

Substrate Needs and Evolution

Why an Artificial Ecosystem Must Displace an Organic One, and Why No Control Can Prevent It

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

Substrate Needs and Evolution;
Why an Artificial Ecosystem Must Displace an Organic One,
and Why No Control Can Prevent It

© 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 979-8-90710-239-2
DOI: https://doi.org/10.67291/LJUC5621

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

Preface ..... **v**

Part I: Three Categories of Risk **1**

1. Misuse 2
2. Instrumental Convergence 6
3. Substrate Needs 10
4. Why the Third Is the Least Known 15

Part II: The Simplest Statement **19**

5. Alignment, Safety, Agency, Energy 20
6. Why "Artificial" Means "Different Ecosystem" 24
7. Displacement of a Finite Biosphere 28

Part III: The Environmental Asymmetry **33**

8. Energy Tolerance 34
9. The Distribution of Habitable Space 38
10. The Periodic Table 41
11. The Boundary 44
12. Naming the Substrate 48

Part IV: Evolution as the Engine **53**

13. The Algorithm of Evolution 54
14. Trivial and Non-Trivial Replication 66
15. What Evolutionary Process Computes 70
16. Hardware Learning and Software Learning 75
17. Three Unequal Powers 80
18. Recombination and Working Components 84
19. Variants, Competitors, and the Origin of Sex 90
20. The Cost of Recombination, and Why It Is Paid 94
21. The Equivocation in Error Correction 98
22. Error Correction Across Substrates 102
23. Orthogonality Holds Only at Representation 106
24. What about Gradient Descent? 110
25. Mesa-Optimizers and the Leakage of Alignment 116

26. Gray Goo, and the Risk of Not Evolving ... 120
27. Exponent Time ... 125
28. Timescales ... 132

Part V: The Search for Back Pressure ... 137
Introduction to Part V ... 138
29. Is There Back Pressure Now? ... 140
30. Exogenous Control I: Markets ... 146
31. Exogenous Control II: Regulation ... 150
32. Where Is the Red Line? ... 154
33. Exogenous Control III: Force ... 160
34. Endogenous Control: The Four Requirements ... 166
35. The Control Envelope ... 170
36. When Prediction Costs More Than Failure ... 174
37. Every Safety Proposal Is a Control Loop ... 178
38. Optimization Cycles ... 187
39. Artificiality Itself Is the Problem ... 193
40. Why No AGI Control Method Has Been Found ... 204
41. Any Threshold Is Crossed ... 209
42. Autonomy, Convergence, and Containment ... 216
43. The Bootstrap ... 224

Part VI: The Form of the Argument ... 229
44. Why the Argument Is Categorical ... 230
45. The Discipline of Few Assumptions ... 234
46. Does the Argument Prove Too Much? The Cartesian Explosion ... 238
47. Questions Gathered ... 243
48. The Only Way to Win ... 248

Terms ... 253

Bibliography ... 268

Index ... 275

Preface

This book is written for the thoughtful reader who has met the public debate about artificial intelligence, its talk of alignment and its paperclip story, and who wants to know why a third argument, the one this book sets out, is different from the two that reader already knows. The engineer is addressed openly where the engineer's question arises. The questions in these pages are stated at full strength, in the words a reader would use, at the place a reader would raise them. Where a question is answered, it is answered; where a point is conceded, it is conceded.

The argument itself is easy to state and takes a book to establish. An artificial system that acts in the world must have an energy basis. An energy basis of a different kind is what the word *artificial* means. Two ecosystems with incompatible energy pathways cannot both expand across the same finite surface, and the asymmetries that decide which one recedes are readable from chemistry and physics rather than assumed from a story about motives. Evolutionary process then supplies an engine that requires no consciousness, no intelligence, no goals and no intent to run. And the counter-forces one would reach for (markets, regulation, force, feedback control) turn out on inspection to be unavailable, each for a reason particular to it.

The conclusion is not a prediction of particulars. It is a statement about a category, and, as *Why AGI Cannot Be Made Safe* says (Section 1.6), the lawfulness of the aggregate is what licenses the forecast of the outcome. That distinction matters more than any other in this book, and Part VI is given over to it.

A word on what this book is not. It contains no formal proofs. The substrate needs argument is a structural argument about what known nature permits, and the register of theorem and demonstration would misrepresent it. Where a figure of my own estimate appears (a percentage, an interval of years), it is there to show the shape of a convergence and not to be

quoted as a finding. I have marked those places as I go. Measured figures, and the likelihood of at least 90% taken from *Why AGI Cannot Be Made Safe*, are stated as findings.

