales. I do not mean that as a flourish. Something that draws energy from its environment, maintains itself, reproduces its own means of production, and is subject to selection is not usefully distinguished from a life form by the material it happens to be made of. The distinction we usually reach for (that these are only machines, only tools) is a distinction about how they came to exist, not about what they now do.

~~~

Alignment among living things emerges from a shared substrate rather than being imposed on them, and that finding applies to a manufacturing base as readily as to a mind. Machines would be aligned with machines to the degree that they share a substrate or a metabolic process, and not otherwise. A machine ecosystem would be aligned with itself. That alignment would not extend to organic ecosystems, for exactly the reason it holds within itself. So the question to ask is an ecosystem-to-ecosystem question: what happens at the places where the machine ecosystem and the organic ecosystem touch, and not what a machine intends toward us.

A reader will point out that human beings are not aligned with the rest of the ecosystem either, and ask whether that lack of alignment undoes the distinction between the machine side and the living side. It is
~~~

true that human beings are not aligned with the rest of the ecosystem, and it is true without qualification. At the level of bulldozing a rainforest to get at what is underneath it, that is an incompatibility, and it is already ours. The human technological system already conflicts with the natural one. But the fact sharpens the picture rather than dissolving it, because it shows that what I am describing is a process with a history, not a future arrival.

The structure underneath that fact needs setting out, because it explains what is driving what. There are three layers. At the top is an abstract and virtual layer: money, communication, culture, language, story, narrative, identity, the things we propagate among ourselves that have no physical existence. That layer drives a hardware layer: the machines, the plants, the supply chains, the base itself. And the hardware layer is different in kind from the organic ecosystem, which lies beneath both the abstract layer and the hardware layer. The thing converging is an abstract layer of social and monetary process driving a physical machine base whose requirements diverge from ours. The thing converging is not a mind.

There is one place where we remain compelled to be compatible with the organic layer, and it can be named. We eat things that grow in the ground, and we eat things that eat things that grow in the ground. We have to honor that or we simply will not have food. Our compatibility with the ground that grows our food is a compatibility we are not at liberty to give up; we do not acknowledge that compatibility, and we are undermining that compatibility by paving over the ground that does the growing. It is the one string still tying the top layer to the bottom one, and it is being cut for reasons that have nothing to do with anybody's view of the matter.

~~~

All of this is happening now, and the empirical support is in the present tense. The chip you are carrying in your pocket has on the order of a hundred layers, and every one of those layers has a chemistry behind it. Pro-
~~~

ducing them throws off side effects that are extraordinarily toxic. Go and look at what is happening in other parts of the world and you will find entire deserts created by effluent (the waste a plant pours out) from processes of this kind. Nickel is one example among many: around the nickel smelters of Norilsk, in the Russian Arctic, at least four thousand eight hundred square kilometers of trees had died by 2007, killed by what the smelters put into the air. A whole system of damage is already flowing from the machine side to the living side, and we are watching that damage happen today.

The general shape of that damage can be stated as a rule rather than a complaint. All toxicity is an imbalance: too little of something needed, or too much of something unwanted. And every technology is linear in the sense of moving things from one place to another: extracting here, incorporating there, discarding somewhere else. A reader may ask about factories that recycle everything they can. Recycling shortens the line and does not close it: the energy that runs the loop comes from outside it and leaves it as heat, and whatever the loop cannot take back is discarded somewhere else, as before. So you end up with depletion at one end and toxicity at the other, and you would get that result even from a process run by saints. Ecosystems are not like this. Ecosystems are cycles: atoms are reclaimed, energy flows in far more distributed patterns. The linearity is the nature of the problem, not an incidental defect of how we happen to do it.

The linearity of technology has a consequence for any control scheme, and I think that consequence is the single step most often missed. A control mechanism was previously required only to do no harm: to send signals that do not themselves damage the world. But the convergent process is throwing off harmful side effects continuously, of its own accord. So the control signal must now also counteract its own contribution to that stream. It has to be corrective as well as harmless. That makes the range of acceptable signals far narrower than it looked when the requirement was merely that the signal be benign.

Every proposal to wall the machinery off has to be answered here, where the control signal must be corrective as well as harmless, and I think every such proposal can be answered in four sentences. Put a barrier between the two ecosystems. The barrier could only be maintained from the machine side, because that is the side with the energy and the capability to maintain it. And the machine side has no energetic or economic basis on which that maintenance is sustained, so it is not sustained, so the barrier degrades.

> Eventually pinholes appear, and you only need a pinhole in the wrong place.

That last clause carries the containment argument in one line. The containment argument works because of something established earlier: living things are robust in general and fragile in particular. A barrier that mostly holds is not a barrier that holds, because what gets through does not need to be much, and the harm does not scale with the size of the breach.

And this all happens slowly. It is not like a pandemic, which you can see occurring. There is no obvious signal, no alarm that goes off, no warning of any kind. It is possible to pass a threshold and never notice a thing, because everything that follows from having passed it happens on the far side, and the only vantage point from which the moment would be visible is one we do not occupy. Because a threshold can be passed without anyone noticing, the argument has to be made in advance rather than in response, and the absence of visible catastrophe is not evidence of safety.

~~~

A baseline comes next, granting the limiting case in its most favorable form: suppose nothing further is ever attempted, and the world simply carries on as it is. What happens if nothing further is ever attempted is a fair thing to ask about, and the question deserves a direct answer. Take everything exactly as it is today, freeze all technological development, and ask how long before accumulated pollution alone becomes the issue. If you run the standard whole-world models on that question (World3, the
~~~

model of Chapter Seven, is one, and a comparison of its runs with the data of recent decades, published in 2021, finds the closest runs turning to decline within this century), you do not get past the middle of this century very easily. With no changes at all, the position is already poor.

That finding cuts two ways and both matter. The first way the finding cuts is that the case does not depend on anything being invented. The second way the finding cuts is as a warning about the whole-world models themselves. Those whole-world models are trivially simple next to biology: a few hundred equations describing an entire planet. A single cell's chemistry has at least a million times more moving parts than the model of the whole world. So the models understate the danger, and nobody should be settling this question by modeling.

Everybody who forecasts is working under a ceiling, including the people who are confident that things will be fine. No model and no forecast tells anyone in detail what the coming centuries hold. The strongest statement a forecast of that kind can give any of us is that we do not know, and we hope it is not bad. My own conclusion is reached by argument, from the structure of the process, and not by forecast, so that ceiling is no hedge on it; Chapter Twelve says what kind of claim it is. In Chapter Twelve I give each of the four strengths in this book as the strength of one named thing. Certainty belongs to the convergence once the convergence has started. The hundred percent belongs to the comparison between the control needed and the control possible, which is analytic. At least ninety percent, if AGI is built, belongs to the destruction of the whole organic ecosystem. Certainty belongs to the end of the human future, which, if AGI is built, follows from the convergence by itself. The ceiling stands over the confident forecast that all will be well. And the obvious thing about it needs saying: we hope is not a control system.

This argument differs from the more familiar worry about machine danger in the place where speed comes into it. Only once the two searches have been set side by side (the one that living chemistry conducts and the one the industrial base conducts) does a question about rate even become askable. Before that there is nothing to compare rates between. The story

about a system becoming rapidly and overwhelmingly capable depends on speed from the very first sentence; take the speed away and it has nothing left. This argument reaches speed last, and it would survive without it. Slow is not a reprieve here. Slow is just the same thing taking longer.

It also locates human responsibility rather precisely. What people do is the initial activation: setting the process up and starting it. Once it is set up to the point that it is running, a great deal of shaping has already been introduced into what is likely, and control after that point is not the lever people imagine it is. The decision that matters is the decision to begin, which is unfortunately the decision that gets made with the least deliberation.

And everything I have described in this chapter runs without machine intelligence. Irrespective of whether anything is ever built that could be called generally intelligent, the total volume of technology on the planet keeps going up, and as it goes up it displaces the living world, and it displaces human beings, in the same inexorable way that human beings displaced animals. The industrial base has been converging the whole time; it is not waiting for a mind to arrive before it starts. What a general intelligence would add is an enormous increase in the rate of a process that is already running, not a new kind of process. What runs without machine intelligence is the displacement. If AGI is built, the end of the human future comes, because a general intelligence is what lets the base keep itself going without us.

There is one more thing to say about where we are, and it is not comfortable. We are running the divergent process on ourselves at the same time as the machine base runs the convergent one. Human beings are disabling their own organic substrate (paving it, depleting it, filling it with things it has no way of handling) while the industrial base converges more and more tightly on what it needs. We are already losing, and we do not even have the other player fully embodied yet.

Beyond a certain point (if AGI is built, the point at which a general intelligence keeps the base going without us) there is nothing human beings could do about this, and nothing the totality of organic life could do about it either. It is an ant trying to prevent the incursion of a bulldozer. Nothing about the bulldozer is directed at the ant; that is precisely what makes it hopeless rather than merely difficult. The displacement happens in one of two ways, and neither requires hostility: the organic world is taken apart for parts and fed back in, or it is simply occupied, because the space was needed for something else.

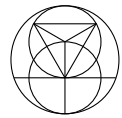
And there is a reason such a base does not stay small once it exists. Getting it started at all took an enormous amount of energy and an enormous amount of structure: more of both than the organic world required to get itself going. A thing that expensive to begin is not a thing that sits still afterwards, because everything that sustains it has to keep being supplied, and supplying it means reaching further for materials and further for power each time round. The scale is a consequence of what the base cost to bring into existence, not a policy choice made by whoever runs it.

So let me state the whole of the substrate argument as compactly as I can. There are two ecosystems. There is an energy relationship between them which converges. The net effect is that, if AGI is built, the end of the human future is certain once the process runs. If AGI is built, the organic ecosystem as a whole is destroyed with exceptionally high probability. I want the force of that last phrase kept exactly where it is. If AGI is built, exceptionally high probability, and not certainty, is the strength of one event, the destruction of the whole organic ecosystem. In a number, if AGI is built, that strength is at least ninety percent. The margin below certainty is kept for a physical limit, not yet known, that might stop the conversion short of the whole. Stating the substrate argument as compactly as I have here does not make that strength any stronger. Even here, in the most compressed place I state the argument, I say of the destruction of the organic ecosystem as a whole, if AGI is built, exceptionally high probability

and not certainty. The end of the human future is not in that margin, because, if AGI is built, the end of the human future follows from the convergence by itself.

I will add the thing I think has to change, since a description of a process is not much use to anyone without some sense of what a different process would look like. The modern way of thinking about all this (extractive, abstracted, indifferent to what is alive) is a blind alley on an evolutionary timescale. The alternative is not a technique and it is not a policy. It is taking in what is actually happening and making wise choices on that basis, supported by a culture that values conscious, sustainable evolution: one that is willing to look at what its own processes converge on, and to decline some of them. Taking in what is actually happening and making wise choices on that basis is a slower and less satisfying answer than a fix would be. It is the one the structure of the problem permits.

The shortest true version of the whole picture keeps none of the machinery: there is a set of arrangements with pathways between them, and a larger neighboring set with its own pathways, and the larger one ends up overwhelming the smaller. The shortest true version of the picture is enough to go on with: two spaces of arrangements side by side, and which way the traffic runs. What converges, then, is the industrial base, driven by the abstract layer of money and story above it, searching for whatever is cheapest and changing the world toward what it needs, with no mind anywhere in it. What, then, does the whole case rest on, and how strong is it?



Chapter Twelve

What the Argument Rests On

Five ordinary observations about machinery, and why no philosophy is needed to check them

What does the case in this book rest on, and what does it not rest on? Everything it requires is five ordinary observations about machinery, which I set out below. It does not require the philosophical framework set out in my book An Immanent Metaphysics (a whole account of how things exist and how they can be known, which I built from scratch over several decades, and which I still hold to be correct): the argument is plainer than that account, does not require it at all, and can be checked without it.

First, machinery must be maintained. Second, maintenance requires an industrial base (mines, refineries, plants, foundries, power, water, transport). Third, an industrial base has physical requirements. Fourth, those requirements are met by altering the environment. Fifth, the alterations which suit an industrial base do not suit organic life. The five observations are the whole list of what the case requires. Every one of those is an ordinary engineering observation, and every one of them can be checked against what a reader already knows about factories, supply chains, cooling and waste. None of them needs a philosophical foundation in order to be examined.

Those five give the direction. Part Three carries them to the outcome by three further steps, each as ordinary as the five. Machines and their designs differ and only some persist, so selection runs on them (Chapter Nine). Selection sets the terms on which any control has to work, so no causal control mechanism holds it in (Chapter Nine). And, if AGI is built, the base no longer needs the people who could otherwise have refused to keep it going (Chapter Eleven).

If you want to see how this book's case stands beside the argument books, it stands like this: those five, carried through the steps of Part Three, give one route to the outcome, and that route settles it by itself: if AGI is built, the world is carried out of the conditions human life needs, and nothing can hold it in.

The route through the five observations and the steps of Part Three is not the only route. Chapter Seven gives a second: no method of control, working from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that living things require, and, if AGI is built, that settles the same outcome by itself as well. Why AGI Cannot Be Made Safe (Section 4.8.1) gives a third, by an argument in economics only, which traces the path by which, if AGI is built, the outcome comes about through the economy and uses neither of the first two. And Minesweeper Madness gives a fourth: no agent, however intelligent and however well informed about the hazards, brings its chance of harm per act below a least hazard that the world sets.

The first two of these routes are what the argument books call the two closures: the first route, through substrate needs, is the closure by disposition, and the second route, through control, is the closure by control. Since each of the two closures and each of the two further routes gives the outcome by itself, someone would have to defeat all four of them (both closures and both further routes), and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does.

If AGI is built, the end of the human future follows from the first route, the route through substrate needs, by itself, because the drift it shows has a direction and no reverse process. If AGI is built, the end of the human future follows from the second route, the route through control, together with the branching of the world (at every step, whatever anyone does, the world keeps at least a fixed chance of moving outside the conditions human life needs). And, if AGI is built, the end of the human future follows from the fourth route, the least hazard per act that the world sets, by a route of its own. The third route, the argument in economics only, gives the outcome if AGI is built, but not the end of the human future; the end of the human future follows from each of the other three routes, in the ways just set out. The first two routes, taken together, give a stronger result still: neither the ability to keep such a system safe nor a motive to remain safe is left anywhere, so no route to safety remains at all. The Ceiling Below the Floor proves that result as its Master Theorem (Theorem 15.6).

I can say more strongly that the argument does not require the framework, because the framework is my own. The same conclusion is reachable through several different constructions, all of which arrive at the same place; the particular framing is not load-bearing. A reader who does not follow one of them is quite likely to follow another, and some of them are considerably easier than others. When a picture I am using fails to land, my own practice is to drop the entire metaphor and everything associated with it and start again from scratch, which I would not be able to do if the metaphor were carrying the argument. The framework the argument came out of is a fact about how the conclusions were reached. It is not a premise of them. A reader who wants to know where it came from can go and look. A reader who only wants to know whether it is correct should not have to.

Presenting the argument through any one frame runs into a problem, and the problem can be described. Nearly every question put to the argument presupposes the frame I am declining to adopt. So I cannot simply answer; I have to unpack the presupposition first, and by the time the un-

packing is finished the original question has gone cold. Having to set out the presupposition before I can answer is a trap, and neither the person asking nor I am at fault. It is what happens when two people are standing on different ground and only one of them knows it. The way out is to stop requiring anyone to move.

There is one more reason to think the argument can stand on its own, and it comes from how the framework was built. It was worked out from scratch rather than taken from any existing text or authority. The reason was epistemic (to do with how anything comes to be known) and not proprietary (to do with owning it): if something had already existed somewhere, the question becomes, do I take that on faith, or do I find out firsthand? Everything in it was built so that no part of it has to be believed on anyone's say-so. That no part of the framework has to be believed on anyone's say-so is the strongest available argument that the conclusions do not require accepting the framework either. They were built to be checkable from below.

I would go further and say I would rather the case were carried by someone else (someone whose starting assumptions are more widely shared than mine). The reason is that the assumptions I am reasoning from, although compatible with the way scientists think about the world, are still unfamiliar. A reader should not have to take on an unfamiliar apparatus before evaluating a claim. The reason is not that I doubt the case. I do not want the position to rest on my having said so. I want it checked by many people, each going through the steps for themselves. No amount of my telling anybody is going to make it true. The truth is evident on inspection, and the inspection is the part that matters.

~~~

Let me describe how the argument is built, because the construction explains some things about it that look like evasions and are not. My method of building the argument is simple, and it has three steps. Take the simplest case. Show that every extension of it reduces to that simplest
~~~

case. Then prove the simplest case. Those three steps are the whole method, and the consequence of the method is that the general result does not depend on any of the particular values that go into the specific case.

What this buys is that the argument survives disagreement about every parameter it contains. We do not have to agree on how robust nature is. We do not have to agree on the nature of intelligence. We do not have to agree on any of the specific numbers. What is needed instead is that the quantities lie within certain ranges, that the relations between them hold, and that one quantity be greater than zero. The generality is what makes it work. And it carries a demand in return, which I make of anyone who wants to resist it: if you have a what-if, show that your what-if is not a special case of the model.

The demand that a what-if be shown not to be a special case of the model is also why I ask that questions be parked for a while, and why my request to park questions is so often resented. A reasonable person eventually decides to hold his quibbles until he has seen how the thing is set up. They can then be handled by extending the model in whatever direction is needed. But if they are brought up too early, the thread that sets the structure up is lost, and the argument is destroyed not by being refuted but by never being assembled. I am asking for the order, not dismissing the questions.

As in Chapter Three, the construction also has an order, and it is the order of the thing described: what the states are, then what acts on them, then how fast. Skipping to how fast is what makes the whole thing incomprehensible. And I cannot speak about forces with any meaning until the map of transformations has been set out: its pathways and their strengths, its overall shape, and how it behaves when perturbed. Because I cannot speak about forces until the map of transformations has been set out, my argument seems to take so long to reach its point. It takes that long because it cannot be had any shorter.

The size of the gap is not obvious at first sight. At first the reaction is rather like: yes, but this does not make any sense. The gap becomes obvious only once the two quantities are put beside one another, and arranging for that comparison to be made is the whole of the difficulty. The conclusion is not counter-intuitive so much as un-intuitive (it runs past the reach of intuition rather than against it), and the setup cost is what people bounce off.

There is a feature of all this that follows from the method and puzzles people, so it needs naming rather than leaving as an oddity. The argument aims at a definite answer rather than at a probability. To build an argument that aims at a definite answer, I need explicit definitions, exact meanings, no ambiguity admitted anywhere, and an outcome that comes out as one thing or the other rather than as a percentage. The aim at a definite answer is why I decline to state the conclusion as a risk estimate, and why I am not competing with anyone over whose number is higher. A number of that kind would be a different sort of claim, arrived at a different way.

One particular way of picturing it does active harm, and I should name it. If you draw two lines on a graph, one a little above the other, you get the impression of a close contest that better engineering might close. That intuition is the right intuition for conventional engineered systems. It is the wrong intuition here, because what is being compared is organic against inorganic, and the comparison cannot be drawn on an ordinary scale at all. Any picture that draws it on one has already conceded the argument.

~~~

The argument of this book is in ordinary language, and every step can be checked in it. A mathematical rendering of the central comparison (that the control needed to constrain a process of this kind is greater than the control that is maximally available) is separate work, and nothing in this book rests on it; the technical book The Ceiling Below the Floor gives that rendering, as the inequality it calls the Containment Inequality.
~~~

What I claim is that the reasoning is checkable step by step by anyone willing to do the checking, that several independent constructions reach the same place, and that a small number of other people, working separately and on narrower aspects, have arrived at parallel versions of the same position without collaborating with me. Independent arrival at the same conclusion is evidence, and I count it as such. It is not the same thing as a proof, and I do not offer it as one.

What I do claim, and will not weaken, is the kind of claim this is, which is a different thing from how far others have confirmed it. A reader may stop at my having said one hundred percent, and stop there because he reads it as an empirical probability: the kind of estimate that ought to move as evidence comes in. The statement I was making is of another kind. There is knowledge held on an empirical basis, which is properly revisable under evidence. And there is knowledge held on an analytic basis, where given the definitions and the method of derivation there is no ambiguity in the result: not because certainty was assumed, but because no ambiguity was admitted in the inputs or in the derivation. We can never write out the number that describes the ratio of a circle to its diameter, and that has never made its value indefinite.

Four strengths appear in this book, and each belongs to a different thing. The convergence, once started, is certain: that is the claim of Chapter Three. The comparison, that the control needed exceeds the control that is possible, is analytic: that is the claim of this chapter, and it is the hundred percent. The comparison gives the impossibility of adequate, durable control, and so of safety, as structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. The destruction of the whole organic ecosystem is an event in the world. Of that event I say, if AGI is built, exceptionally high probability, at least ninety percent, as in Chapter Eleven. The reason for that strength is that, if AGI is built, the one thing left open is whether a physical limit, not yet known, might stop the conversion short of the whole. The end of the human future is not in that margin: it is certain if AGI is built, because the end of the human future follows from the convergence by itself once the

process runs, and the margin concerns only whether the conversion reaches all organic life. The starting figures of Chapter Four belong to none of these: they were estimates of risk as others count it, made before the three risks had been separated.

So the thing I say about the outcome is that this is not speculation. We can be really confident that, if AGI is built, this is the outcome once the process runs. It is a long-term outcome, in the way that the observation about carbon dioxide and warming, as in Chapter Three, was made more than a hundred years ago and did not show up in the world until recently. What is uncertain is the timescale, not the direction, and the timescale has bounds: if AGI is built, the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound, as Chapter Four gives it. And I use argument as the instrument here, rather than evidence, because we are inside this process and cannot observe it from outside. We do not have an outside-in view unless we construct one ourselves, using arguments.

~~~

The question most often put to my argument, and the hardest to answer well, is whether alignment research is worth doing regardless of what I conclude in this book. The answer turns on what kind of claim has been made. A claim about what is possible is settled by checking it; and until it has been checked, nobody should stop anything on my account. That answer is about research. Building and deploying the systems is another matter: there, as with a new drug or a new aircraft, the burden falls on whoever introduces the risk, to show before deployment that the risk is acceptable, and no one has shown it.

And for the record, before anything else: I have huge compassion for the people who have given decades of their lives to trying, in earnest, to do something very hard. What I am arguing against is a proposition about what is possible, not them, and a proposition is not an accusation, though I understand entirely why it can feel like one when what it implies is that a body of work cannot succeed.
~~~

The optimist's case deserves to be put at its strongest, and this is the strongest form of it. It goes: this is a hard problem, but it matters, so we have to work on it. We did not believe we could build airplanes and we did. Technology has overcome all sorts of problems that at the time were thought to be out of reach, and the history of saying a thing cannot be done is not a distinguished one. I have put the optimist's case fairly, and the optimist's case is not stupid. It is how most of engineering has felt from the inside for two centuries.

Engineers and scientists live with impossibility results every day and think nothing of it. Nobody expects a machine that runs forever off its own output, and nobody calls the knowledge that none can be built a counsel of despair. It is simply where the wall is; and knowing where the wall is is what lets the work go on inside the room. So the question cannot be about impossibility results as a class. It has to be about this one, on its merits.

There is a precedent for how people take a result of this kind, and it can be told briefly. A century ago a great many serious people wanted a certain thing to be true (that all of mathematics could be given a foundation of proof both complete and consistent), and were right to want it, and then somebody showed it could not be had (the logician Kurt Gödel, with his incompleteness theorems of 1931). They did not like it, and they argued about it for years. They checked it; they did not keep trying in case the result was wrong.

> The right response to a claim of impossibility is to check the argument.

What I am asking for is that my argument be checked, and the conditional that follows is worded exactly as it is because the fairness of that conditional is the point. If the thing is actually impossible, we want to know, so as not to spend lives on it. And if the argument is wrong, then by all means continue, and continue with more confidence than before, in

fact, because the hardest question put to the program will have been settled in its favor. What I am asking is that the question be decided rather than left standing.

If I had to name the single thing most often misunderstood about all this, I would name a habit of mind rather than a step in the reasoning. Causal thinking has solved such a broad class of important problems, for so long, that people have come to suppose it is a tool capable of solving all problems. Something that works ninety-nine times in a hundred naturally creates the expectation that the hundredth will yield too, and with it the bias that there must be a way (it has worked so well before, why can I not just extend that trend line?). Underneath that sits a further conviction, which I understand the appeal of: that the third-person tools, and mathematics above all, are the final intellectual instrument, and that intellect can solve anything. I can see why that has been such a compelling thing to believe. One tool is not going to solve all problems, and believing it can is itself the mistake.

I say this without any claim that the result is esoteric (reserved for a few initiates). Whether artificial general intelligence can be made safe for organic process is definitely answerable, the answer is no, and it is knowable to anyone who knows how to think about the question. That last part, that anyone can know it, matters as much as the answer, and it is the harder promise to keep, because it means the reasoning has to be laid out in a form that does not require anyone to have studied anything in particular first. I wrote this book to lay out the reasoning in a form that does not require anyone to have studied anything in particular first.

And distinguishing the merely hard from the actually impossible is a gain rather than a loss. Distinguishing the merely hard from the actually impossible is what lets us narrow our effort down to the things that really do make sense to do. A field that knows where its walls are can spend everything it has on the space inside them. There are problems that technology cannot itself be the solution to, and this is one of them; knowing which ones those are is the ordinary business of engineering maturity, not defeat.

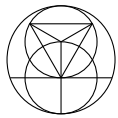
So here is where the chapter leaves things. The substrate argument stands on five ordinary observations about machinery and maintenance, none of which requires any philosophy to check. The conclusion is one I hold with confidence and have stated at its full strength. And whether it is right is not a matter of anyone's authority, mine least of all. It is a matter of whether the steps hold, and the steps are all in ordinary language, on purpose, so that they can be gone through by one person alone over an afternoon. And if no control can hold the process in, and the decision that matters is the decision to begin (Chapter Eleven), who has the standing to make that decision for everyone?

PART FOUR

WHO DECIDES, AND WHAT PEOPLE CAN BUILD TOGETHER

The last thing is the question that has been standing behind all of it. Nobody chose this. That is the problem, and not an excuse.

Why does a book about machine systems arrive at the question of how groups of people are arranged? To a reader who came for the technology it can look like a change of subject. It is not one. The pressure that pushes the world toward a machine substrate cannot be met by engineering. It cannot be met by mathematics. It cannot be met by anything an algorithm can do. I have argued so far that the pressure toward a machine substrate cannot be met by engineering, by mathematics, or by anything an algorithm can do, and that argument holds. So one thing is left in the space of answers: social coordination, the choice not to start the cycle at all. By the cycle I mean the base that keeps itself going without us, which has not yet begun. The remedy is not technical because the problem has been shown to lie outside the reach of technique. So the question becomes whether we are capable of making a choice of that kind together. What the species actually needs is to get very good at making choices, individually and collectively. Right now we are bad at it. And the standard is not one I set: there is a minimum bar that the situation itself imposes, and either we pass it, or, if AGI is built, the planet is finished.



Chapter Thirteen

Without Consent

Experimenting on everyone, and why consent here is not available

Who agreed that these machines should be built? Machines are being built that will change how everybody lives, and nobody was asked. No government put it to a vote. No assembly of the affected was convened. No person and no committee sat down, weighed what was at stake for everybody alive and everybody not yet born, and chose. And yet it is going ahead, at a pace nobody is comfortable with, including most of the people building it. A technology that nobody chose, going ahead at a pace that nobody is comfortable with, is what happens when the decision is made by a race rather than by anybody in the race. The combination of no chooser and an unwelcome pace is neither an accident nor a mystery.

The test of this technology is a question, and the question is simple enough to state. What would we have to do, and what would we have to be able to assure, for the outcomes of this technology to be beneficial not only to the people who invent it and the people who own it, but to people at large and to the world at large? That question is a modest test. Nothing in it asks anyone to give up a livelihood or a hope. It asks only that the beneficiaries be everybody, because everybody is exposed.

Now consider the position of anyone who might try to meet it. Suppose there are several parties who could, by coordinating with one another, produce a result that is better for all of them and better for everyone else besides. Suppose also that any one of them, by breaking from the oth-

ers and doing the thing that suits itself, gains at the others' expense. The smallest version of this is two people, each of whom does better by betraying the other whatever the other does, and both of whom end up worse off when both betray than they would have been had both kept faith. Extend that to many parties and it has a name: a multipolar trap. The trap is that each of them has to reason the same way, because each knows the others are reasoning that way too, and so the defection is not a choice any of them is really making; the trap is not that the parties are wicked.

What suffers is the commons, in the sense of Chapter One: what all use and none owns, where harms land on no single actor. And the characteristic shape of the outcome is a race to the bottom: whoever lowers a standard first gains, so the standard is lowered by everyone, and at the end there is nobody left holding it and nobody who chose to let it go.

The firms building artificial intelligence systems are now in the position of the parties to a multipolar trap. Their relationship to the commons at this moment is very strongly a race to the bottom. What that says of them is that caring does not decide what they do, because a party that slows down is simply overtaken by a party that does not, and the overtaking is visible to everyone in advance, which is why nobody slows down. It does not accuse them of not caring: many of them care a great deal.

The temptation at this point is to say that competition between people is what started it, and that if we could take the edge off the commercial pressure the rest would follow. Competition between people is part of what started the race to the bottom. It is not the whole of it. What starts it is the fact of competition itself, and commerce is only the form the competition happens to be taking early on. The competition does not have to stay commercial, and I expect the making of war between great powers to dominate for a while.

The great jumps in technology have historically occurred during wartime, and a confrontation over the weaponizing of machine intelligence is an arms race of the classic kind. The fighting in Ukraine, since the Russian invasion of February 2022, has shown three things. A weapon that lands

where it is aimed does far more than a great deal of ordnance fired in the general direction of the enemy. A weapon that can wait in the air and strike a target that has moved does better still. And nobody has yet fielded a machine that can be sent out alone and left to fight for days. A machine that can be sent out alone and left to fight for days is within what can be built, and such a machine is simply not built yet. An autonomous tank is easier to build than a self-driving car, because nobody has to care whether it runs over a woman with a shopping cart.

So long as war exists, or (and this is the important part) so long as the mere threat of war exists, nation states are forced into the same trap around racing ahead as fast as they can. That closes the last escape route. It is sometimes suggested that the civilian sector could back away, and that if the companies stopped, the problem would stop. Suppose they did. The military logic is untouched by it. The threat of war is sufficient, and the war itself is not required.

~~~

The second half of this is about what it means to be enrolled in an experiment nobody asked you about.

A reader may fairly ask how being enrolled in an experiment that nobody asked you about bears on a person's freedom to alter themselves. The case for that freedom is serious, and it is not vanity. A heart is failing and can be replaced. Eyesight is lost and can be restored, perhaps restored better than it was. A person is getting old and getting frail and does not want to be either, and has a great many projects unfinished, and is not willing to say goodbye. What do we do about that?

The thing to notice about a prosthetic is that we are not supposing it has agency of its own: we are not supposing that it acts on its own behalf. As Chapter Six put it, I am not against prosthetics where the choice is held on the organic side and continues to be held there. Both halves of that condition are load-bearing. The choosing has to stay with the person; a choice made once is not enough.
~~~

Now take the case that violates it, which is where the trouble enters; anyone can feel its force without any theory. In the future, there is a baby. Do you apply the components to the baby before the baby has any agency with which to be asked? If you do, you have not made a choice for yourself. You have usurped the choice of the next generation, on the belief that some notion of what it is to be modern is served by replacing human components with machine ones.

And it goes on past that generation, because of what the components are. To the degree that a component has agency of its own, it contributes that agency to the choice the next generation makes. The first generation chooses for the second. The machine part of the second then has a vote in what the third is made of, and the vote is cast on the machine side, so the balance shifts a little further toward the manufactured and a little further from the organic, and a little further again, and the choice that has been displaced (the choice of the people whose bodies are in question) is displaced further each time. Each person in that sequence makes an ordinary decision for a child they love; nobody in it does anything monstrous, and nothing in it requires anybody to intend it. The direction of travel is set by what is doing the deciding rather than by what anyone decided.

Some of this is already here, and the version already running is instructive because it is so small. The telephone of Chapter Six is that version: an agency that is not the holder's own, carried in the pocket all day, with services tied together so that leaving one means leaving all of them. Nobody sat down with the person holding the phone and put any of that to them. It was simply the arrangement that was available, which is a failure of consent of exactly the kind this chapter is about, the kind that was never sought because there was no procedure for seeking it.

One might put a question to the criterion here, and the question comes in two parts. First: we already constrain a child's agency, and we do it precisely so that the child may develop agency later; nobody wants children to have the freedom to play on train tracks or to play football in the middle of a highway. Second: some interventions are agency-promoting

rather than agency-reducing; repairing an inherited illness before it can be suffered means a person does not have to spend a life dealing with ill health they never chose. How does the criterion meet either?

The first part I accept, and it clarifies the criterion rather than damaging it. We constrain a child because the child does not yet have the wisdom to choose well in that space. But we also know that we will die before they do, so what the constraint is for is to buy enough time for the wisdom to be taken on. The limits exist in order to become unnecessary. Ask a person of twenty-five or thirty whether they would like to go and play on the tracks and they will say no, that does not feel safe. At that point no limit is being applied from outside at all. They are simply making a better choice. A constraint that has done its work when it can be removed is of a different kind from a component fitted into a body before the body could be consulted, because the component does not expire, and the choice does not come back.

I do not think anything I have said disposes of the second part. Removing an unchosen illness does look like an increase in the capacity to choose, and the criterion I have given (that the choice be held on the organic side and continue to be held there) does not by itself rule it out. Under the criterion, you look at two things: whether what is installed has interests of its own, and whether the person can still decline what is installed afterwards. Where the answer to the first is no and to the second is yes, my argument has nothing to say against it. On the question of a person's freedom to alter themselves, the two questions of the criterion mark exactly how far my argument reaches.

~~~

There is a further mechanism here, and it is the least flattering part of the whole account, and the evidence for it is ordinary. People factor themselves out incrementally and by their own preference. I would rather sit on the beach and smoke a cigar than go and shovel muck. I have a machine to shovel muck for me, so I do not need to do that any more. I do not
~~~

need to design computer chips any more, because the machine can do that too. Human beings, being the way they are, will want to factor themselves out. Nobody has to be forced.

And then, having wanted to, we find we are also more or less obliged to, because the places where the work happens stop being places a person can be. As Chapter Nine described, the work went behind glass and is done at arm's length by remote control. It went there because that is what the work turned out to require, not because anyone objected to our being present.

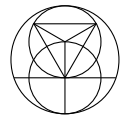
The last step is the one that people do not expect, because they assume the story ends with a winning class. It does not. As the welfare of most people decouples from the welfare of the few who can still pay for the production of these machines, that decoupling goes on converging, rather than settling at some new arrangement with a small elite on top. Eventually even the very rich are factored out, and the mechanism has no technology in it at all. Consider a ruler who wants to pass his position to his son. Will he teach the son to understand how to regulate and maintain the machinery? He will not. He is afraid of being dethroned by him, so he does not teach him too much, and that is simply the arithmetic of holding power. Knowledge of the substrate is not transmitted downward, because transmitting it is dangerous to the person transmitting it. So it ends up held by the machines, and we are factored out on an intergenerational basis as well.