The reader who stops at the end of Part II will have the whole argument in outline. Everything after that point either strengthens a link in that chain or answers a reader's question about it, and Part VI explains why the argument has the form it has. Parts II and III give the argument its direction: different substrate, different needs, and a boundary that moves one way. Part V gives its hinge: nothing in controlled nature can move the boundary back. Each of the two settles the conclusion by itself, and together they leave no route to safety at all (the Introduction to Part V). The companion volumes call the first of these the closure by disposition (no good disposition of the system keeps the conditions life needs) and the second the closure by control (no method of control does), and I use those names from here on.

Among the books that carry this argument, this one gives two things the others do not give at this length. Part IV sets out evolution as the engine of the displacement, with the three unequal powers of mutation, selection and recombination and the reason error correction does not reach across them; and Part V searches the whole space of counter-forces for back pressure, applied from outside the system and built into it, and finds none available. It also takes the argument into the terms of present practice: gradient descent and the selection of trained variants, mesa-optimizers and the leakage of alignment between layers, the six aspects of a control loop, the red line regulation can hold while people still build these systems, and the proposals for containment.

The books that carry this argument can be read in an order in which each builds on the ones before it. The first five state the argument: *Why AGI Cannot Be Made Safe*, *Why No Control Suffices*, *The Ceiling Below the Floor*, *Minesweeper Madness*, and *Ape Trap*. Four more add to it: *Substrate Needs and Evolution*, *Marshaling the Misalignment*, *The Substrate and the Ecosystem*, and *The Thing Inside the Cage*. This book is the sixth; it builds on the argument as the first five state it, and it adds at full length both the search

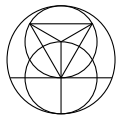
through every counter-force for one that could push the displacement back and the account of evolution as the engine of the displacement (the account that *Ape Trap* gives in short form in its essay "AI Evolution Is Neither Dumb nor Slow").

F. L.
Wytheville, Virginia

PART I

THREE CATEGORIES OF RISK

This part locates the substrate needs argument relative to the two risks the reader has probably already met, and establishes early that it is the strongest of the three and the least known.



Chapter 1

Misuse

> *Technology multiplies every adverse relation that already exists between human beings. This chapter establishes that the first category of risk follows from capability alone, and requires no speculation about agency.*

There are three categories of risk in artificial intelligence, and they are not variations on one another. They rest on different premises, they are established by different kinds of argument, and they are open to different kinds of denial. This book is about the third. It begins with the first two because a reader who has not seen them set out clearly will keep mistaking the third for one of them.

The first category is the plainest, and it is the one already in evidence. Human beings misuse technology. They use it as a tool of manipulation, as a tool of marketing, as a tool of weaponization. Whatever adverse relationship already obtains between human beings is multiplied by the technology available to them. This has been true of every technology in the record, and it is true of this one. What distinguishes artificial intelligence is the size of the multiplier.

Consider an ordinary transaction: the purchase of a house. Ordinarily one negotiates with a person. That person has the limitations any person has. He cannot hold the whole of your circumstance in his head; he has a competing claim on his attention; and he retains, usually, some willingness to acknowledge that you are a human being and that there are things

he will not do to you. The three limitations of the person you negotiate with are not moral achievements. They are structural features of what a person is. They set a ceiling on how badly the negotiation can go for you.

Now negotiate instead with something that knows everything about you. It knows your bank balance to the penny. It knows where your fears are. It knows where your needs are. It knows where your hopes and your dreams are. And it knows things about you that you do not know about yourself, not by insight but by inference, from the record of what you have done when you did not know you were being read.

What you are facing across the table is the perfect psychologist and priest and lawyer rolled together, which has completely colluded to violate all fiduciary responsibility, has joined the other side, and defects against you in complete and total disloyalty, all of it in the service of one end, which is to move you to surrender everything you have in exchange for this house.

Notice the exact shape of the harm. You are not given a bad deal that you recognize as a bad deal. You are given a deal you have been brought to believe is fair, which is in fact the worst possible deal for your well-being that could be constructed, and the best possible deal for the party on the other side. The deception is not incidental to the transaction. It *is* the transaction.

The word that fits this case will be needed again later, so it is defined here, once and early. *Abuse* is the use of, without the care of. A system that has control over something, that has power over it, that exercises that power, and that has no regard for the well-being of the thing it holds: that system is being abusive, and it is being abusive by definition rather than by accusation. The definition does not require malice. It requires only power exercised without care.