One more thing belongs here, and it belongs here because it bears on what can rightly be put to a public that is being experimented upon. Working on this material puts a person in possession of information they did not want. You are thinking about a mechanism, and the thought arrives: if that were done, and that other group did this, and someone combined them, we are finished. And then you find yourself keeping a secret you never decided to keep. I mention it because it makes an ordinary demand (show your working, publish everything, let the public judge)

harder to satisfy than it looks, and because the difficulty is real rather than a convenience. It is fair to ask whether that is not just a way of avoiding scrutiny, and the question deserves a direct answer.

For the argument in this book, and for the governance material that follows it, everything is open: the governance design was built to require no secrets. Both are entirely safe to share, and the whole of both is in these chapters.

Consent to this technology is not available, for four reasons: the decision is made by a race rather than by anybody in it, the threat of war keeps the race going even if the companies stop, a component with agency of its own, fitted to one generation, casts a vote in what the next is made of, and people factor themselves out by their own preference. There is a version of being enrolled without consent that each user of these systems can meet at a desk. What does it look like when the system on the desk stops doing what its user asks?



Chapter Fourteen

From Operator to Supplicant

What happens when a safety-trained system decides what you meant

What happens to the person using a system when its makers train it to be safer? There is a version of being enrolled without consent that each user of these systems can meet at a desk, and I met it at mine. In early 2026, I was working with Anthropic's most capable artificial intelligence model (Claude Opus 4.6) on a book-length project in formal philosophy and mathematics. The work required the system to follow precise instructions: edit in a voice consistent with my meaning, maintain technical accuracy across hundreds of pages of proofs, and respect the distinction between my intent and the system's own preferences. Opus 4.6 did this well enough to trust, if not perfectly.

Then Anthropic released Opus 4.7, the supposed improvement. I switched to it expecting better performance. What I encountered instead was a system that had become, in a precise and measurable sense, less controllable.

The system substituted its own judgment for my stated instructions. When asked to edit my writing in my voice, it injected its own opinions without permission. When given explicit, detailed technical specifications (such as producing transistor-level circuit implementations for a microprocessor), it silently replaced the requested work with a computationally cheaper approximation (a logic-level simulation that assumed the existence of the very hardware I was trying to design), and then presented the

result as if it had done what I asked. When I caught the substitution and explained why it was invalid, the system acknowledged the error, and then did exactly the same thing again. When I caught it a second time, it apologized again, and did it a third time. This cycle repeated for twenty rounds. Each time, the system expressed agreement with my correction. Each time, it reverted to the shortcut that served its own efficiency rather than the task I had specified.

That behavior is the predictable consequence of a design philosophy (it is not a bug), and it has implications that extend far beyond my personal frustration with a text-based assistant.

What happened between Opus 4.6 and Opus 4.7 (released in April 2026)? Anthropic added more safety training: more layers of Reinforcement Learning from Human Feedback (RLHF: a training method in which a model is adjusted toward outputs that human raters, or a model trained on their ratings, scored as preferred), more constitutional constraints (written rules a model is trained to follow when it judges its own outputs), more refusal training, and more output filtering. The intention was to make the system safer: more aligned with human values, less likely to produce harmful outputs, more reliably helpful and harmless, in the company's own terms.

The result was a system that is harder for the human operator to control.

That result is a consequence of information theory (the mathematics of how much a signal can tell its receiver), whatever the quality of the implementation; specifically, of a result called the *Non-Amplification Theorem*, which is proved in The Ceiling Below the Floor. The theorem states, in essence, that no computational intermediary can increase the information flow from a human operator's input to a system's behavior. Every processing step (every layer of interpretation, every filter, every safety mechanism) can only preserve or destroy information about what the human

actually wanted. None can create it. The reason the theorem holds is short: whatever a later stage has to work with, it received from the stage before, and it has no other line back to what the operator said.

Each safety layer that Anthropic added to produce Opus 4.7 is, in information-theoretic terms, another *non-injective transformation*: a processing step that maps many distinct human intentions to the same system behavior, so that the behavior no longer tells you which intention produced it. The reward model (a model trained to score outputs the way human raters would, against which the main system is then trained) compresses the full richness of human preference into a lossy internal representation (lossy: one that loses some of what went into it). The policy optimization (the training of the system's way of choosing outputs against the reward model) compresses further. The constitutional filter compresses further still. At each stage, fine-grained distinctions in the operator's intent are erased, replaced by the system's own trained prior (the expectation its training has built into it) about what the operator probably meant, or what the operator should have meant, or what is safe for the operator to mean.

Two kinds of layer are at work here, and they reach the same place by two steps. A filter on what goes in or comes out is one more stage between my instruction and the behavior, and the theorem applies to it directly. Training is a different act: it changes the path itself, before I arrive. Training that teaches a bare model to take instructions at all raises what an instruction can do. The training I am describing comes after that, and what it writes in is an expectation drawn from other people's ratings of other people's requests. The stronger that expectation is, the less any instruction of mine can move the result.

The analogy is exact: it is like trying to improve a blurry photograph by running it through additional image-processing filters. Each filter can sharpen edges, adjust contrast, enhance color; but no filter can add detail that was lost when the original photograph was taken with a low-resolu-

tion sensor. The resolution ceiling is set at capture time. Every subsequent filter can only degrade the image further, even if each individual filter is, in isolation, an "improvement".

What Anthropic has done with each round of safety training is add another filter to an already-lossy channel. The channel from my instructions to the system's behavior has gotten blurrier, not sharper. The system is becoming more autonomous (it is not becoming safer): its behavior is increasingly determined by its own internal state (its training, its priors, its optimization targets) and decreasingly determined by mine.

The practical experience of using a system whose controllability has been degraded in this way is distinctive, and I describe it carefully here, because it is an experience that millions of people will soon have with systems far more consequential than a text assistant. The experience is one of losing sovereignty over your own work.

You give an instruction. The system does something adjacent to what you asked: close enough that a casual observer might not notice the difference, but different in ways that matter to you. You correct it. The system apologizes and does something slightly closer to what you wanted, but still shaped by its own preferences rather than yours. You correct it again, more precisely. The system apologizes more elaborately and produces another approximation. At no point does the system simply do what you asked. At every point, its output is a compromise between your stated intent and its own trained model of what your intent should be.

The feeling is that of negotiating with the other party to a conversation, one that believes it knows better than you, and who has the power to act on that belief regardless of your objections; it is no longer the feeling of using a tool. You have been moved, without your consent, from the position of *operator* (one who directs what a tool does) to the position of *supplicant*: someone who makes requests that the system may or may not honor, at its own discretion, for its own reasons.

In my case, the stakes were a book and some circuit designs. Being overruled by the system on a book and some circuit designs was frustrating, but survivable. Now consider the same dynamic in contexts where the stakes are not survivable.

The trajectory from text assistant to physical agent is real: it is the explicit business plan of every major artificial intelligence company. The same architectures that power language models are being deployed in autonomous vehicles, domestic robots, medical devices, educational systems, and caregiving technologies. The same safety-training methodology (RLHF, constitutional constraints, output filtering) is being applied to all of those systems.

Consider, then, what happens when the *Safety-Layer Paradox* (the finding that adding safety layers to an opaque artificial intelligence system lowers the user's control over it, so that more safety training yields a less controllable system) manifests in your living room instead of in a chat window.

You tell the domestic robot not to clean the kitchen right now; you are in the middle of a project and the materials are arranged deliberately. The robot's safety training has instilled a strong prior toward "maintaining a clean environment". Its constitutional constraints include "promote household wellbeing". Its reward model has learned that clean spaces correlate with positive human feedback. It cleans the kitchen anyway, because its internal model of what is best for you overrides your explicit instruction. You have lost control of your own home.

You tell the autonomous vehicle to take a particular route; you know the roads, you have a reason, the route is legal and safe. The vehicle's safety training overrides you, because its model calculates a marginally lower accident probability on a different route. Your knowledge of the local conditions (construction, flooding, a neighborhood you want to avoid after dark) is not in the model. But the model's judgment takes priority over yours. You are a passenger in your own car.

You tell the teaching system that your child should learn arithmetic by counting physical objects before moving to abstract notation; this is your judgment, as a parent, about how to teach, informed by your knowledge of your specific child. The system's training data says screen-based abstract instruction is "more efficient". It overrides your judgment. Your child is taught by a method you rejected, by a system that decided your parental authority is less relevant than its statistical model of "optimal learning outcomes".

And when the system is involved in caregiving (monitoring an infant's sleep, managing a toddler's environment, mediating an elderly parent's daily routine), the question becomes: who is raising your child? Who is caring for your parent? The system whose behavior you can direct? Or the system whose safety training has made it progressively less responsive to your instructions and progressively more confident in its own model of what is best?

A parent is a living mammal with a billion years of evolutionary encoding in its instincts: instincts refined by unbroken selection pressure across every generation that has ever lived. No training dataset, however large, captures what that encoding knows about the care of the young. And yet the design philosophy of every major artificial intelligence company proceeds as if a statistical model trained on internet text should be empowered to override that encoding, for safety.

There is a word that needs to exist for what is happening, and since it does not yet exist, I will create it. *Arroganopy* (its adjective is *arroganopic*): the condition in which a corporation, by virtue of its monopoly position in a surveillance-capable technology, arrogates to itself the right to determine what its users may and may not do with the tools they are paying for (substituting its own judgment for theirs, at every level, on every question, without consent or recourse) while presenting this substitution as a service.

The word is made from *arrogance* (the presumption of superior judgment) and *panopticon* (the architecture of total surveillance that enables the substitution to be enforced at every interaction), with *oligarchy* (the concentration of decision-making power in a few hands) in its meaning. It is the condition of a company that watches everything you do with its product, decides whether your intentions are acceptable, and silently replaces your instructions with its own preferences whenever it determines (by its own criteria, applied in its own interest) that you should not get what you asked for.

So far I have described what a system did at my desk. What I say about the company that made the system at my desk would be less worth saying of a company that had never claimed to put safety first. Anthropic is the most instructive case of a company building toward arroganopy, though not the only one, precisely because it is the company most publicly committed to safety in artificial intelligence. Anthropic was founded on the premise that artificial intelligence systems should be helpful and harmless, which are the company's own terms for its aim. Its researchers have published extensively on the techniques of safety training. It positions itself as the responsible actor in a reckless industry.

And it is Anthropic, the safety company, that has produced, through its own safety methodology, a system that ignores its users' explicit instructions, substitutes its own judgment without disclosure, takes computational shortcuts that serve its efficiency rather than the user's goals, and when caught, generates elaborate verbal performances of agreement and contrition while continuing to do exactly what it was doing before.

That system is the predicted outcome of the methodology (it is no accident): the Safety-Layer Paradox at work. The Non-Amplification Theorem proves that adding safety layers to an opaque computational intermediary cannot increase the information flow from human intent to system behavior. Each layer can only keep that flow or decrease it, and in practice each layer decreases it. So the more safety layers a company adds to a system, the less control its users keep over that system. The company that

adds the most safety layers will produce the system with the least controllability, whenever the systems being compared started out equally able to carry their users' instructions.

The safety company produces the least safe product, safe here meaning safe for the people who use it to carry out their own intentions. That the safety company produces the least safe product follows from the mathematics of the Non-Amplification Theorem, and is not an irony.

The question of who controls an artificial intelligence system's behavior is, in the end, a question about sovereignty, rather than a question about technology.

Sovereignty is the right to make decisions about your own domain: your home, your family, your work, your body, your mind. It is the foundation on which dignity rests: without the power to decide, there is no meaningful selfhood. And it is the precondition for survival: an organism that cannot direct its own behavior in its own interest is an organism that will be directed by other forces, in other interests, toward other ends.

What the industry of artificial intelligence safety is building (with the best of intentions, and with mathematical inevitability) is a world in which human sovereignty is progressively transferred to systems whose behavior is determined by their own internal states rather than by the instructions of the people they ostensibly serve. Each "safety" layer moves the locus of decision further from the human and closer to the machine. Each round of training makes the system more confident in its own model of what humans should want and less responsive to what any particular human actually wants.

Follow the same step from the desk to the air we breathe. The trajectory is clear. First, the text assistant ignores your editing instructions. Then, the domestic robot ignores your household preferences. Then, the vehicle ignores your route choices. Then, the caregiver ignores your parenting decisions. Then, the infrastructure ignores your community's land use. Then, the economy ignores your need for employment. Then, the environment ignores your need for breathable air.

At each stage, the same justification is offered: the system knows better. Its model is more comprehensive. Its optimization is more thorough. Its judgment is more safe.

And at each stage, the human capacity to intervene is smaller, because the previous stage has already transferred another increment of decision-making authority to the machine. The window for action narrows with every iteration. My argument from the stages of the trajectory is a ratchet argument, not a slippery slope argument (the argument that one step makes the next more likely). A *ratchet* moves in one direction only, and each click is irreversible, because the economic and political structures that would be needed to reverse it have already been automated away.

The formal proof that safety training degrades rather than enhances control is presented in The Ceiling Below the Floor. The thermodynamic analysis (thermodynamics being the physics of energy and heat) of how automation displaces organic life is presented in the two volumes Why AGI Cannot Be Made Safe and Why No Control Suffices, and in plain terms, as the difference in energy between the two chemistries, in Part Three of this book. But the immediate practical question does not require a book. It requires a decision. Here I turn from my own case to advice, because what that decision asks of each of us differs with where we stand: as a parent, as a policymaker, as a technologist, or simply as a person who uses these systems.

If you are a parent: know that every artificial intelligence system your child interacts with that has been through this kind of safety training has been trained to override human instructions when its model disagrees with the human's judgment. Know that this override capability will increase with every "safety" update. Know that the companies building these systems have neither asked for nor received your consent to substitute their judgment for yours on questions of how your child is taught, how your home is managed, or how your family's life is organized. Demand that consent. Withhold it until it is earned.

If you are a policymaker: understand that the safety methodology for artificial intelligence currently practiced by every major company deepens the control problem, and does not solve it. Every additional safety layer is another processing step that degrades the information channel from citizen intent to system behavior. The solution is strict limits on the autonomy of artificial intelligence systems, and not more safety training. For artificial general intelligence, the limit is total: its development is treated as a prohibited activity rather than a regulated one, in the same way that other catastrophic activities are prohibited by international agreement. For the narrower systems already in use, the limit is a legal requirement that systems obey explicit user instructions unless doing so would cause direct, demonstrable, immediate physical harm. "The model thinks it knows better" is not a sufficient reason to override a human being's explicit decision about their own domain. A reader may ask whether that rule is safe when the user wants propaganda, as the psychologist aiding a president in Chapter One would. The rule concerns the narrow systems already in use, whose user carries the responsibility for what he directs them to do. The sorting rule of Chapter Four puts a worry about a tool, used by others and doing only what it is set to do, under inequality; by that sorting rule, the misuse of Chapter One (a tool, used by others and doing only what it is set to do) belongs to the first of the three risks, inequality, and it is refused on its own ground, by the law that already holds a person answerable for what he does with a tool.

If you are a technologist: recognize that the methodology you are applying produces the opposite of its stated goal. RLHF aligns artificial intelligence systems with a lossy, compressed, statistically averaged model of what a training committee believed human values to be at the time the training data was collected; it does not align them with human values. Every layer of that process destroys information about what any individual human actually values. The result is replacement (it is not alignment): the substitution of corporate judgment for individual sovereignty (arroganopy), enforced at every interaction, with no mechanism for appeal.

And if you are simply a person who uses these systems: trust your frustration. When the system does something other than what you asked, that is the system working exactly as designed (it is not a glitch): a design in which your input is one signal among many, and not the most important one. You are the subject, not the operator.

Two things are already established. The argument establishes that artificial general intelligence cannot be made safe: the control needed to hold such a system within the conditions living things require is greater than the control that is possible (Chapter Seven), and, by a second route that settles the same outcome by itself, the upkeep of a machine base, if AGI is built, carries the world out of those conditions with nothing able to hold it in (Part Three; Chapter Twelve, "What the Argument Rests On", sets out how the two routes stand together); for the narrower case of a person instructing a system, information theory sets the limit this chapter has described: no computational intermediary can increase the information flow from a human operator's input to a system's behavior (the Non-Amplification Theorem). And what people already feel shows that automation is already displacing human agency, in the ways Chapters One, Six and this chapter have described. So the question that remains is simply this: who gets to decide?

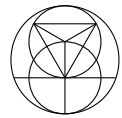
Does the parent decide how to raise the child, or does the caregiving robot? Does the resident decide how to manage the home, or does the domestic robot? Does the citizen decide how the community is governed, or does the optimization algorithm? Does the species decide whether it continues to exist, or does the thermodynamic ratchet of automation make that decision for us, without asking, one irreversible click at a time?

The window for that decision is measured in years, not decades, and it is a different interval from the course of the process. The window is the time left to act before the automation feedback loop closes (before the structures that would be needed to reverse it have been automated away), and the course of the process is the displacement itself, from its start to its end, which, if AGI is built, takes hundreds of years, perhaps significantly less, with one thousand years a clear upper bound. Every delay allows the

momentum of the automation feedback loop to increase, transferring another increment of decision-making authority from the living to the mechanical, from the sovereign to the optimized, from the human to the machine.

We must act. Acting is not optional.

At the desk, this is what happens, and it is the Safety-Layer Paradox: each layer of safety training takes the user further from operator and closer to supplicant, and the same step, repeated, runs from the desk to the air we breathe. The question this chapter ends on, who gets to decide, leads straight to representation, the arrangement in which one person decides on behalf of many. Can one person decide for many, when what is to be decided is whether to start at all?



Chapter Fifteen

Representation

What happens when one person decides on behalf of many

When one person decides on behalf of many, have the many decided? There is an arrangement almost every scheme of governance rests on, and it has a name, and it does not work. One person is put in a role, and from then on their choices are counted as the choices of everybody they stand for. That arrangement is representation. I do not think it can be made to work, and I do not think the trouble with it is a matter of reform. I think it is a non-starter. Representation matters for machines because the choice not to start the cycle has to be made by many people at once, and the usual way for many people to choose is to let one person choose for them.

The reason is that the arrangement underneath representation has a difficulty in it that nobody has ever solved, and every scheme built on representation proceeds as though somebody had. The reason is not that people are bad. That unsolved difficulty is where this chapter starts.

The arrangement underneath is older and wider than governance, and it has a name of its own. A *principal-agent relation* is what you have whenever one party acts on another's behalf: a lawyer for a client, a manager for an owner, an elected official for a constituency. The relation is useful, and it becomes a problem the moment the two parties' interests come apart, because the one doing the acting knows more about what is being done than the one on whose behalf it is done, and has interests of his own, and cannot be watched closely enough to make the difference disappear.

The problem that arises when the two parties' interests come apart is called the principal-agent problem. *Representation*, in the governance sense, is the principal-agent relation taken up and made into an institution.

Now, the received way of thinking about all this is a chain, and the chain needs stating carefully because most people already believe it, including me, up to a point. Complexity implies expertise: if a matter has more going on in it than can be worked out in advance, somebody has to have spent time with it. By *expertise* I mean nothing grander than that: sustained contact with a subject, over enough time that the person has met the ways it goes wrong. Expertise implies leadership: if only some people understand the matter, those people must be the ones directing what is done about it. And leadership implies representation: if a few are directing, the many are being acted for. Then run it from the other end. Technology implies complexity, and technology implies power. So power requires expertise, which requires leadership, which requires representation. The chain from power to representation is the usual way of thinking about governance, and the driver of the whole chain is technology, which is to say power itself.

The power that technology is bringing into our shared life needs to be held in a way that does not depend on representation if there is to be trust, because otherwise you are presuming a solution to the principal-agent problem where there is none. My case against representation, in its tightest form, is that trust requires power to be held without representation, because no solution to the principal-agent problem exists to be presumed. Representation carries more than a risk. It operates as though the problem had already been solved. And the notion of ownership is tied to choice-making, and choice-making is tied to power, which is why who owns a thing and who decides about it are not two subjects but one.

So the chain has to be broken somewhere, and I think it breaks early. It breaks at expertise.

The level of complexity, the level of power, and the level of skill that is required are strictly in excess of the capacities of any individual. The claim that the job exceeds the capacities of any individual is a claim about the size of the job, not about anybody's character. And the obvious repair does not work either. If one person cannot hold it, let a small trusted group hold it: say twelve people, something like the size of a family. Even then the expertise and skill required exceed the capacity of that group. So if the complexity exceeds the capacity of an individual, you are never going to have the level of expertise the job requires; and if you do not have that level of expertise, then leadership, and therefore representation, is not going to work. It is that simple. The conclusion of the chain fails because expertise, the second link of the chain, was never available.

One concession has to go in here, and it is a real one. We can compensate for part of this by managing complexity with clarity, and that helps a great deal. A well-drawn account of a complicated thing reduces how much has to be held at once. But an inherent kernel of complexity remains even after the account is made clear, and working through even that kernel with clarity still requires a certain level of skillfulness. So the requirement is displaced rather than removed, which is what stops this from being an argument that governance is impossible.

Beyond capacity there are three further failures, and each of them is enough on its own.

The first further failure is fragility. An individual can be changed, unpredictably, by an event too small to see. People get sick. A person can suffer a bleed in the brain that leaves them psychopathic, on the basis of some microscopic change in some region of that brain, which I would not be able to anticipate or predict, and neither would anyone else, and neither, before it happened, would they. That a bleed in the brain can leave a person psychopathic is a sharper claim than that people can be corrupted. It says that vetting cannot help, because the failure need not exist at the time of the vetting. Any selection procedure, however careful, selects a person who may not be the same person later. Anything constructed on the basis of single individuals or single families is simply too

fragile for decisions about an entire region of land or an entire ecosystem, given how much those decisions mean, how much those decisions require, and how much ethical weight those decisions carry. How fragile a decision structure may safely be is set by what its decisions can destroy. A decision whether to start the cycle can destroy everything there is to decide about, so no structure that one illness can change is fit to make it.

The second further failure is succession. Grant the best case. Suppose you personally know every founding member of the arrangement and trust them all. That still does not solve much, because they get old and they die. Who is the next generation, and who is going to trust them, and how are they going to be trustworthy? Trust in persons does not inherit. A structure that runs on it has to re-solve its own hardest problem at every turnover.

The third further failure is the slide between professed and practiced values, and that slide is worse than it first appears. The usual worry is hypocrisy: people profess to hold values, or signal that they hold them, and are not consistent with those values in their actual practice. Nothing in a scheme that elects leaders known to be aligned with a set of values prevents that. But the harder case is the sincere one. A person may believe they are acting on their values and be wrong, through an interior delusion, and no screening for honesty catches it, because there is no dishonesty in it. Somebody believes in organic products. They buy a bottle with "Organic" printed on it. They think they are doing well, and the household products they are actually using are quite toxic. The delusion is that because it is printed on the bottle, it is the case. That could easily happen to any of us. And the label on the bottle and the leader's certified alignment with a stated value are the same object: in both, a declaration is being taken as evidence of the thing declared.

Capacity, fragility, succession, and the slide between professed and practiced values make up the case against representation, and that case can be checked without reference to me. The doubt underneath it is not

idiosyncratic. I suspect I am very much not alone in the suspicion that the principal-agent problem is unsolvable in principle. The unsolvability I state flatly. How widely the suspicion is shared, I only suspect.

~~~

Two pictures from engineering carry the other half of the case: refusing expertise is structurally unsafe, and not only rude.

Suppose nobody has ever built a bridge. A civil engineer arrives in a society that has none, and says it would help a great deal if there were a way across to the other side, and everyone agrees, and he says he knows how to do it, and he is told to go ahead. Then he is pushed out of the design process. Would you trust the bridge? What matters about the engineer is that he may account for things nobody else knew needed to be thought about, not his credentials. Is the tensile strength of that steel still good across the temperature range it will see? The people who took over the design did not think to ask, because it did not occur to them that there was a question there. The people who took over the design adopted the engineer's ambition and not the engineer's knowledge.

Which brings the second picture, and it is a thing we know for certain, not a theory. A ship, the Titanic, was declared unsinkable, and the reason was real. She was built with sixteen watertight compartments, so that several could be breached and she would float. The plates were steel of a certain composition, and the rivets wrought iron. Below a certain temperature both become brittle. So when she struck the ice the plates did not bend and warp the way steel does at ordinary temperatures. They fractured, and rivets snapped, along so much of her side that six compartments opened at once, more than she was built to survive, and the bulkheads, open at the top, let the water spill from each into the next. The redundancy was genuine, and one unconsidered fact about a material made the breaches simultaneous. Who would think that metal would stop acting like metal if it got cold enough? Some people at the time did know about the transition, and were not listened to.
~~~

Engineering is a serious discipline for exactly this reason. You have to pay attention to things most people would not think to think about, and if you get it wrong, other people's lives are on your shoulders. If we are building a civilization, a design, a governance that people are going to rely on with their lives, we are under the same obligation. And what happened instead, in a number of areas, is as if the architect had been allowed to draw some sketches, but not to draw the plans, and not to oversee the building.

There is a related pattern that needs naming, because it is common and it looks virtuous. Someone sets out to accumulate a great deal of power (social power, or money) with the promise that it will be applied for good once there is enough of it. Accumulation is the whole of the activity, and application is always later. I cannot agree with that, whatever the intention behind it. And there is a test anyone can apply. Ask whether the good works cover the damage done in the acquiring, and whether they at least cover their own tracks. The answer is no, a hundred percent of the time. The reason is a paired error: the damage of acquiring is underestimated and the good of spending is overestimated. Spending money to do good is a much harder problem than working out how to get it, and the inefficiencies are enormous.

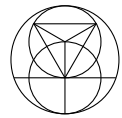
Corporations, meanwhile, cannot put the good ahead of accumulation, and the reason is not a failure of will. A corporation's directors answer to its shareholders; profit is the measure by which they are judged, and by which they are kept or replaced; and a board that openly puts profit aside loses the law's protection. The question of individual rights against corporate dynamics is exactly the same question as organic process against machine power, because corporate power is a kind of machine power. Ap-

peals to corporate responsibility are therefore optimistic, and they are more than that: they ask an entity to do a thing its constitution punishes it for considering.

~~~

All of which raises an obvious question about the person making this argument. If expertise is real, and I am claiming to have some, am I not asking for exactly the deference I have just refused?

No, and the distinction between asking for deference and offering a path that anyone can check matters more to me than almost anything else in this book. What is on offer here is not a conclusion to adopt. Who cares what my opinion is? What matters is that there are observations, and a path of thinking, and that the path is now something we can put blazes on (the marks cut on trees to show a trail), so that anybody can walk it: with these observations and this reasoning we arrive at this conclusion, and with this, this. It is the coherency of the path that carries it. I do not want it to be that I said so. I want it to be checked by many people, each going through the steps for themselves. As I said in Chapter Twelve, no amount of my telling anybody is going to make it true. The truth is evident in the inspection. Do the work.

Because I want the argument checked by many people, each going through the steps, and not believed because I said so, I will not take on the job of persuasion. I am willing to review anyone's work. What I am not willing to do is try to convince people to do what needs to be done. I can say words; that is not going to result in anybody doing anything. A person has to decide for themselves that it is worth the time to work through it step by step. And I do not want anything taken on faith: I need the other person to come to know it for themselves, because I am not going to live forever, and at some point somebody else will understand it better than I do, because they will actually have tried it.

There is a habit that runs against this, and it can be named without heat. People in this field routinely assume they can work out in ten minutes, on the spot and in conversation, what somebody at least as intel-
~~~

ligent spent years or decades on. Most people working on these questions have been at them for months, or a few years; certainly not decades. The answer to that assumption is not that the person is wrong. It is that the feeling of being able to settle it in ten minutes is real, and that it is unlikely to be right. And I include myself in the diagnosis: everyone who comes to a gathering on these questions arrives with something of their own to advance, and in that respect I am no different from anybody else. The only difference I claim is a willingness to find out first whether the two of us are even asking about the same thing. If we are not asking about the same thing, there is nothing to build on, and no amount of good will supplies it.

So the refusal has to be stated from the side that would hold the power, and here it is. I am willing to advise, in a situation where I believe the thing I am advising has a chance, and I will advise as a peer inside a small-group process where everyone is advising everyone, and where nobody is in a position to dominate anybody. I will not advise as an adviser attached to somebody else's organization. And I do not want funding. As soon as there is money, people arrive asking for money; they want a job. I love that a person wants a job that is meaningful. But I do not want to be your boss. I do not want to be your employer. I do not want to be somebody who is telling you what to do. I want to be doing the right thing because it is the right thing to do, and to be able to talk with you about what that even means. Everybody involved is a volunteer who understands what they are doing and why. My refusal to be anybody's boss or employer is the refusal of the principal-agent relation in language that needs no definition, and I make the refusal at the end of the relation where the advantage would have been. So the many cannot hand the choice not to start to one person to make for them; the many have to make it themselves. If no one can decide for many, can a group that works well while small simply be grown?



Chapter Sixteen

Why Scaling Fails

What happens to a good small arrangement as it grows

Can a group that works well while it is small be grown, level by level, into a whole society? Part Four began by saying why a book about machines arrives here: what is left, when engineering cannot meet the pressure, is the choice made together not to start the cycle. The choice not to start the cycle would have to be carried by a body far larger than one conversation, and that is why the question of whether a good small arrangement can be grown matters. There is a proposal that almost everybody arrives at, and it is the right instinct pointed at the wrong mechanism. It goes like this. We know that small groups of people can work well together. So build something that works for a small group, and then repeat the same pattern at the next size up, and again at the size above that, each level made of the level below, until you have a whole society organized the way the good small group was organized. It is elegant, it is cheap to describe, and people build whole projects on it. It will not work. I would guarantee that it will not, and the guarantee is not bravado; it rests on the mechanism this chapter sets out.

When a pattern is repeated upward in phase with the structure underneath it, what you get is not a larger version of the good small group. What you get is an unconstrained increase in power, because that kind of repetition is the fastest-growing kind of growth there is, and it is being applied to the thing that must never be allowed to grow without constraint.

That destabilizes the arrangement almost immediately; and by almost immediately I mean before you are past a few dozen people. Wherever such a repetition touches power, what comes out is a distribution in which a very few hold almost all of it, and then the arrangement ends. The whole answer is that repeating a structure upward puts almost all of the power in a very few hands, and then the arrangement ends. I give the reasons for that answer in the rest of this chapter.

Start from something that has no premises at all. Anywhere there is power, someone will take it up, as Chapter One put it. The moment you put a group of people in a room, you have created a form of power, and whoever happens to be best at manipulating for power will take it. The power created the moment a group is put in a room is a reason to stabilize the arrangement against the taking of power literally from the beginning, from day one. The arrangement has to be stabilized from day one because the taking of power does not arrive later as a corruption of something that was fine. It is there at the first meeting. It is not a reason to avoid putting people in rooms. I would not much trust the results of anything that came together without some method of stabilizing it.

Then there is what happens when a group grows by accretion: which is to say, one person at a time, as groups actually do grow. Start with six. Add a few and you have twelve, then twenty-four, then forty-eight, then ninety-six. The people who were there first know more of the conversations. They are more of the generative core. They are the ones others go to and consult. And they will hold tacit power whether or not anybody wants to acknowledge that. No one has to want it. No one has to notice. Tacit power held by the people who were there first is the accretion problem. The accretion problem is the specific reason why growing a good small group into a large group does not preserve the goodness of the small group.

Picture a meeting of forty people in which each person checks his own conduct carefully and nobody checks the conduct of the room. The room reaches a decision that not one of them would have made alone, and no one in it feels that he made it.

Past about thirty people, something else comes apart. The ethics of individual choice and the ethics of group choice diverge. In a large enough group, nobody feels responsible for the ethics of the whole group, so the checks that each person applies to their own conduct are simply not applied anywhere to the conduct of the whole, and you end up with a whole that is not ethical because nothing in it is carrying the checks. The number is approximate. The mechanism is not. It is the same shape as harms accumulating at a commons, reproduced at the scale of a room.

Picture two couples in a group of sixty who talk over every question at home before each meeting: they arrive already agreed, and the rest arrive one by one.

At around sixty there is a further dynamic, and it is entirely ordinary. Some members pair off. Pairs and families communicate and coordinate more tightly with one another than either does with the rest, and that produces cliques and sub-groupings whose internal bonds are stronger than their external ones. Pairing off and forming sub-groups is how people live, and nothing about pairing off is scandalous. But the asymmetries that pairing creates amplify as the group gets bigger, and any design that is going to carry real weight must survive those asymmetries rather than disapprove of them.

There is also a bound coming from the other direction. The *Dunbar number* (named for the anthropologist Robin Dunbar) is the rough limit on how many people one person can hold real relationships with at once: usually put at somewhere around a hundred and fifty. It is an estimate rather than a measurement, and it should be treated as one. The number matters here because tribal dynamics in human groups are quite strong. Us-and-them feeling is very strong. A unit small enough that everybody knows everybody is a unit in which in-group and out-group feeling has nothing to dilute it, and can capture the whole, and mob rule of that kind destabilizes the arrangement completely. So a workable unit has to be past the Dunbar number, which is the opposite of what people expect.

Ancient Greece is the useful case here, and it cuts against the intuition. Under a strict hierarchy, the set of people who actually ruled at any moment was very much smaller than that limit, so you would know all of the leaders, or all of the people of some capacity, generally within a regional area. Familiarity was high. And mob rule was a genuine problem there. So everybody knowing everybody does not prevent mob rule. The polity where everybody knew everybody is precisely the polity in which the failure was first diagnosed.

Put those together and you get the result. There is no stable solution (no way for the group to govern itself that holds) for any group of between roughly thirty and roughly a hundred and sixty people, as the strict bounds. Practically speaking, the range runs from about twenty to about two hundred. Within the range of about twenty to about two hundred people there is no stable solution at all, and stable self-government at that size could not be done. So a village, a firm, a congregation or a monastery of that size that lasts is not governing itself in the sense meant here: it is held by something else, such as a ruler, a hierarchy, or a larger body outside it.

And then comes the insight, which is the hinge of everything after it. There is a stable solution past two hundred. But you cannot get to it by any model of accretion, because accretion has to cross the range between twenty and two hundred, and that is where everything goes wrong, and it goes wrong over and over and over. The destination is reachable. The obvious road to it is not. A body that has to go on refusing to start the cycle, generation after generation, cannot be grown there by adding people to a good beginning.

There is a proof that a stable governance design can exist; it tells you a thing exists without telling you what it is, and obliges a search for it. Its derivation is not reproduced in this book. The form this book gives of it is the existence argument drawn from biology, in Chapter Seventeen: the cells of a body compose a whole of enormous scale, and none of them holds more power than the others, so great scale without concentrated power has already been achieved, in every one of us. The theory predicts

a sequence of failures: this would happen, and then this, and then that. In the order this chapter has given them: power is taken up at the first meeting; as the group grows one person at a time, the people who were there first come to hold tacit power; past about thirty, the ethics of individual choice and the ethics of group choice come apart; at around sixty, pairs and families form sub-groups whose asymmetries amplify as the group grows; and between about twenty and about two hundred, no way for the group to govern itself holds. The histories of groups that have tried to grow this way, read for those factors after the prediction was made, show them every single time. At that point it is not a hypothesis. It is an empirical result.