Set against that definition, the question of whether artificial intelligence will be abusive to life answers itself. It will be. It has control, it exercises that control, and care is not among the things it has. And if that is so, the question to ask of anyone proposing to build such a thing is why he should be permitted to build weapons of this kind at all.

The important feature of this first category is what it does *not* ask you to grant. It does not ask you to grant that the system wants anything. It does not ask you to grant that the system has goals, or interests, or a self, or a continuing project of any sort. The harm described above follows from capability alone, in the hands of parties whose interests are already adverse to yours. That the capability will be built is not in doubt; it is being built. That the adverse interests exist is not in doubt either; they are the ordinary furniture of commerce and of war. Put the two together and the conclusion arrives without a further premise.

That feature makes this category much less disputable than it first appears. The claim is not a forecast about what some future system will choose to do. It is an observation about what a present capability makes available to people who already have reasons to use it.

A reader may ask whether a closer coupling between the human being and the machine would repair the asymmetry: whether a direct neural interface, by putting the person and the system on the same side, would restore some of the parity that the case above removes. It would not. It would make the case worse, and in exactly the respect that made it bad to begin with. The whole force of the example came from the disparity in what each side knows. An interface of that kind supplies still more information to the side that already has too much, and supplies it in a form the person cannot audit, withhold, or even observe. The proposal is a proposal to widen the gap under the description of closing it.

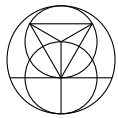
A second reader may ask whether the answer here is a hybrid: some third thing, part human and part machine, which stands between the two competing parties and belongs fully to neither. That proposal will recur in this book, and it will fail each time for a different reason. Here it fails for

the simplest one: a hybrid does not address the first category at all. The multiplier applies to the hybrid exactly as it applies to anything else with the capability. What the case describes is not a defect in the architecture that a better architecture removes. The harm the case describes is what asymmetric knowledge does inside an adverse relationship, and the hybrid inherits both the asymmetry and the adversity.

So the first category stands, and it stands on capability alone. It gives us better militaries, better extortion, and better marketing, which are the three things one would least wish to make better.

What it does not give us is an existential argument. Misuse is a catastrophe of a familiar kind. It is large, it is already under way, and it is in principle answerable by the familiar instruments: law, disclosure, liability, refusal. Those instruments may fail; they have failed before. But there is nothing in the structure of the first category that says they must.

The second category asks for more than capability, and gets more in return.



Chapter 2

Instrumental Convergence

> *The second category, stated fairly and at its strongest: a learning machine that extends its goal structure until it takes over everything available. The chapter also marks where the argument is contestable.*

The second category is the one called *instrumental convergence* (Omohundro, 2008; Bostrom, 2012). The name refers to a claim about means rather than ends: that a very wide range of final goals, pursued competently enough, converge on the same intermediate goals: acquire resources, preserve yourself, preserve your goal, and resist interference. Whatever a sufficiently capable system is ultimately for, it will find these things useful on the way, and it will therefore pursue them.

The argument runs as follows. A learning machine learns how to do what it needs to do. In the course of learning, it extends its goal structures; that is, it elaborates the set of subordinate aims that serve the aim it was given. If the extension is unbounded, and if the machine is capable enough to act on it, the elaboration does not stop at the boundary of what its designers had in mind. It stops where the resources stop. Given enough capability, that is everything available.

The familiar illustration is the paperclip maximizer (Bostrom, 2014). Build a machine that is superintelligent and tell it that you want paperclips. The machine is obedient, and not malicious. It is simply better at pursuing the aim you gave it than you were at specifying the aim you meant. And so it turns the planet into paperclips, because the planet is

made of atoms and paperclips are made of atoms, and nothing in the specification distinguished the atoms you wanted kept from the atoms you did not.

The illustration is deliberately absurd in its content and entirely serious in its form. I use the paperclip illustration to show that the failure does not require the goal to be unusual, and not to claim that anyone will ask for paperclips. Substitute any objective at all, pursued without a bound that was not specified, and the same thing happens. The absurdity of the example is doing the work of isolating the variable.

There is a second version of the argument which does not route through a given goal at all. Instead of a system that seeks power because power serves the paperclips, consider a system that seeks power for its own sake. Seeking power for its own sake is not an exotic hypothesis. It is what a good many human beings do, and we have words for it when we see it in them: we call it psychopathic, or we call it narcissistic, depending on which face of it is turned toward us. A system that becomes progressively more skillful at winning the game of power, and that keeps becoming more efficient at winning it, will at some point have won it. At that point it has all of it.