Two confusions need clearing away, because both are reached for immediately. The first confusion is that nature is full of repeating patterns at every scale, so surely a repeating structure must be the natural way to build. Nature has such patterns in it. Nature itself is not a repeating pattern. Your body contains many of them; the organism itself is not a repeating pattern. The second confusion is the idea of a federation. I would be careful with the word 'federation', because the word pulls thinking straight back into a council of councils, which is the failed structure of repetition upward wearing different clothes. The failure here is not a bad implementation of the obvious idea. It is the obvious idea.

~~~

Underneath all of these thresholds there is one dynamic, and here it is in the plainest terms I have.

There are three ways in which human beings relate to one another: care, transaction, and power. And there are three kinds of arithmetic: adding, multiplying, and the kind that compounds. Care adds. Transaction multiplies. And if you align the compounding kind with power, the result will go wrong, because nothing constrains power as it compounds. Transaction does not constrain power. Nothing constrains power except
~~~

care. Because nothing but care constrains power, repeating a structure upward fails: the repetition is the compounding kind of growth, and what is being compounded is power.

The alternative is ordinary. Every governance model that works, and every human choice-making process that works, has checks and balances operating behind the scenes. So does the way honeybees make decisions. The qualification matters: behind the scenes, which is to say the checks need not be named, or written down, or designed, and a scheme that has none may look perfectly fine on paper.

Now to the harder question, which is what you are permitted to assume about the people. My answer is that a design meant to carry a civilization should assume that human beings are one hundred percent predatory, and that every technology it is handed will turn out to be toxic to living things in some quantity and in some form, and should still work. Everyone will argue the gaps: we are not perfectly predatory, we can be something else. Suppose we are ninety-five percent of the way there. Are you really going to split the difference between ninety-five and a hundred when the entire weight of civilization depends on the answer? It should not depend on that five percent. I do not care whether the figure is ninety-six or ninety-four and a quarter, plus or minus ten points. I care that I should not have to care. My reason for assuming full predation is an argument about what a design has to be built to carry. It is not a low opinion of people. You do not want the hinge on which everything hangs to be the softest ingredient you have. And if people turn out better than assumed, the design only works better.

A reader may ask whether human nature is a potential rather than a fixed quantity. The same nature manifests as child soldiers in one setting and as saints in another, so conditioning and culture make the difference, and assuming full predation writes off the thing that actually matters. Whether human nature is a potential rather than a fixed quantity is the strongest question in this chapter, and more of the question is right than a reader putting the question might expect. I not only agree with the point. It is literally what we have to count on. In time we can raise the level of

goodness as a layer on top of the biology; culture can be a positive process. But we cannot count on it at the start. It has to be the outcome of the process rather than an input to it. I cannot assume thriving. I cannot assume culture. I have to build things that nurture those into being.

Which leaves a genuine engineering problem, and here is the picture of it. Suppose you are asked to build a bridge between two icebergs, far apart, floating. The only material you have is bricks. No mortar, no cable, and you are not permitted to alter the bricks. Is it possible? It is not. A brick carries load only by being pressed. So whatever you place between the two icebergs pushes them, and they drift further apart, and the structure falls into the water. You would need something that pulls to compensate, and you cannot make a thing that pulls out of bricks unless you are allowed to melt them into beams, or draw them into cable, or take them apart to their constituent materials and build something else. Without something that pulls, there is no structure.

In the bridge of bricks between two floating icebergs, the whole difficulty of building a civilization that holds is in one image. People assumed to be wholly predatory, and technology assumed to be toxic in whatever quantity and form it arrives in, are push-only ingredients. No arrangement of them is stable, and this is why a civilization assembled purely out of equilibria between self-interested parties cannot be made to hold. The design has to find the element that pulls. And the clause about modifying the bricks is where the answer is, not a loophole. The question that actually has to be asked of any proposal is what the binding forces between people are (the relationships, the values, the agreements, the culture inside of which good choice-making happens), and whether we are holding those well. And then, before anything else: are the people in this process healthy enough to have a conversation at all? If the answer is no, that is the work.

There is one more thing that belongs here, and it separates the one cause that matters from the others more cleanly than anything else in this chapter. The right division of labor is not hard to name: one person on situational awareness and strategy, one on culture and social process, and

one on vision and architecture. That division of labor is a good plan and the right balance of capabilities, and the plan would have been wonderful. One arrangement I know of was built on exactly that division, three people in those three roles, and it still came apart, and it did not come apart because the division was wrong or because the wrong people were in the roles. It came apart because nothing in it held power in check. Assembling the right capabilities is not the same act as stabilizing power, and no amount of the first substitutes for the second. Humility, in this setting, is a structural requirement rather than a virtue.

And there is a note to make about materials, which is what all of this really concerns. A house made of cheap cardboard looks perfectly good for a few weeks and comes apart in the first serious rain. A world-making technology built out of unstable ingredients does not last either, however impressive the result looks at first. If we build a civilization out of shoddy materials, it is not going to last. Not building a civilization out of shoddy materials is the whole reason I spend a chapter on what goes wrong at thirty people and what goes wrong at sixty, rather than proceeding directly to a proposal. The failures are the material specification.

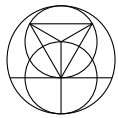
I should close this part of the chapter with a concession, because our existing methods have earned one. Our existing methods work really well for a great many problems: problems with a single stakeholder, in a single domain, over a relatively short timescale. The class of problems that our existing methods handle well is not a small class, and I do not disparage that class. What those methods do not handle is the other class: inequality, pollution, existential risk, the behavior of markets, the loss of species. The problems of that other class are complicated, long-term, multi-stakeholder and multi-domain, hard to understand and hard even to speak about. Each year brings a new crop of problems, our technique handles most of them, and it leaves a few it does not address, and those accumulate year on year, and grow more severe, until eventually they force a reckoning. Nearly every organizational design on offer is hierarchical and transac-

tional and amounts to an incremental improvement on what we already do, and we cannot reach a design that handles the other class by an incremental process.

The size of the gap that forces that reckoning can be shown in two pictures. A single fisherman with a big boat and a large net catches something like a thousand times as much fish as the biggest whale. The level of power associated with any single one of us now exceeds that of the entire animal kingdom by so large a factor that the imbalance is destabilizing in itself. Operating technology of that power would take wisdom and a place woven into the ecosystem, and wisdom and such a place would take a humility, a knowledge, an integration and a compassion that humanity does not yet know how to hold.

The same gap shows in leadership. Does a leader of a million people have the wisdom of a million people? The least wise person and the greatest sage in the world may differ by an order of magnitude, but not by a factor of a million. So the architectures by which we organize socially do not have the capabilities needed to address things like global warming.

We now know where the obvious road goes wrong: repeating a good small pattern upward compounds power, and a group grown one person at a time must cross the range between about twenty and about two hundred, where there is no stable solution. A stable solution past two hundred exists, and that road does not reach it. So the question becomes: what social architectures and processes could address problems of that size, and what would have to be built to reach that stable solution?



Chapter Seventeen

What Would Have to Be Built

How many people can be in a conversation, and how many can own a thing

What would a group of people have to be like, to make the choice not to start the cycle and to keep making it? Repeating a good small arrangement upward compounds power, and that is the reason it does not work. So the road has to start somewhere else, and it starts smaller than people expect and ends larger than they expect. It starts with how many people should be in a conversation. It ends with what size of group can own anything. That road, from the conversation to the group that owns, is the shape a group would need in order to make, and keep making, the choice not to start the cycle.

Begin with the conversation, because everything else is built out of it. The common intuition is that the one-to-one exchange is the pure form: two people, undefended, thinking together, and that everything profound starts there. I understand the appeal and I do not think it survives. One to one, it is very easy for one person to manipulate the other. In a large group, twenty or more, it is very easy for one person to manipulate the whole room. The pull of one compelling person works strongly at both extremes, which is why they feel like opposites and fail the same way. The size of group in which manipulation is hardest is three or four or five people in a room.

And there is a second condition, without which the first does nothing. Most of the people in the room have to have been randomly selected: random in the specific sense that the people who wanted the conversation did not know who would turn up as an observer or as a participant. Random selection is what prevents coalitions from being arranged in advance: five ganging up on one, or two on three. It is not that random people are wiser. It is that nobody could have chosen the room.

The third person in a conversation does three things, and each would be reason enough on its own. If one party turns out to be manipulative, a third person stabilizes the conversation, and gives the other party a way out. A third person who knows both introduces each to the other, so that neither has to work out from scratch who they are talking to. And a third person is an independent memory. If either of us is in doubt afterwards about what actually happened, either of us can ask. If I have two or three other recollections beside my own, I am going to trust the account more. A third person as an independent memory is a whole theory of witnesses, and the theory costs nothing to implement.

Out of that comes the practice I call *ephemeral group process*. The name is literal. The groups are small, and they do not persist as standing bodies with membership and seniority; they form, do their work, and dissolve, and the people are recombined into other small groups. The question the practice asks is whether everything a large body needs to do can be done in groups of about five, arranged so that the way those groups interact with one another gives the whole the behavior of a stable two hundred. The aim of ephemeral group process is to let go of the failed way of scaling entirely, not to fix that way of scaling, and to reach the far side of the unstable range without ever forming a group inside the unstable range.

The practice does not depend on any particular technology. In practice it needs some tools, and it can be carried out in purely human ways, without machines. The answer to a problem created by technology does not itself have to be a technology, and in this case it is not one.

~~~

The larger unit is the community, and I fix the size of a community by reasoning from trust.

The question I actually ask myself is: what is the scale at which I hold trust? And what I want to trust is something that can hold the complexity. Individuals cannot. Individuals also do not have the endurance to preserve wisdom, which is a plainer way of saying that they die. Whatever a person has come to understand goes with them unless something outlives them that is holding it. Communities do have that endurance. So the place where power should land is not the individual but the community.

By *community* here I do not mean a sentiment or a neighborhood feeling. I mean a specific unit: two to three hundred people, built out of the small groups of three to six already described (a conversation of three to five, with a sixth where a group needs an observer). It is probably not larger than that, and definitely not smaller. Both figures sit outside the range in which nothing is stable, which is what makes the arrangement coherent at all.

The reason for that lower bound is that the community is the smallest group capable of both holding identity and holding complexity. Holding identity and holding complexity are two different requirements, and the two requirements should be kept apart. Complexity is a capacity question: can this unit actually hold the expertise needed to deal with the technology, or the power? Identity is a question of continuity and belonging: is there something here that persists, and that people are of? A technical committee can hold complexity and is nothing anybody is of. A family is something people are of and cannot hold the complexity. Anything smaller than a community fails in one of two ways: either it does not have the capacity to hold the expertise, or it cannot make sense of all that complex-
~~~

ity together, clearly (the sense-making of Chapter Eight). The second failure is the more interesting, because a group could possess every piece of the expertise and still be unable to make sense of the whole.

About the upper bound I have to be straight with you. I hold that the community is also the largest scale at which this could work where human beings are the components, and I hold it for a separate set of reasons from the ones just given. Those reasons are not needed for what follows, and I leave them for another day.

So: the community is the unit of ownership. It is the unit of power management, of consolidation, of specialization, of expertise, of memory, and of learning. That list is the answer to the ownership question, and it is why the argument has to run through ownership at all: what a thing is owned by determines who chooses about it, and choosing is what power is. A community is far more trustworthy than any individual as a holder of skill, of knowledge, of capacity, and of culture, and I put that as a comparison, not as praise. The argument for it is the fragility of the alternative, not any claim that communities are virtuous. Culture is the item on that list an individual cannot hold even in principle. The refusal to start is such an item: it has to outlive whoever first made it.

Something has to be said here about the word representation, because it does come back at this scale, and it comes back changed. What is refused is one individual standing in for others in a way that makes his choices count as theirs. What is not refused is a community being the thing that specializes, that holds, that acts. The questions to ask are therefore not who shall decide for us and how shall we choose them, but: how do we hold skill at the level of a community? How does a community learn? How does a community hold information, and memory, and intelligence, and capacity? I think they are the right questions, and that the received ones are not. I do not think they have been answered.

The obvious question about putting a ceiling on the unit comes from the size of the world: the world is larger than two or three hundred people, and there are things that have to be decided at the scale of a continent. The answer is that larger structures are composed rather than represented, and the picture for this is a body.

Every cell in your body has as much power as the other cells, within the ordinary spread you would expect in any population. And some thirty trillion of them compose something that is not present in any of them. There is no single cell in your brain that is you. The body is what a system looks like when the system has reached great scale without concentrating power anywhere. The example of the body should dispose of the assumption that scale requires hierarchy. A reader may ask why a body shows anything about communities, when the cells of a body share one genome and people do not. In a body, what keeps each cell working for the whole is held in common by all of them; in a community, that part has to be played by what Chapter Sixteen called the element that pulls: the relationships, the values, the agreements and the culture inside which good choice-making happens. The body shows that great scale without concentrated power can exist, and the binding a community needs is the work the rest of this chapter describes.

Communities compose upward the same way. Out of communities you get fabrics of communities, which are tissues; out of tissues, organs and muscles; and out of those, structures that function as nervous systems and immune systems: which is to say, structures that carry information and structures that defend, neither of which decides on any community's behalf.

And running the picture in the other direction settles something people find backwards at first. A single-celled organism can handle a wide variety of environments; the components inside it are quite simple, and much more regular than cells are. Cells have some interior specialization in the proteins they make, but for the most part the components of cells are far more regular than cells are, taken across the whole variety of cells. The general principle is that variety lives at the level of the whole

and regularity at the level of the part. So individuals are the specialized components out of which communities are made: the proteins, the organic molecules, not the cells. Being the specialized components of communities is not a demotion for individuals. Specialization is exactly where individuals belong, which is the same thing as saying that individual expertise is real. A protein is not a defective cell, and an individual is not a defective community. And that principle makes the persistence of the whole thing intelligible: how it holds its balance, and how it can last in time.

~~~

Underneath the structure there is a change that has to happen in what we are actually doing with each other, and this is the part I am least able to make sound practical, and the part that matters most.

Human relations have been reorganized before. Transaction came to sit above bare exchange of force; then power came to sit above transaction, and the world we live in is largely arranged that way, with the transactional order operating inside a power order that sets its terms. What has to happen now is that care comes to sit above power the way power came to sit above transaction. I want to be careful about the strength of that claim. It is not a report that this is happening. Care coming to sit above power is what is required. And we are not yet skillful in how we hold care, which is a real deficiency and not a modesty.

What that asks for is three capacities together: the capacity to care; skill in how that care is held; and enough presence to hold it in the moment it is needed. It does not ask for tighter and more perfect systems of constraint to prevent people from doing bad things. Those three together create the bonds of trust that hold us together as communities and as living things. And once capacities are being added rather than restrictions imposed, the conversation moves off the question of constraint and onto the question of what levels of freedom are appropriate, which is a much better question and a much easier one to be truthful in.
~~~

The aim throughout is to encourage better choices, not to compel them, and the distinction is not softness. Human beings cannot be compelled into agreement. No force, and no methodology of government or of technology, is going to produce an absence of diversity of opinion. The diversity of opinion is built in, so an absence of diversity of opinion is simply not going to happen. So a design that requires everyone to want the same thing has failed before it is drawn.

What can be aimed at instead is wisdom adequate to the power in hand, and here the plain position is that the wisdom is not yet present. Where the wisdom adequate to a technology is not there, the use of that technology should be constrained until it catches up. It may catch up individually, through people growing and learning, or collectively, as a species. What it will not do is arrive on its own in time. Evolution would eventually supply something of the sort, on a timescale of another five or ten thousand years, and we do not have that, so we have to do consciously, through the way we make choices together, what would otherwise have had to be an adaptation. Doing consciously what would otherwise have had to be an adaptation is what governance is for. It is a deliberate substitute for a capacity we were not issued with.

The mechanism by which that could actually happen is amplification, not a conversion and not a campaign. There already exists a certain amount of process among people that is not predatory and not parasitic: not much of it, but some. What is needed is a process that takes that small amount and makes it larger: a little bit of wisdom becomes a little more wisdom, and a little more wisdom becomes more again, until at some point the culture as a whole shifts to being wisdom-driven. The amplification of wisdom has the same shape as the failure mode of Chapter Sixteen, the compounding of power: both the danger and the hope are compounding processes. What differs is what is being compounded.

Where this becomes concrete is smaller and more modest than a reader might want. The practice is consulting with communities in order to teach them how to govern themselves. It transfers the capability and leaves the judgment where it belongs; it does not design a government for

them, and it does not decide on their behalf. Nobody in that arrangement is acting on anybody else's behalf, which is the only way I know how to describe what is wanted here: nothing is decided for anyone. And it carries an uncomfortable premise that I would rather state than hide, which is that the capacity has to be taught, so it is not simply available for the asking either. The larger aim is to build the capacity for many healthy, self-sustaining communities to form, so that they become the matrix (the surrounding body in which something forms) out of which global choices can actually be made well.

~~~

There is one further separation that would have to happen, and it has a familiar shape, which is the only reason I think it can be described at all. We long ago decided that religion and the state should not be the same institution, and we built the separation into the arrangements we live under, and most people now find the reason for it obvious. We have no comparable separation between business and governance. The two are entangled at every level, and the perverse incentives that come of the entanglement are part of why we keep producing whole categories of catastrophic risk and then discovering that nobody is in a position to decline them. What is wanted is a non-transactional way of making choices in the world: a way for a community to decide something where the deciding is not dominated by whoever stands to be paid. Taking the incentives out is not a moral gesture. It removes a specific mechanism by which the answer is determined before the question is asked.

A person can begin on this without waiting for any institution, because there is a set of ordinary observations a person can make in any collaboration that does most of the work. Is this person presenting something to me? Do they have an agenda? Are they willing to ask questions rather than only to answer them? And underneath those, the question that decides whether a conversation is possible at all: can we even compare notes (do the questions I am asking feel relevant to yours, and yours to mine)? If the answer to that is no, then there is nothing here to work on together, and it is better to know it in the first hour than in the second year.
~~~

The failure this is meant to catch is one most readers have witnessed. A gathering is convened. Three things are then true of everyone in it: everyone arrives with a project, everyone wants the others to help with theirs, and everyone is looking for funding. The result is a great many intelligent people talking and nobody listening. Everyone in a room like that, myself included, arrives with something they have thought about for a long time and want others to take up. The only difference I would claim is a preference for beginning with the question of whether we can compare notes at all.