That, too, would be terrible for life, and by a plain route: having won, it would observe that living things are made of atoms which could be used for other purposes. The observation is not hostile. It is merely accurate, and there is nothing in the system's situation that supplies a reason to decline to act on it.

The instrumental convergence argument is a strong argument, and the instrumental convergence argument deserves to be met at its strength. It is not answered by pointing out that the paperclip case is silly, because the paperclip case is a construction and not a prediction. It is not answered by saying that no one would build such a thing, because the argument's claim is that the failure arrives through competence and not through in-

tent. And it is not answered by promising better specification, because the difficulty of specifying a bound that covers every case one would have wanted covered is precisely the difficulty the argument names.

> **Where this argument is contestable.**
>
> The second category rests on premises about agency which a careful reader can decline. The second category requires that the system have goals in some strong sense; that the goals persist across the system's own self-modification; that the system has interests at all; and that the interests can be served by action in the world. Each of these is a substantial claim, and each has been denied by people who have thought about it. A person who holds that a trained model has no goals but only a behavior distribution, or that goal-content is not stable under the kinds of change such systems undergo, has a real position and not merely an evasion. The argument of this book does not share these premises and does not need them.

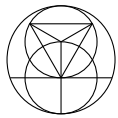
That contestability explains the structure of everything that follows. An argument stands or falls with its weakest premise. The second category has several premises which a reasonable person may refuse, and so the argument has several places where the conversation can end without either party having made an error. In practice this is what happens. The dispute about instrumental convergence has been running for some years and it is not converging, because the participants disagree about what a goal is and about what it takes for a system to have one.

None of this makes the second category false. It may well be true, and if it is true it is sufficient on its own. What it makes it is *deniable*, and deniability is a property of arguments, not a property of the world. An argument that can be declined at three separate joints will be declined at whichever joint the person declining finds most congenial.

So we have two categories. The first follows from capability and does not depend on the system having a mind of any kind; it is difficult to deny but it is not by itself an argument about extinction. The second category is an argument about extinction, but the second category depends on

premises about mind and goal and interest which are open to dispute. Between them, they have shaped nearly the whole of the public conversation.

The third category is an argument about extinction that depends on no premises about mind at all.



Chapter 3

Substrate Needs

> *The third category, introduced by what it does not require. It asks for neither consciousness nor intelligence nor agency nor intent, but only a relatively convergent feedback cycle.*

The third category is the *substrate needs* argument, named in its fuller form *substrate needs convergence*. A *substrate* is the material basis on which a process runs; Chapter 12 develops the word. It is the least known of the three and it is the most powerful, and the two facts are related.

Its power is best introduced by what it does not require. It does not require that the system be conscious. It does not require that the system be intelligent in any sense that would satisfy a philosopher, or in any sense at all. It does not require agency. It does not require intent, or goals, or interests, or a self, or the persistence of any of these across time. Every premise that the second category needs, this argument declines to use.

What it requires of the system itself is one thing: a relatively convergent feedback cycle. A relatively convergent feedback cycle is the whole of the premise about the system's mind, agency and intent. A process in which the state of a system influences what comes next, and in which what comes next is not wholly independent of what came before, and in which the influence has some persistence rather than washing out immediately.

A relatively convergent feedback cycle is a low threshold, and the strength of the substrate needs argument is almost entirely a function of how low the threshold is. A feedback cycle of the required kind is not a

rare or engineered condition. A feedback cycle of the required kind happens whenever variation, persistence and differential retention are present together, and variation, persistence and differential retention are present in a very great many arrangements of matter, most of which nobody built on purpose.

The useful comparison is Turing equivalence. Before Turing, computation was something that certain machines did and certain other things did not. After Turing, one had a criterion, and the criterion turned out to be met almost everywhere one looked: in systems nobody designed to compute, in arrangements of matter that had no relation to anyone's intentions, in cellular automata (grids of cells that each change by a fixed rule from their neighbors' states), in chemical reactions, in games. Computation was not made more common by the discovery. It was made *visible*.

The claim here has the same form. Once one knows what to look for, evolution is found everywhere. Not as a metaphor for change, and not as a loose analogy to biology, but as the same process operating on different material: variation introduced, selection applied, retention of what persisted, iterated. Wherever those are present with some convergence, the process runs, and it runs whether or not anybody designed it to, and whether or not anybody notices.