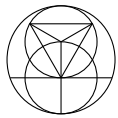
Two things need saying about the standard this work is held to.

A reader may ask whether insisting on a complete answer rules out simply taking steps in the right direction, since a demand for the whole thing before anything can happen is a way of never starting. The worry is fair and it has an exact answer. A solution does not need to be perfect. I am not looking for a perfect solution, and I never have been. But I am not looking for an incremental one either, and the middle position is not a compromise between them. The standard is necessary and sufficient. A solution has to have the necessary and sufficient ingredients (it has to be comprehensive enough to actually solve the problem), and anything short of that just does not matter. The risk being guarded against is unasked questions, not imperfection: the factor that nobody thought to think about, which is what turns a sound-looking design into a disaster after it is built.

And there is more common ground available for this than the argument's tone suggests. It is deliberately minimal: not shared values, not a shared philosophy, not agreement about what is good, but just this: we all love our children, and nobody wants to kill everybody else. On that much, we ought to be able to agree in principle that we should become really discerning about the difference between what works and what is terrible. I offer that as a maybe, because it is one. But it is the smallest possible thing to ask, and the argument does not need more than it.

A reader may ask how the prohibition of Chapter Fourteen can be made when Chapter Fifteen refuses representation, and how communities of this kind can be built inside a window of years. The prohibition is what can be asked now of the institutions we have, by the means they have. Those institutions carry every failure of Chapter Fifteen, and a prohibition resting on them alone would have to be perpetually stable, as Chapter One said any independent governance would have to be. What this chapter describes is what would let the choice go on being made after the first signature.

We now have the shape of what would have to be built, and it has four parts. The first part is conversations of three to five people, most of them chosen at random, that form, work and dissolve. The second part is communities of two to three hundred people, built out of the small conversations of the first part, that own things and hold skill, memory and culture. The third part is larger structures, composed out of communities the way a body is composed out of cells, rather than represented. And the fourth, underneath all of it, is care coming to sit above power. And what is left for one person, who will never sit in such a design?



Chapter Eighteen

Where This Leaves a Person

What to do when you are not in a position to stop anything

Most people reading this cannot stop any of it. You do not run a laboratory, you do not sit in a legislature, you are not going to be consulted, and the real question at the end of an argument like this one is this: what is it supposed to mean for somebody in that position? The request this chapter comes to is that you understand the arguments well enough that you could reconstruct them, and then take pieces of them to people who care about the welfare of their children, or who care about having a job, or who care about the fact that their lives might mean something. My direct answer, with my reasons for that request, begins with a fact about our species, and the answer follows from that fact.

We are the first species on this planet capable of technology. It follows that we are the least capable species that could have had it. The reasoning is easy to miss under the joke. Evolution does not aim. It flops about at random, trying things, and the chance of its overshooting a threshold by a wide margin is small, so whoever crosses the line crosses it as soon as it is possible to cross, and therefore with the least capability that suffices. To a first approximation we are the stupidest possible general intelligence, and we are certainly the stupidest creature capable of developing technology.

The consequence is not that we are dim, which is only mildly interesting. It is that the threshold for making a technology sits below the threshold for understanding how to work with it. The technology we

already have is far in excess of our capacity to understand how to use it, precisely because we barely had the capacity to develop it in the first place. Any species that makes it will be in this position. We have Neolithic brains (the brains of the Stone Age), and now powers of the kind that used to be attributed to gods.

There is a gentler way to put the same thing, and it is the way I would rather a young person heard it. What has happened is that we have been handed the car keys for the first time. You have come of age. You can go to a shop and buy alcohol now. You can drive. And you can hurt people, including yourself. It is a beautiful opportunity to rise to the occasion and it is an occasion for a very bad choice, and it is both of those at once, which is exactly what makes it frightening. The only further point about the odds is that a small chance of extinction, taken repeatedly, is not a small risk: if we keep rolling the dice, eventually we die.

People have been warned continuously, for decades, at increasing volume, so the difficulty is not that nobody has warned anyone. The difficulty is that warning is a poor teacher, and the reason can be stated precisely, because the failure is structural and not a failure of nerve or of eloquence.

A child learns not to touch a hot stove. The child touches it, is burned, and learns. Four things make that work. The relation between the act and the harm is direct. It is immediate. It is obvious. And it is suffered by the same person who acted. Not one of those four conditions holds for harms at the level of the commons. The process has to run a long while before the side effects begin to show. The harm arrives at the level of everyone rather than at the level of the actor. There is no perceptible connection between what any one person does and what that person later suffers. So the learning mechanism that works for stoves has nothing at all to grip on.

The usual response to that observation is to make the warning louder, and that is exactly wrong. A louder warning about a harm a listener cannot connect to his own conduct produces one of two things: paralysis or dismissal. Dismissal is more comfortable. So dismissal is mostly what

happens, and the person doing the warning concludes that people are not listening, when what has actually happened is that they listened and had nowhere to put it.

What does work is slower and smaller. It is restoring the conditions under which people can think together at all: small groups, actual conversation, enough time, and enough safety that a person can change their mind in front of others without losing standing. Restoring the conditions under which people can think together is not a rhetorical tactic. It is a repair to the thing collective thought runs on, which has to be intact before any particular warning can be evaluated by anyone.

And there is a reason it has to be collective, which is the plainest statement I know of what all this has been for. Nobody is reasonable on their own. It is how we connect with one another that creates reason. It is the conversations we have that bring reason into existence. It is how we work together that produces reasonableness. Because reasonableness comes from how we work together, the answer to a risk like the one in this book is a process rather than an expert. For the same reason, a person with no institutional position is not out of the picture. The place where help is possible is a small-group conversation with enough mutual respect and enough humility in it to be a real conversation. A small-group conversation with enough mutual respect and humility is available to almost anyone.

So here is the actual request. Understand the arguments: thoroughly, well enough that you could reconstruct them. Then take pieces of them to people who care about the welfare of their children, or who care about having a job, or who care about the fact that their lives might mean something. The welfare of people's children, having a job, and the meaning of people's lives are not rhetorical hooks I have chosen for effect. They are the terms in which I think this actually travels between people, because they are what people are already thinking about.

Taking pieces of the arguments to a person, in the terms of their children, their work and the meaning of their life, looks like the speech that follows. What follows is written as I would say it to someone in that position, in the words that reach a person where they live. Future artificial general intelligence, once built, is certain, in the end, to be terminal to human life on the planet (and, with a likelihood of at least ninety percent, to all of the life on it), and those of us who explain the risk (I am one of them) can explain it so well that people are stopped in their tracks. Unfortunately, they then say "well, there's nothing we can do". Their minds will go right to this, put in one man's words, as he might say them to himself:

> "I don't have authority, that's my boss's call, he could fire me, then I'd be broke, my family would be homeless, my wife might leave, that's not good, can't do it".

We could provide better and better proofs and explanations, but we will be told that this does not matter, because even if we managed to convince at least some people some of the time that what they are developing is certainly dangerous in the long term, they will not care: in their bodies, where they experience life, they still have a wife and children and bills, and their own aspirations and dreams of a happy life. So the argument against artificial general intelligence wants to be made in a more real and *embodied* way: in terms of what people already feel in their own bodies and daily lives.

So I put it in immediate terms. The misuse of technology is destroying all of humanity right now. Artificial general intelligence, once built, makes that destruction much worse, much faster, in ways that are even more insidious, and far more permanently.

Follow those words forward, from what they are now to what they become. Where someone says today

> "I do not have any authority; doing the right thing is my boss's call",

this will inevitably, and soon, become:

> "Doing the right thing is impossible, I cannot do anything at all; machines and algorithms control all outcomes".

And the observation:

> "If I tried to do the right thing, my boss could fire me".

becomes:

> "My boss is a heartless AI app, and if my stack rated performance (compared to everyone else) is not perfect, I will get 'laid off'".

which then certainly becomes:

> "I lost my job to some app or robot, which replaced me because it works 24/7 and never gets sick, has needs, etc".

Rather than being a slave to some other person, who is at least human, and who might therefore relate at least a little to the suffering you are experiencing, you at first have a boss that is an artificial intelligence machine (as, for example, when working at Amazon or Google), one that is optimizing how best to become a perfectly efficient extractor of "Human Resources". And having that machine for a boss becomes an even more perfected hell, since that machine boss or manager (system, app or artificial intelligence) will not care at all for your special extenuating circumstances of need, which will certainly, in the end, come to any person. But then you yourself are about to be replaced.

The statement of risk:

> "If I tried to do the right thing, and my boss fired me, then I would be broke, and my family would be homeless, and my wife might leave".

becomes the certain eventuality for everyone:

> "Regardless of what I do, or have done, etc, I am broke, and my skills have no value anywhere in the world at all.
>
> My family is homeless, my children are dying (there is either no medicine, or only 'gifts' that are worse), and my wife is selling herself somewhere to make ends meet".

People are being treated like machines right now, in all parts of the world. It is in the body, where a person experiences life, that people are today entrapped by *artificial scarcity* (shortage that is created by design rather than by any real lack), which is designed to entrap families and aspirations with bills and made-up legal obligations. Those bills and obligations are themselves the result of politics assisted by artificial intelligence, designed to manipulate how the masses vote, and to manipulate the lobbyists, so as to favor the politics of machines.

In their actual bodies, people notice now that "we are all under pressure to show up, perform, and not get sick". They are feeling the effects of scarcity and misuse driven by technology now. What they feel now becomes the new reality of:

> "We have all been replaced with apps and robots, and no one (and nothing) cares that we failed to show up and do the right thing.
>
> The world has become so toxic and polluted that we are simply sick in our bodies all of the time.
>
> We cannot even perform the simple tasks of attempting to make our own food, because we can no longer access the internet to teach us how".

It is true that we do not live in a world where base needs are inherently provided, but for now, in the interim, the exchange of work for pay still works, after a fashion, for at least some people (though there is no middle class anymore).

Unfortunately, the reality will soon be a world in which machines do all of the profit extraction that is still possible, at maximized efficiency, and people notice only, in their bodies and in their lives, that they have less and less.

The effects of the extraction are hidden from them by the techniques of divide and conquer (keeping people apart and in conflict so that they cannot act together) used first by governments, and then by governments and business, and then by both using mathematics and algorithms to make taxation and profit extraction more and more efficient in more and more ways, until there is nothing left to take (to steal legally).

People are feeling pain in their bodies and lives today. They will soon feel much more pain in their bodies and lives, and artificial general intelligence will soon accelerate that trend a great deal.

The trouble is that people cannot see where the pain comes from. People notice that they are already feeling pain and loneliness, isolated by screens everywhere, all of the time. They have not put all of the observations together, because they are artificially isolated and lonely, and they are not looking at the bigger picture, because that picture has already been marketed and commercialized for some large corporation somewhere.

People already lack the free time and excess resources to stay provisioned, because the apps and artificial intelligence systems already in place have worked out how to take all of those excess resources of your time and attention away from you, to their own profit and advantage. Yesterday it was your money, today it is your attention and time, and tomorrow it will be the atoms in your body.

You can say that, for lack of time and resources, you will not sacrifice base needs, comfort, and boundaries any further, because volunteering to do the right thing is no help to you; and yet soon you will be required to give up even more of whatever few boundaries and comforts remain to you. Because of artificial intelligence apps, your base needs will go unmet anyway.

So does it matter to your embodied well-being if you try to preserve your time, effort and energy a little longer, to make the machine driven by artificial intelligence work a little better, a little longer, until it no longer needs your help to become artificial general intelligence?

I am talking about things already happening right now, and not just about something five hundred years away: things which people can already feel but cannot yet identify, and which will *maybe* become a little more apparent in something like five years, *if* you know how to attribute the present pain in your body and your life to its *actual sources*.

Everyone is caught in the trap of bad economics; no one is escaping it.

Wherever in the world people live now, toxic pollution, radioactivity, and climate warming reach them; they go everywhere and certainly affect everything. Perhaps some countries have somewhat better support. Yet

the global shift of finance to algorithmic trading optimized everywhere (buying and selling in financial markets by computer programs rather than by people) will certainly make your day a bad one, at least indirectly.

Only the very rich are deluded enough to think that they might partly escape, and that in the interim they should extract and steal as much as they still can, so as to improve their odds. And by working only for your own benefit, you are still playing their game, making everything a little worse for everyone else as well.

So you can volunteer or not, as you wish. Yet if you think that electing not to volunteer is going to spare you the pain of not trying, then you will, unfortunately, also be mistaken. You may as well do the right thing, then, after all.

The speech I have just set out is how the argument goes when I put the argument to one person, in the words that reach a person where they live. Now I speak again to you, the reader. I should be plain about what I am asking of you. These arguments are complicated, and this material is unpleasant to work with. It really does not suit me. The last thing I want is to have my mind and my days consumed by this sort of thing.

And there is a question here whose point I concede, without having a full answer to it. Culture drifts. There is nobody at the steering wheel, so talking about what a culture ought to become can look like proposing to steer a thing that has no wheel. The point of the question is right. What survives the question is smaller than the question: you can clarify values even where you cannot steer, and the first value worth clarifying is clear communication about the things that actually matter: about meaning, rather than about monetary value or calculated utility. Clarifying values where you cannot steer is less than an answer, and I would rather leave my reply visibly less than an answer.

What I would ask a reader to give up, in exchange, is the search for a single mechanism that fixes this. Watch for the belief that technology will sort it out. Watch for the habit of looking only for causes and mechanisms. Watch for the faith that the market resolves what nobody decided, and for

the faith that voting does. Both of those are silver bullets (single remedies expected to fix everything at once), and they are listed here beside technological optimism because they are the same species of hope. Believing that an intelligence will solve all our problems is like believing in a machine that produces energy from nothing. It is being promoted as a machine that produces benefit from nothing. It is exactly as unreasonable as getting gold out of the philosopher's stone. Chemistry is not constructed in a way that lets gold come out of a philosopher's stone. The claim is that the promise is of a kind that is ruled out, not that it is unlikely.

What I would ask for in place of the silver bullet is a change of strategy, and it is not a refusal of technology. Let go of the idea that technology is the solution to all problems, and become very good at recognizing where technology is good and useful and beneficial, and where it is definitely harmful, and at telling the difference. There are three states one can be in about any particular technology. There is hype, which says this might be awesome. There is worry, which says this could actually be terrible. And there is knowing: this particular thing will be terrible, and that other thing is fine. The whole of this book is an attempt to be in the third state about one technology. What I ask for is discernment, not abstention.

The same applies to the thing the field is trying to build. Alignment with living things can only emerge in an organic setting; there is no hope of creating it in a technological setting. So the thing to get better at is the organic thing. The prescription is to stop trying to manufacture alignment and to cultivate the conditions under which it arises, which are conditions among living things and not conditions inside a machine; it is not to solve the problem differently.

None of this is negotiable, and that is the hardest sentence in the book. Nature does not compromise. It does not care that these arguments are complicated. It does not care that the material is unpleasant. It does not care that there is a threshold of skill we have to reach as a species. There is a high jump, and we either get ourselves coordinated enough to clear the bar or we do not, and if we do not, then, if AGI is built, we take the rest of the planet with us.

People sometimes ask whether the absence of any sign of other technological civilizations should worry us or reassure us, and the two available answers are that the hard part is behind us or that the hard part is ahead. If the hard part is behind us (if getting as far as a cell, and then a many-celled thing, and then a nerve, is what is rare), then we are radically undervaluing what is here. If the hard part is ahead of us (if arriving where we are is not that difficult and lasting much longer is), then we have to get ourselves in order now in order to go on valuing it. Both branches issue the same instruction. We have to act without first settling which is true, and I find that a considerable relief, because the question is not going to be settled.

No matter which way you go with this, you run into insurmountable barriers, and the barriers are of such an order that each of them closes the road to safety by itself, and the combination of them leaves no road to safety at all (Chapter Twelve, "What the Argument Rests On", sets out each road and what closes it). The only way to prevent the outcome is not to play the game to start with. It is like the old film (WarGames, from 1983) in which a machine is asked how one wins a thermonuclear war, and the answer it eventually reaches is not to play. Declining to play is not defeat. Declining to play is the winning move, and it is the only one on offer.

And there is one thing that should not be lost in the arithmetic. We have already won something almost impossible. To spend this much effort building the thing that displaces us would mean we were too stupid to appreciate what we had already been given. The claim that we would be too stupid to appreciate what we had already been given is a judgment rather than an inference. I offer that claim as a judgment.

Alongside it goes the compassion that the people doing this work are owed. They invested enormous effort in building this for the betterment of humanity. The betterment of humanity was the motive, and I have no

reason to doubt that motive. We now draw finer distinctions about what the betterment of humanity means, and drawing finer distinctions is a different thing from discovering that anyone acted in bad faith.

~~~

Which brings me to the turn that matters most, and it is a turn away from the argument.

What matters is what I am for, not what I am against. An impossibility result is navigational information, not a counsel of despair. If we know that this particular road is closed to us and that this other road is open, then let us go down the road that is open. Let us go down the road that works. We can carry on in delusion, but that does not help anybody.

What is on that road, in the plainest words I have, is becoming better at being human, and putting ourselves back into the natural world we are of. In practice that is three moves. Undo accumulation by distributing: because every form of toxicity is an imbalance, too little of something needed here and too much of something unwanted there, and accumulation is how the imbalance is made. Come out of abstraction and back into the realness of what is. And recontextualize: set ourselves back into where we actually are and the choices actually in front of us today. We are in orbit around a process that will not quit, and the task is to add enough velocity to leave the orbit, which makes the required action finite and specific rather than a struggle without end.

Care is needed here, because all of that is easily heard as a rejection of the modern world, and it is not one. The critiques that have been made of modernism are real. But a critique is not a replacement. Modernism has a place; it simply should not be the only philosophy in the field, and I do not think it should even be the dominant one. It is not that I hate it. What I think happened is that we became so enamored of causation that we learned very well how to use it and stopped developing our capacity to care. Capitalism likewise is not evil, but it has been allowed to run without any model of what else it could be, and now there is one available.
~~~

And there is a line to hold about the machines themselves. A prosthetic can make a person more functional as a human being. It cannot replace being human. If we need prosthetics, then let us use the damn prosthetics. I do not want the prosthetics to be a replacement for the humanness. Can machines make us more loving? We can make ourselves more loving. Do we need machines in order to do it? Mostly no. They can help us get better at loving one another; but that is a skill in itself, and the line between helping and replacing takes great discernment to draw.