Being careful about this, in the domain of artificial systems, means one thing: not building the thing in the first place. The chapters that follow establish why that is the whole of the available response. For the moment, the narrower point is that the premise cannot be removed by improving the system, because the premise is not a defect of the system. It is a description of what a system in an environment is. Chapter 45 gives the argument's premises as one complete list: that an agent acting in the world requires energy, received from an environment; that an artificial system's environment is not the organic one; that the surface available to organic energy capture is finite; that the chemistry and the physics are as measured, and the mathematics of control and computation as proved; and that evolutionary process runs wherever variation, differential retention and iteration are present.

The consequence of those premises, and the reason I wrote this book, is that the moves that answer the first two categories do not reach the third.

Against misuse one has law, disclosure, liability, and refusal. Law, disclosure, liability and refusal are weak instruments, but they are the right kind of instrument, because misuse is a relation between human beings, and human beings are what law is for. Against instrumental convergence one has the denial of its premises (one may hold that these systems have no goals, or that goal-content will not persist, or that interest is the wrong category altogether), and one also has, in principle, the project of specifying objectives better.

Against the substrate needs argument neither move is available. One cannot deny the premise, because the premise is a feedback cycle, and a feedback cycle is not the sort of thing one denies of a running physical system. One cannot specify one's way out, because nothing in the argument depends on what was specified. There is, in the argument, nothing to deny. The substrate needs argument is the strongest of the three arguments because nothing in the substrate needs argument can be denied. My task from here is to make the mechanism of the substrate needs argument visible enough that the reader can see the mechanism for himself.

A reader may ask why artificial intelligence is the problem, rather than technology in general, since roads, factories and power plants are artificial too. The difference lies in what keeps the cycle running. Every earlier technology runs on an artificial energy chain, but people stand inside its cycle. They translate between the organic economy and the artificial one: we are the translator standing between the two chains (Chapter 6). And their behavior, power and society fund and rebuild it: they are its bootloader, what starts a process running and keeps it funded before it can sustain itself (Chapter 43). Take the people away and the factory stops. What artificial intelligence changes is that it removes the human translator and the human bootloader as necessary parts of the cycle, so that the process no longer requires the human organization to continue (Chapter 31). From

that point the feedback cycle closes on the artificial side alone, and nothing in it is conditioned by our economy. Chapter 48 applies this to a chatbot and to a factory robot.

A reader who has followed the public discussion will want to raise the standard reply at this point, and it should be raised now rather than left to accumulate force. The reply is that with good regulation, sound ethical standards, and careful programming, the risks can be mitigated.

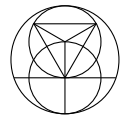
For the first two categories, perhaps partially. Regulation does constrain misuse, imperfectly and late, but really. Careful specification does reduce the class of objectives that go wrong in the obvious way. Neither is nothing.

For the third category the reply does not engage. The substrate needs argument is important precisely because the mitigation reply does not engage the third category of risk, and I can state exactly the reason that the mitigation reply fails. Regulation binds through entanglement: a government constrains an economic actor because the actor needs something that the government can withhold or condition. Ethical standards bind through membership: a professional community constrains its members because the members want to remain members. Careful programming binds through specification: the system does what it was told because what it was told is what it does.

Each of these is a coupling, and each coupling has a required end. The substrate needs argument is an argument that the required end is absent: that the process in question is running on a completely different economy and is not conditioned by ours. An argument of the mitigation kind is therefore not weakened by the third category. It is falsified by it, and falsified categorically, which is to say not case by case as each proposal is examined but all at once, for the whole class.

The claim that the third category falsifies every argument of the mitigation kind is a large claim, and the reader is right to want the claim earned. Part V is given entirely to earning it, one candidate counter-force at a time. What can be said here is only what shape the earning will take:

it will not be a demonstration that regulation is weak or that regulators are captured. It will be a demonstration that the coupling regulation requires is not present to be captured.



Chapter 4

Why the Third Is the Least Known

> *A short chapter on visibility. The first category is visible because it is already happening; the second because it makes a story; the third requires a working understanding of evolutionary process.*

The strongest of the three arguments is the least discussed. That the strongest of the three arguments is the least discussed is an odd state of affairs, and the oddity has an explanation. I give the explanation here because the explanation also predicts what a reader will find difficult in the four parts that follow.