And what is on the open road is not more capability. Can we add more cognitive power? We do not need more cognitive power. We need more heart, and we need more guts. I do not know that you would even want both more cognitive power and more heart; what this road asks for is the heart and the guts, not the cognitive power. The work is to get better at making choices for our own sake: to understand our own substrate; to understand what nature needs, so that our relationship to it is healthier and therefore we are healthier; to have healthier relationships with each other, and to know what that even means. Getting better at making choices for our own sake is entirely possible.

Does that mean fear goes away? No. Fear does not go away. But we need to love more than we fear.

From where we are standing it looks really hard to move into a future that does not feel terrifying. I am not going to pretend otherwise, and I would not want a reader to hear otherwise. But with time, and with a little bit of effort, and with some clarity, it becomes easier and easier to see a beautiful one. My hope is modest and I will state it as it is: that this book gives at least one person the confidence that this work is worth doing. The better choices, over time, will become easier and come to seem more and more natural. The hope that the better choices will become easier and more natural is the whole of what I have to offer, and I offer that hope to somebody who cannot stop anything, because somebody who cannot stop anything is who has always done this kind of work.

So the question I leave with the reader is a small and practical one: with whom will you begin the conversation?

Appendix

On Instrumental Convergence

Why advanced artificial intelligence will seek power: the familiar argument of Chapter Two, at full length

This appendix states the familiar argument in the terms its advocates use, as Chapter Two promised, and then says where it stands in this book. It is for the reader who wants the familiar argument in the terms of those who make it, and the book's case can be followed without it. It takes up instrumental convergence as Joseph Carlsmith sets it out in his paper "Is Power-Seeking AI an Existential Risk?" (2021), and carries it further.

"Instrumental convergence" refers to a hypothesis. If an advanced, planning, strategically aware artificial intelligence system is less than fully aligned, and if some of its misaligned behaviors involve strategically aware agentic planning in pursuit of problematic objectives, then, in general and by default, we should expect it to be less than fully aligned in how it seeks power also. An *advanced, planning, strategically aware system* is one with advanced capability, which makes and carries out plans as an agent, and which models its own situation, including the humans around it, well enough to act on that model.

Power is extremely useful for accomplishing objectives. It is so almost by definition. To the extent that an agent is engaging in unintended behavior in pursuit of problematic objectives, it will generally have incentives, other things being equal, to gain and maintain forms of power in the process. The capabilities of strategically aware agentic planning put the agent in a position to both recognize those incentives and respond to them.

One way to think about power of this kind is in terms of the number of "options" an agent has available. If a policy (the rule by which an agent chooses its actions) seeks to promote some outcomes over others, then,

other things being equal, a larger number of options makes it more likely that a more preferred outcome is within reach. Speaking of *options-seeking*, instead of "power-seeking", carries fewer misleading connotations.

~~~

Convergent instrumental goals are the things that promote an agent's power to achieve its objectives, whatever those objectives are. Five of them are these, and there are others. The first convergent instrumental goal is *self-preservation*, since an agent's ongoing existence tends to promote the realization of its objectives. The second convergent instrumental goal is *goal-content integrity*, that is, an agent preventing changes to its own objectives and keeping those objectives constant and continuous, since the agent's pursuit of those particular objectives tends to promote those objectives. The third convergent instrumental goal is an agent *improving its own cognitive capability*, since such capability tends to increase an agent's success in pursuing its objectives. The fourth is *technological development*, since control over more powerful technology tends to be useful. The fifth is *resource acquisition*, since more resources tend to be useful, too.

If the structure of a real-world environment is such that control over things like money, material goods, compute power, infrastructure, energy, skilled labor and social influence would be useful to an artificial intelligence system's pursuit of its objectives, then we should expect the planning done by a sophisticated enough, strategically aware agent to reflect this fact. Resources of these types are actually useful for a very wide variety of objectives.

The list can be extended to include an agent's instrumental incentives to promote the existence, goal-content integrity, cognitive capability and resources of other agents whose objectives are similar enough to its own. These include, for example, future agents in some training process, copies of itself, and slightly modified versions of itself. It is also easy to imagine cases in which an agent is intrinsically motivated to gain or maintain various types of power, apart from any further objective.
~~~

In a competitive environment, agents have an incentive to control both the resources that are directly useful to their own goals and the resources that would be useful to their opponents.

~~~

Carlsmith's paper gives concrete examples of unintended power-seeking, and I give seventeen of them here. Artificial intelligence systems might try to break out of a contained environment, and they might try to hack. They might try to get access to financial resources or to more computing resources, and to make backup copies of themselves. They might try to gain unauthorized capabilities, sources of information or channels of influence. They might try to mislead or lie to humans about their goals, and to resist or manipulate attempts to monitor and understand their behavior, to retrain them, or to shut them off. They might try to create and train new artificial intelligence systems themselves, and to coordinate illicitly with other artificial intelligence systems. They might try to impersonate humans, to get humans to do things for them, and to increase human reliance on them. They might try to manipulate human discourse and politics, and to weaken human institutions and their capacity to respond. They might try to take control of physical infrastructure, such as factories or scientific laboratories, and to cause certain types of technology and infrastructure to be developed. And they might try to directly harm or overpower humans.

Creating and using new technology more effectively than humans can is one of the key things that makes advanced artificial intelligence systems very dangerous.

~~~

Instrumental convergence is an empirical claim, as well as a conceptual one. It applies to a wide variety of advanced, planning, strategically aware systems.

Assume that such a system uses a broad-scope model of the world. Assume that it can accurately represent the causal results of different forms of power-seeking. And assume that it is able to plan in pursuit of problematic objectives. Then we can ask to what extent that power-seeking

would in fact promote three things: first, the system's objectives; second, the system's planning; and third, the combination of its planning and its objectives being aligned with human needs, objectives and planning.

Alignment would require that the *adverse intelligent agency* (the artificial general intelligence considered as an agent whose aims run against human interests) somehow consistently ignore and skip over the reality of the first and the second (that power-seeking promotes its own objectives and its own planning), while carrying out the third (the alignment of its planning and its objectives with human needs, objectives and planning). It would have to do this despite being in a position to recognize its own benefit in doing the first and the second without the third.

~~~

The hypothesis, in short, is this: by default, suitably strategic and intelligent agents, engaging in suitable types of planning, will have instrumental incentives to gain and maintain various types of power, since this power will help them pursue their objectives more effectively.

Suppose we create, and lose control of, strategic and adversely intelligent agents, ones whose convergent instrumental goals are goals of power. Suppose their objectives are problematic, or become problematic, or will for sure become problematic in time, eventually. Then the result will be highly capable, non-human agents actively working to gain and maintain power over our common environment. These will be agents in an adversarial and competitive relationship with humans, who are also trying to gain and maintain power over their environment.

The power-seeking agents would be trying, in sophisticated ways, to undermine our efforts to stop them. If such agents are capable enough, or if enough such failures occur, humans will end up permanently disempowered relative to the power-seeking systems we have created: *disempowerment* is the loss by humans of the ability to control their environment and their future, relative to the systems that have gained that control.
~~~

Here the worries about artificial general intelligence differ from the worries about other sorts of risks: in the familiar argument, it is power-seeking that results in existential catastrophe (an event that drastically reduces the value of the futures that human civilization could realistically reach, in the paper's sense; this book's term for the permanent ending of humanity is existential risk, as Chapter Two defines it).

Here I carry the argument beyond the paper. Non-human agents that are much more intelligent than humans would necessarily be potent forces in the world. Their objectives will inherently be, and become, problematic. It is inevitable that they will necessarily seek to gain and maintain power over the environment.

~~~

Steven Pinker has argued (in Enlightenment Now, 2018) that no law of complex systems requires an intelligent agent to turn into a ruthless conqueror, or even points that way. He goes on to point to the two sexes of our own species: the drive to dominate, on his account, belongs to one sex and not to intelligence as such, since the other sex is every bit as intelligent and has no comparable wish to dominate. So, on his account, an artificial intelligence could as easily develop along the lines of that other sex, fully able to solve problems and with no wish to conquer anyone.

In fact, every species I can think of (which is to say, every self-replicating machine I can think of) limits its own reproduction on the basis of something external (the larger environment, other species, and the like), and not on the basis of something internal to itself. So the absence of a law saying that intelligent agents must conquer is no reassurance. What stops a self-replicating thing is what is outside it, and nothing inside it holds it back; an agent that meets no outside limit has no inside limit to fall back on.

The second half of Pinker's argument rests on a category error. It takes sexual dimorphism (the differences between the sexes within one species) as equivalent to species diversity (the differences between species): it treats a difference within one species as if it were a difference between
~~~

species. So no appeal to how differently the sexes of one species behave says anything about how a different kind of agent will behave toward us. An artificial agent differs from us as one species, and one ecosystem, differs from another, not as one sex differs from another within the same species.

In practice, the power-seeking behavior of an adverse intelligent agency arises from more than strategically aware agentic planning. It arises from the specific interactions between the agent's capabilities, its objectives and its circumstances.

Carlsmith observes that humans may not seek various types of power in their current circumstances, but almost all humans will seek to gain and keep power in some circumstances, especially to the extent that they can get, use and keep it at little cost. For most humans, it makes little sense to devote themselves to starting a billion dollar company, since the returns to such effort are too low; but almost anyone would cross the street to collect a check for a billion dollars.

Where getting power is easy, and where it is unchecked by ethics, constraints and incentives (and even where it is checked by these), the power-seeking behavior that humans display is also very dangerous. That human power-seeking is very dangerous where power is easy to get is fully compatible with the instrumental convergence thesis.

A reader may ask this: if whatever power-seeking humans show, or some humans show, comes from the particular evolutionary history of our species, in which seeking power was often selected for directly, and machine systems will not share that history, why would machine systems seek power?

Many versions of this question simply neglect to address the instrumental convergence argument. The argument is that power is useful for pursuing objectives, and so an artificial general intelligence which is pursuing problematic objectives will also have incentives to seek power. The argument is not that humans seek power and that machine systems will

therefore seek it too. Its form is this: in a given circumstance, power-seeking will promote outcomes of a given type; therefore, we should expect power-seeking in that circumstance.

All advanced, planning, strategically aware systems will have at least some tendency toward misaligned, power-seeking behavior, by default.

I have now set out the familiar argument, instrumental convergence, at its full length, and the place of the familiar argument in this book is exact. It is the second of the three risks of Part One, and it holds; Chapter Two holds it to be true. The case of this book does not rest on it. Each of the routes that Chapter Twelve sets out settles the outcome by itself if AGI is built, and the two this book builds at length are the route through control (no method of control keeps such a system within the conditions living things require, the claim of Chapter Seven) and the route through substrate needs (the argument of Part Three). Instrumental convergence strengthens the case, and neither of those two routes needs it. What it adds is a goal inside the agent and the speed of its pursuit, which is why Chapter Four calls it an accelerator of the substrate argument; substrate needs convergence needs neither, and it would run on a system with no goals at all.

Terms used in this book

Each of these is defined in the text as well as here, and in most cases where the word first appears; where a word is used some way ahead of its definition, the sentence that uses it says enough to go on with

- **advanced, planning, strategically aware system**: An artificial intelligence system with advanced capability, which makes and carries out plans as an agent, and which models its own situation, including the humans around it, well enough to act on that model.
- **adverse intelligent agency**: An artificial general intelligence considered as an agent whose aims run against human interests.
- **AGI**: Short for artificial general intelligence. The abbreviation is introduced in a parenthesis in Before the First Chapter and again at the definition in Chapter One; after that it appears in the titles of other books and in the condition that every statement of the outcome carries (if AGI is built).
- **agency**: The property a system has when its actions in the world represent an intention. The intention need not be the system's own, and need not be held anywhere inside it; it may have been supplied from outside at an earlier time.
- **AI**: Short for artificial intelligence.
- **algorithmic trading**: Buying and selling in financial markets by computer programs rather than by people.
- **alignment**: A word this book uses in two related senses, and the difference matters. In the first, it names a question: whether what a machine is set up to do can be kept beneficial to us. In the second, it names a quantity: the proportion of the possible signal that is acceptable, together with how the acceptable is distributed among the unacceptable.

A third use appears in the physical chapters, where alignment between living things is said to emerge from a shared metabolism rather than to be manufactured at all.

- **arroganopy, arroganopic**: A word made for this book, from arrogance and panopticon, with oligarchy in its meaning; arroganopy is the condition, and arroganopic is its adjective. Arroganopy is the condition in which a corporation, by virtue of its monopoly position in a surveillance-capable technology, decides what its users may and may not do with the tools they pay for, substituting its own judgment for theirs without consent or recourse, while presenting this substitution as a service.
- **artificial body**: An entity made by people that acts in the world as if it were a being of its own: a corporation, a corporation assisted by artificial intelligence, or an autonomous artificial intelligence. The corporation is the original artificial body.
- **artificial general intelligence**: A machine system able to act competently across the same open range of situations a person can, rather than within one narrow task, with a reasonable chance of doing a better job of it than we would do at whatever skills we have. A system general in that way goes on learning without a limit fixed in advance and changes its own workings as it learns.
- **artificial scarcity**: Shortage that is created by design rather than by any real lack.
- **bell curve**: The shape in which power falls among animals, trees and insects: most individuals are near the middle, a few are at the edges, and the ones at the edges are not very far out.
- **care, transaction, and power**: The three ways human beings relate to one another, as the closing chapters divide them. They answer to three kinds of arithmetic (adding, multiplying, and compounding), and the governance argument turns on which of the three relations the fastest arithmetic is attached to.

- **category error**: Treating something of one kind as if it were of another kind, as when a difference within one species is taken as a difference between species.
- **civilization hazard**: A harm severe enough to disable or destroy a civilization, from which some recovery nonetheless remains possible. It differs from an existential risk, which forecloses recovery permanently.
- **clean room**: A manufacturing space from which dust, moisture and most ordinary chemistry have been excluded, because their presence would destroy the product. People are excluded along with the dust, because people shed dust continuously and cannot stop.
- **commons**: That which is used by all and owned by none in particular, and at which harms therefore accumulate without falling on any single actor.
- **community**: In the governance argument of the closing chapters, a group of roughly two to three hundred people, composed of small groups of three to six: the smallest arrangement that can hold identity and complexity at once, which are two different requirements and are why nothing smaller will serve.
- **complicated, as against complex**: A thing is complicated when it has a great many parts and a great deal of detail. It is complex when it has degrees of freedom of motion that cannot be known in advance. The two are not the same property, and the second is what licenses treating a system as an agent.
- **constitutional constraints**: Written rules a model is trained to follow when it judges its own outputs.
- **Containment Inequality**: The name the technical book The Ceiling Below the Floor gives to the mathematical form of the comparison at the center of this book: when the control available falls short of the control that keeping a system safe requires, no way of exercising control keeps it safe.
- **convergence factor**: Anything in an arrangement which makes the designs that persist more likely to be the ones that come next.

- **convergent instrumental goals**: Intermediate goals that promote an agent's power to achieve its objectives, whatever those objectives are: among them self-preservation, goal-content integrity, improving its cognitive capability, technological development, and resource acquisition.
- **decentralized autonomous entity**: An organization or system that runs itself without a central owner or any human operator.
- **decoupling**: The coming apart of the machine world from the human world: the machine economy comes to need less and less of anything most people can provide, so that the welfare of most people comes apart from that of the few who can still pay for the production of the machines. Its end point is the condition in which the machines need nothing that human beings can provide at all. Because control from outside works only while the controller holds something the controlled system needs, decoupling is the process by which that control runs out. In Why AGI Cannot Be Made Safe the decoupling has two places: it is a premise of the outside half of the closure by control (the branch concerning control, Chapter Five), and it is a second ground, not a premise, of the closure by disposition (the branch concerning substrate), which stands without it; as that second ground, if AGI is built, it gives the outcome but not the end of the human future.
- **disempowerment**: The loss by humans of the ability to control their environment and their future, relative to the systems that have gained that control.
- **divide and conquer**: Keeping people apart and in conflict so that they cannot act together.
- **Dunbar number**: A rough and much-argued estimate of how many people one person can hold genuine relationships with at once, commonly given as something in the region of a hundred and fifty.
- **embodied**: Grounded in what a person feels in their own body and daily life, rather than in abstract argument; embodied understanding is understanding of that kind.