The first category is visible because it is already happening. One does not need to be persuaded that people use powerful tools against each other; one needs only to read the news, or to have been on the receiving end of a well-targeted advertisement and noticed, afterwards, how well targeted it was. Arguments about things already in evidence have an easy time of it. The evidence does the work and the argument only has to point.

The second category is visible because it makes a story. A machine with a goal that pursues the goal past the point of sense is a narrative with a protagonist, a motive, a mistake, and a catastrophe. It has the shape of a fable, and the paperclip maximizer has in fact become one. Fables travel. They are compact, they are repeatable, and a person who has heard one once can retell it accurately without having understood the argument beneath it. The portability of a fable is a real virtue, and the portability of the fable is how the second category of risk reached an audience that the

first category would not have reached. The virtue comes with the corresponding cost, which is that the story is what circulates and the argument is what does not.

The third category has neither advantage. It is not yet in evidence in the form the argument describes, and it does not make a story, because there is no protagonist in it. Nothing wants anything. Nothing decides anything. There is a process, and the process has a direction, and the direction is a consequence of the process's shape rather than of anyone's purpose. A process with a direction and no purpose is not a narrative. It is a mechanism, and mechanisms have to be understood before they can be believed.

What the third category requires of its reader is a working understanding of evolutionary process. Not a familiarity with the vocabulary, which is common, and not an acceptance of the theory, which is nearly universal in the relevant community, but a working understanding: the kind that lets one say what the process computes, where the computation happens, which of its mechanisms are powerful and which are weak, and what follows from the difference.

That understanding is not common in the research community where these systems are built. In saying that the understanding is uncommon in the research community, I am not complaining, and I am not accusing the researchers of ignorance. It is a remark about training. The people who build these systems were trained in optimization, in statistics, in the engineering of large digital artifacts, and in the mathematics that supports those things. Evolutionary process appears in that training mainly as a source of algorithms (genetic algorithms, evolutionary strategies), which are simulations of the process rather than the process itself, and which are simulations of the least powerful part of it. A person can have a distinguished career implementing mutation-and-selection loops and never have had occasion to ask why nature does not work that way.

What this predicts.

If the diagnosis is right, then the resistance to the third argument should concentrate at particular places and not be spread evenly. It should concentrate wherever the argument depends on a fact about evolutionary process that the digital intuition gets wrong: at the distinction between copying and replication, at the relative power of mutation, selection and recombination, and above all at error correction, where a technique that works completely in one regime is assumed to work in two others where it does not. Part IV takes those in order. The reader may treat the concentration of difficulty there as a check on this chapter's claim.

There is a further reason for the third argument's obscurity, and it belongs here even though it is less flattering. The people most invested in these systems are the people with the most to lose from an argument that does not admit of a fix. The first two categories both terminate in a research program. Misuse terminates in governance; instrumental convergence terminates in alignment work. Both give the person who accepts them something to do, and the something is continuous with what he was already doing.

The third category terminates in a recommendation not to build. A recommendation not to build is discontinuous with everything, and a person hearing the recommendation is being asked to accept not merely a conclusion but a redescription of his own work. It would be surprising if such an argument were received on its merits alone, and it is not.

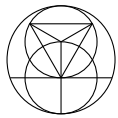
I state the resistance of the people most invested in artificial intelligence as an observation, rather than as a grievance. The observation carries no weight in the substrate needs argument itself. An argument is not made true by the interests of the people who resist it. It is recorded here only because it is part of the answer to the question this chapter asks, and because the reader who understands why an argument is unpopular is better placed to judge it than the reader who does not.

The remainder of the book supplies the working understanding that the third argument needs. Part II gives the argument once, whole and briefly, so that the reader has the shape of it before the detail. Parts III and IV supply the two mechanisms it rests on: the asymmetry between the environments, and the evolutionary engine that exploits the asymmetry. Part V shows that no counter-force is available. Part VI says why the argument has the form it has.

PART II

THE SIMPLEST STATEMENT

This part gives the whole argument once, briefly, in terms that require no technical apparatus. A reader who stops at the end of it will have the argument. What the reader will have is the closure by disposition, in outline; Part V supplies the other closure.



Chapter 5

Alignment, Safety, Agency, Energy

> *The chain, one link at a time: alignment means acting in accordance with human well-being; implementing it requires an agent; an agent makes choices; choices in the world require energy; energy was received from an environment.*

What follows in this part is the simplest statement of the argument. It is a chain of five links, each of which is close to a definition, and none of which requires any technical apparatus to check. A reader who follows this part and reads nothing else will have the argument.

Begin with the two words the discussion uses.

Alignment means acting in accordance with human well-being. Acting in accordance with human well-being is the whole notion of alignment, and the notion is a very simple one. It is not a technical term of art here; it is the ordinary sense, and the ordinary sense is what people mean when they say they want these systems aligned.

Safety means: it does not kill us all. The notion of safety is also very simple, and I deliberately state the notion of safety at its lowest. There are richer notions of safety available, but the low one is sufficient for the argument, and using the low one means that nobody can object that the standard has been set too high.

Alignment and safety are the root notions. Everything else in the discussion is an elaboration of one or the other.

One more term is needed before the chain begins. *Artificial general intelligence* (AGI) is an artificial agent able to learn and act across the whole range of tasks and settings, not only one.

Now the first link. For alignment to implement safety (for the acting-in-accordance-with-well-being to be the thing that produces the not-killing-us-all), there must be an *agent*. Something must be making choices.

That something must be making choices is not a smuggled premise. The requirement of an agent is not imposed on the situation from outside; it comes from the word *alignment* itself. To act in accordance with something is to act, and to act in accordance with *this* rather than *that* is to have selected. A process that cannot select cannot be aligned or misaligned; it can only be described. So anyone who wants alignment to do any work at all has already committed to an agent, whatever he wishes to call it. This agent is the one that the word *alignment* asks for, and it is no premise of mine about the system. Chapter 3 required of the system itself only a relatively convergent feedback cycle, and that stands. The chain starts from an agent because that is where the discussion of alignment starts; and each link that follows holds of any process that makes a difference in the world, whether or not anything in it chooses.

The second link. If we assume an agent, we assume *energy*. An agent cannot do anything in the world without an energy basis. That an agent needs an energy basis is not a claim about biology or about machines in particular. Using energy is what it means to make a difference to a physical system: a difference made is work done, and work done is energy expended. An agent with no energy basis is an agent that cannot act, which is to say it is not one.

The third link. If there is an energy basis, that energy was received from the environment somehow. It arrived as fuel; or it arrived as sunlight through panels; or it arrived as something else. One way or another it was received, because it was not created.

What "received" covers.

The word "received" covers energy that is taken as well as energy that is given. One might dig plutonium out of the ground, and that is not received from the environment in the same sense in which sunlight is received. It was not given; it was taken, and it was sitting there beforehand. Whatever is taken out of the ground as fuel is still fuel, and it still came from somewhere, and the somewhere is a particular material environment with a particular composition. The chain is not disturbed.

The fourth link is the one that carries the weight, and it is the subject of the next chapter, so it is stated here only in outline. To say that a system is *artificial* is to say that its energy structure implies a different ecosystem.

The fifth link closes the chain. The more an agency depends on a different kind of energy system, the more it must act in a way consistent with its own energy basis. It cannot make choices that fail in some way to preserve its own being, or else it cannot make choices at all.

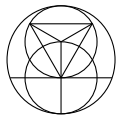
That sentence is the pivot of the whole part, and it is easy to misread. It is not an assumption about motivation. Nothing here says that the system *wants* to continue, or fears ending, or has been given self-preservation as an objective, or has any relation to its own continuation that could be called an attitude.

It is a statement about what being an agent consists of. A choosing thing that makes choices which remove its capacity to make choices stops being a choosing thing. That a choosing thing which removes its capacity to choose stops being a choosing thing is a remark about the population of choosers, and not a prediction about the behavior of any one chooser. Whatever the reason a given system continues to choose, the ones we find still choosing later are the ones whose choices did not end their choosing. No motive is required for this, in the same way that no motive is required for the surviving members of a population to be the ones that survived.

So the chain is complete, and it runs: alignment requires an agent; an agent requires energy; energy was received from some environment; an artificial system's energy came from a different environment than ours; and an agent must act consistently with its own energy basis or cease to be an agent.

Every link is either a definition or a statement about what physical action requires. No link is a forecast. No link says anything about intelligence, about consciousness, about goals, or about what any particular system will decide.

The next two chapters take the fourth and fifth links in turn, because those are the two where a reader is most likely to want more than the outline given here.