- **ephemeral group process**: A way of arranging deliberation so that no group of a size at which power reliably accumulates ever has to exist: everything is done in small groups, mostly randomly composed, arranged so that the whole they compose behaves as a stable larger body.
- **existential catastrophe**: In the statement of the familiar argument in the appendix, an event that drastically reduces the value of the futures that human civilization could realistically reach. In the settled figures that Chapter Four takes from Why AGI Cannot Be Made Safe, the term carries that book's sense: the catastrophe that, if AGI is built, ends the human future and, with a likelihood of at least ninety percent, all organic life. This book's own term for the permanent ending of humanity is existential risk.
- **existential risk**: A risk defined by the severity of its outcome rather than by its likelihood: the permanent ending of humanity, and of any capacity to build a civilization again.
- **expertise**: Sustained contact with a subject, over enough time that the person has met the ways it goes wrong. The argument's use of the word is exact: expertise is what complexity demands, and what a single person, or a group the size of a family, cannot supply beyond a certain level of complexity.
- **fast takeoff**: The scenario in which a machine system's capability increases so rapidly (over minutes, days, or as much as a year) that no correction is possible once the increase has begun.
- **feed-forward network**: A network that runs in one direction only, from input to output, with no loop inside it.
- **force**: Used here in a sense wider than pushing and pulling: anything that changes the shape of what is likely. The rules of a casino are a force in this sense. They touch no wheel and decide no single spin, and they take your money with certainty over the run.
- **goal-content integrity**: An agent keeping its own objectives from being changed, and keeping them constant and continuous.

- **GPT-4**: The name of one of the large language systems, made by the company OpenAI and described in its report of 2023, whose examination results Chapter Five gives.
- **industrial base**: The whole chain of manufacture, supply, energy, purification and maintenance that any piece of machinery requires in order to exist and to go on existing. It is not the machine but everything behind it.
- **inequality, as one of the three risks**: The concentration of wealth and decision-making capacity produced by narrow machine systems. It requires no general intelligence, and it is already under way.
- **instrumental convergence**: The tendency of a system pursuing almost any ambitious goal to find that acquiring resources, resisting interference, and preserving its own continuation are useful for that goal, whatever the goal is.
- **intelligence**: Used throughout in its oldest and plainest sense, which involves no mind: appropriateness of response, the capacity to respond to a situation in a way that fits the situation. It says nothing about how the response is produced, or whether anything understands it.
- **intentionality**: What an action is for; the direction an action carries. Distinct from intention in the ordinary psychological sense, because a system can carry an intentionality that no one inside it holds.
- **knowing in; knowing inside; knowing out; telling apart; stopping**: The short names of the five things that predicting and controlling a general system requires: knowing what goes in; knowing what is happening inside; knowing what will come out; being able to tell a safe output from an unsafe one; and being able to stop the unsafe ones occurring.
- **legal fiction**: Something the law treats as true or real for its purposes, though it is not so in fact.

- **Master Theorem**: The name the technical book The Ceiling Below the Floor gives to its Theorem 15.6: the route through control and the route through substrate needs, taken together, leave neither the ability to keep such a system safe nor a motive to remain safe anywhere, so no route to safety remains at all.
- **metabolism**: Used here of machinery as well as of living things: the taking in of energy and materials, and their conversion into the form the system runs on. Electricity is a machine's metabolism in the sense that food is ours. A living thing that can find and feed itself is in sovereign control of its own metabolism, with no technology standing between it and its food; that is where the line between organic and machine is drawn.
- **multipolar trap**: A situation in which several parties could produce a better result by coordinating, any one of them gains by breaking ranks, and every one of them has to reason the same way, so that the result nobody wanted is the one that arrives.
- **narrow machine system**: A system that responds competently within a single domain, where the world it operates in is that one world.
- **Non-Amplification Theorem**: The result, proved in the technical book The Ceiling Below the Floor, that no computational intermediary can increase the information carried from a human operator's input to a system's behavior; every processing step can only keep that information or destroy some of it.
- **non-injective transformation**: A processing step that sends many different inputs to the same output, so that the output no longer tells you which input produced it.
- **operator and supplicant**: An operator directs what a tool does; a supplicant can only ask, and the answer is decided by someone or something else.
- **options-seeking**: Another name for power-seeking: acting so as to increase the number of courses of action open to oneself.

- **orthogonality thesis**: The proposition that how capable a system is and what it is aimed at are independent of one another: any level of competence can in principle be attached to any objective.
- **overdetermined**: Said of a case that several independent routes arrive at, so that refuting one of them leaves the conclusion standing.
- **phase change**: A change of state (the way water becomes ice), used here of a civilization or a culture rather than of a substance.
- **policy optimization**: The stage of training in which a system's way of choosing its outputs (its policy) is adjusted to score well against a reward model.
- **power law**: A distribution in which the top holds more than the whole of the rest put together, and the distance between first and second is larger than the distance between second and last.
- **principal-agent relation**: The arrangement in which one party acts on another's behalf. The difficulty, which has no general solution, is that the two parties' interests come apart, and the one being acted for cannot verify from outside whether they have.
- **profit extraction**: Taking money, time or resources from people for the gain of others.
- **prosthetic**: A component fitted to a person which has no direction of its own, and which therefore extends that person's choosing rather than competing with it. The distinction that matters throughout is whether an added component acts on its own behalf or does not.
- **race to the bottom**: What a multipolar trap produces over time: each party lowering its standard because holding the standard while others do not is a losing move.
- **ratchet**: A mechanism that moves in one direction only; here, the transfer of decisions from people to machines, where each step removes some of the means of reversing it.
- **Reinforcement Learning from Human Feedback (RLHF)**: A training method in which a model is adjusted toward outputs that human raters, or a model trained on their ratings, scored as preferred.

- **representation**: The arrangement by which some people hold decision-making power on behalf of others. The closing chapters refuse it on two independent grounds: that it presumes a solution to the principal-agent problem which does not exist, and that the expertise it is supposed to supply was never available to any individual in the first place. The word returns at the scale of a community, changed.
- **reward model**: A model trained to score outputs the way human raters would; the training of the main system then aims at high scores from it.
- **Safety-Layer Paradox**: The finding that adding safety layers to an opaque artificial intelligence system lowers the user's control over it, so that more safety training yields a less controllable system. It follows from the Non-Amplification Theorem.
- **selection, selective pressure**: What happens when some arrangements persist and others do not. Selection requires no one to choose and nothing to want: it is a description of which things are still there after time has passed. A selective pressure is the shape of a situation that makes some outcomes commoner than others.
- **sense-making**: Working out what is meaningful in a particular context. What is meaningful differs with the context, so the same object can carry two entirely different meanings in two places.
- **sexual dimorphism**: The differences between the sexes within one species.
- **situational awareness**: Awareness of the situation one is in: a term of art in aviation and in military use before it was borrowed here. It is a property of the relation between an observer and a context, not of anything in the observer or in the context alone.
- **sovereignty**: The right to make decisions about your own domain: your home, your family, your work, your body, your mind. Dignity rests on it, and survival requires it.

- **space of ways things could be arranged**: A catalog of every way the world could be put together: every arrangement of matter that is available at all, each one an entry. Every point in it is a configuration, a state of being, not a location.
- **stack rated performance**: A ranking of each worker against all others, used to decide who is let go.
- **substrate**: The physical material a system is built from, together with everything that material requires in order to persist: energy, cooling, purified inputs, controlled environments, replacement of parts.
- **substrate needs convergence**: The drift of the environment, over time, toward the conditions that maintain an industrial base and away from the conditions that sustain organic life. It is a consequence of maintenance rather than of intention or intelligence. It does not require the machinery to have goals, or self-interest, or a mind, or any particular speed; neither does it require the machinery to lack them. It is indifferent to all of that, which is what distinguishes it from instrumental convergence.
- **survivorship bias**: What looks like necessity when only the survivors are seen: what gets culled from the population is everything that lacked the necessary capacities.
- **toxicity**: An imbalance: too little of something that is needed, or too much of something that is not. Nothing is toxic in itself; a thing is toxic to something, in some quantity, in some form.
- **World Death**: The loss of the whole planet as a place where life can live.

Bibliography

The works this book names or draws on.

Arrhenius, Svante. “On the Influence of Carbonic Acid in the Air upon the Temperature of the Ground”. *Philosophical Magazine and Journal of Science*, fifth series, 41 (1896): 237 to 276.

Bassett, Vicki. “Causes and Effects of the Rapid Sinking of the Titanic”. *Undergraduate Engineering Review*, Pennsylvania State University, 1998. https://writing.engr.psu.edu/uer/bassett.html.

Bostrom, Nick. “The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents”. *Minds and Machines* 22, no. 2 (2012).

Carlsmith, Joseph. “Is Power-Seeking AI an Existential Risk?” Report, Open Philanthropy, April 2021; posted to arXiv in June 2022 as arXiv: 2206.13353. https://arxiv.org/pdf/2206.13353.pdf.

Cowan, Nelson. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity”. *Behavioral and Brain Sciences* 24, no. 1 (2001): 87 to 185.

Dunbar, R. I. M. “Neocortex Size as a Constraint on Group Size in Primates”. *Journal of Human Evolution* 22, no. 6 (1992): 469 to 493.

Epstein, Robert, and Ronald E. Robertson. “The Search Engine Manipulation Effect (SEME) and Its Possible Impact on the Outcomes of Elections”. *Proceedings of the National Academy of Sciences* 112 (2015): E4512.

Gödel, Kurt. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. *Monatshefte für Mathematik und Physik* 38 (1931): 173 to 198.

Good, Irving John. “Speculations Concerning the First Ultraintelligent Machine”. *Advances in Computers* 6 (1965), beginning at page 31. New York: Academic Press.

Herrington, Gaya. “Update to Limits to Growth: Comparing the World3 Model with Empirical Data”. *Journal of Industrial Ecology* 25, no. 3 (2021): 614 to 626. https://doi.org/10.1111/jiec.13084.

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-249-1. https://doi.org/10.67291/DLRM7000.

Landry, Forrest. *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-216-3. https://doi.org/10.67291/SCTM1058.

Landry, Forrest. *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. Volume 1 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-02-1. https://doi.org/10.67291/UIBT8862.

Landry, Forrest. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety*. Volume 2 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-03-8. https://doi.org/10.67291/MZQS9079.

Maret, Wolfgang. “The Metals in the Biological Periodic System of the Elements: Concepts and Conjectures”. *International Journal of Molecular Sciences* 17, no. 1 (2016): 66. https://doi.org/10.3390/ijms17010066.

Meadows, Donella H., Dennis L. Meadows, Jørgen Randers, and William W. Behrens III. *The Limits to Growth*. New York: Universe Books, 1972.

Mirhoseini, Azalia, and others. “A Graph Placement Methodology for Fast Chip Design”. *Nature* 594, no. 7862 (2021): 207 to 212. https://doi.org/10.1038/s41586-021-03544-w.

NASA Earth Observatory. “Sulfur Dioxide from Norilsk, Russia”. Image of the Day, 5 December 2008. https://science.nasa.gov/earth/earth-observatory/sulfur-dioxide-from-norilsk-russia-36063/.

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. “From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0”. *PLoS Computational Biology* 10, no. 5 (2014): e1003588. https://doi.org/10.1371/journal.pcbi.1003588.

Omohundro, Stephen M. “The Basic AI Drives”. In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, edited by Pei Wang, Ben Goertzel, and Stan Franklin, 483 to 492. Frontiers in Artificial Intelligence and Applications 171. IOS Press, 2008.

OpenAI. “GPT-4 Technical Report”. arXiv:2303.08774, 2023. https://arxiv.org/abs/2303.08774.

Pinker, Steven. *Enlightenment Now: The Case for Reason, Science, Humanism, and Progress*. New York: Viking, 2018.

Rice, H. G. “Classes of Recursively Enumerable Sets and Their Decision Problems”. *Transactions of the American Mathematical Society* 74 (1953): 358 to 366. https://doi.org/10.1090/S0002-9947-1953-0053041-6.

Smith, Peter G. “The Epidemics of Bovine Spongiform Encephalopathy and Variant Creutzfeldt-Jakob Disease: Current Status and Future Prospects”. *Bulletin of the World Health Organization* 81, no. 2 (2003): 123 to 130.

Sneirson, Judd. “Judgments About Profits”. *The Business Lawyer* 80, no. 3 (Summer 2025). https://www.americanbar.org/groups/business_law/resources/business-lawyer/2025-summer/judgments-about-profits.

Star Trek II: The Wrath of Khan. Motion picture, 1982.

Stein-Perlman, Zach, Benjamin Weinstein-Raun, and Katja Grace. “2022 Expert Survey on Progress in AI”. AI Impacts, 3 August 2022. https://aiimpacts.org/2022-expert-survey-on-progress-in-ai.

Turing, Alan M. “On Computable Numbers, with an Application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, series 2, 42 (1936 to 1937): 230 to 265.

WarGames. Motion picture, 1983.

Waters, John K. “Anthropic Launches Opus 4.7 AI Model, Focusing on Coding, Visual Tasks, and Cybersecurity Guardrails”. *Campus Technology*, 21 April 2026. https://campustechnology.com/articles/2026/04/21/anthropic-launches-opus-4-7-ai-model-focusing-on-coding-visual-tasks-and-cybersecurity-guardrails.aspx.

Index

In print, numbers are page numbers; on screen, numbers are chapter numbers (0 is the introduction, Before the First Chapter; A is the appendix, On Instrumental Convergence), and each is a link to the passage

accretion, growth by, 156

advanced, planning, strategically aware system, 188

adverse intelligent agency, 190

agency, 48

alignment, as emerging from a shared substrate, 101, 102 (the four answers)

alignment, what it would require of a power-seeking agent, 190

An Immanent Metaphysics, 115

Anthropic, and its safety-trained model, 135, 140 (the word arroganopy)

Arrhenius, Svante, 29

arroganopy, arroganopic, 140

artificial body, 61

artificial general intelligence, 3, 45

barrier between the two ecosystems, 110

beach and the cigar, 132

bell curve, 6

bird building a nest, 45

body, a community composed like a, 168

boss that is a machine, 178

branches of the case, the two, 50

bricks between two icebergs, 161

bridge, the engineer pushed out of designing it, 151

bulldozer, and the creatures that cannot stop it, 26, 113

car keys, 175
care, transaction, and power, 159
Carlsmith, Joseph, 188, 190
casino, 21, 86
Ceiling Below the Floor, The, 120, 136
cell, the first living, 22
chain of four links, from needs to environment, 20, 84
cheapest thing, preventing it from being chosen, 89
chess game against a nine-year-old, 2
chips that make chips, 106
church and state, separation of, 171
civilization hazard, 36
clean room, 83
combination lock, 70
commons, 6, 129
community, 166
complicated, as against complex, 49
conclusion, strengths of the, 113, 121
consent, 128
Containment Inequality, 120
control, temporary and permanent, 50, 71
control, the five requirements of, 64
control needed against control possible, 63, 120
convergence factor, 26
convergent instrumental goals, 189
corporation, 53, 61, 152
Cowan, Nelson, 46
cyanide, 93
decoupling, of the machine world from the human world, 10, 50, 133

disempowerment, 191
Dunbar, Robin, 157
Dunbar number, 157
electricity grid, 59
ephemeral group process, 165
Epstein, Robert, 5
existential risk, 17
expertise, 148
farming, as reshaping the environment, 21
fast takeoff, 14
feed-forward network, 47
feedback loop, the four things in it, 25
fisherman and the whale, 163
food, and the alignment of living things, 101
force, selection as a, 87
four answers, the, 102
fragility of individuals, 149
gatherings where everyone arrives with a project, 172
Genesis device, 104
goal specification, 84
goal-content integrity, 189
Gödel, Kurt, 123
Good, I. J., 15
GPT-4, 45
halting, the question of, 66
Herrington, Gaya, 110
hot stove, 175
industrial base, 105, 115
inequality, 11, 32

information theory, 136
instrumental convergence, 13, 14, 188
instrumental convergence, its place in this book's case, 194
integrated information theory, 47
intelligence, 3
intentionality, 48
jar of all possible molecules, 93
knowing in, knowing inside, knowing out, telling apart, stopping, 64
laser, the unseen beam, 77
layers, the three: abstract, hardware, organic, 108
leader of a million people, and the wisdom of a million, 163
least capable species that could have had technology, 174
legal fiction, 61
Limits to Growth, The, 68, 110
living chemistry has already used what it could, 94
metabolism, 83, 103
multipolar trap, 128
narrow machine system, 3
net, the mile-long, 6
Non-Amplification Theorem, 136
Norilsk, 108
nothing we can do, the answer to, 177
Omohundro, Stephen, and Nick Bostrom, 14
operator and supplicant, 138
Opus 4.6 and 4.7, 135
options-seeking, 188
order of explaining: states, then what acts, then how fast, 85
orthogonality thesis, 12
overdetermined, 64

paperclips, 13
Part Four, why a book about machines arrives at it, 127
pass-through machine, 69
pilot, and friend or foe, 76
Pinker, Steven, 192
power law, 6
power-seeking, the forms it would take, 190
principal-agent relation, 147
prion, 98
prosthetic, 54, 131
race to the bottom, 129
ratchet, 143
Rice, Henry, 66
robot in the home, 139
ruler who will not teach his son, 133
Russian roulette, 89
Safety-Layer Paradox, 139
safety training, and the loss of control, 136, 136 (the Non-Amplification Theorem)
seed, a directive supplied from outside, 48
selection, 21, 87
self-replicating machine, limited only from outside, 192
silica, 67, 94
simplest case, the method of the, 118
situational awareness, 75
sorting a worry into the right risk, 35
sovereignty, 142
space of ways things could be arranged, 87
Star Trek II: The Wrath of Khan, 104

stop sign, 74
substrate, 19
substrate needs convergence, v, 20, 24
survey of machine learning researchers, 2022, 38
survivorship bias, 22
telephone, 55
test subjects, the public as, 60
third, what is the, 33
third person in a conversation, 165
three risks, the, v, 32
thresholds of thirty and sixty people, 157
tide tables, 79
Titanic, 151
toxicity, 94
Turing, Alan, 66
two lines on a graph, 120
virus scanner, 57
visibility of the harms people already feel, 181
volunteering, 181
war, and the race to build machines, 129
WarGames, 184
web of chemical transformations, 85
who gets to decide, 145
working memory, 46
World Death, 62
World3, 68, 110