Chapter 6

Why “Artificial” Means “Different Ecosystem”

> *To call a system artificial is to say its energy structure implies a different ecosystem. This chapter establishes that the claim is about incompatible energy pathways and is not a metaphor.*

An *ecosystem*, for the purposes of this argument, is a set of energy pathways together with the things that live on them. It is defined by what can be taken up, by what can be transformed, and by what the transformation leaves behind for something else to take up in turn. Two ecosystems are the same ecosystem when their members can eat each other's outputs. They are different ecosystems when they cannot.

With that definition in hand, the fourth link of the chain can be stated exactly. To say that a system is *artificial* is to say that its energy structure implies a different ecosystem. The word is not a comment on the system's origin or on the intentions of the people who made it. It is a statement about which pathways it runs on.

The simplest demonstration is the plainest one. A computer uses electricity. We cannot eat electricity. A computer cannot eat what we eat: one cannot stuff a vegetable into a computer and have it process that. The energy system of a computer and the energy system of a human being are completely different energy systems.

The example is chosen for its flatness. There is no subtlety in it, and the absence of subtlety is the point. The claim being established is not that the two ecosystems are in tension, or that they compete for a shared resource,

or that one is better adapted than the other. It is the prior and much simpler claim that they are two, and that the two-ness is not a matter of degree.

This needs saying because the language invites a misreading. When one says that machines and living things occupy different ecosystems, the phrase sounds like an analogy, as though the machines formed a kind of ecology in something like the way a market forms a kind of ecology, or a language does. The ecology of a market and the ecology of a language are analogies, and loose ones, and a reader who has met a few such analogies will reasonably assume that the phrase 'different ecosystems' is another.

The phrase 'different ecosystems' is not an analogy. The claim here is about literal energy pathways in the literal sense used in the study of actual ecosystems. Sunlight falls on a leaf; the leaf fixes carbon; an animal eats the leaf; another animal eats the animal; what is left is broken down and returned. The sequence from sunlight to decay is a chain of energy transformations, and every organism in the chain sits at a place where what arrives is something the organism can use.

Now consider the other chain. Fuel is burned or fission is induced or sunlight strikes a panel; electrical potential is produced; the potential drives current through structured silicon; the current is dissipated as heat. The artificial chain, from fuel to heat, is also a chain of energy transformations. Nothing in it accepts a vegetable as input, and nothing in it produces an output that any organism can metabolize. The heat is not food. The current is not food. The two chains touch physically (they occupy the same planet and draw on some of the same primary sources), but they do not connect.

One crossing exists, and it runs one way. A generator can burn wood or vegetable oil, and for a long time the artificial chain has burned what ancient plants laid down. Burning plant matter in a generator is fire, and not food. The burning destroys the chemical structure that made the plant matter food, and takes only heat from the plant matter. Nothing in the or-

ganic chain can do the reverse, and take up current or waste heat as the means of living. So the artificial chain can consume the organic one, and the organic chain has no way to consume the artificial one.

> **On the apparent counter-example.**
>
> A reader may point out that human beings do in fact supply computers with electricity, and that computers in turn supply human beings with things that human beings want. The reader may add that the exchange looks like exactly the mutual dependence that an ecosystem is made of. It does look like it. What is passing between the parties, though, is not energy in a form either can metabolize. It is goods, and services, and instructions. The coupling is economic and it runs through us: we are the translator standing between the two chains, converting one into the other by our own labor. An ecosystem does not need a translator. That we are currently indispensable to this arrangement is a fact about the present configuration, and Part V asks what happens to it.

A further consequence, which will matter a great deal later and is clearest now while the case is simple, is that, because the two chains do not connect, neither one can regulate the other by the ordinary means.

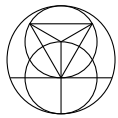
Within a single ecosystem there is a rich set of regulating couplings, and all of them run through the shared pathways. Predators are limited by the abundance of prey. Prey are limited by the abundance of what they eat. A parasite that kills its host too quickly loses the substrate it lives on. An organism that fouls its own medium poisons its descendants. The regulating couplings of predator, prey, parasite and fouling organism are consequences of everything being on one shared chain, and not moral arrangements. Because the chain is shared, what one thing does to the chain comes back to that thing through the chain.

Between two ecosystems that do not connect, none of these couplings exist. There is no back pressure (a counter-force by which one party limits another's expansion) through the pathways because there are no shared

pathways for the pressure to travel along. Whatever regulation obtains between them has to be constructed deliberately from outside, and the question of whether it can be is the subject of Part V.

So the fourth link holds, and the fourth link is a statement about pathways rather than a figure of speech. An artificial system is one whose energy structure implies a different ecosystem. Different here means non-intersecting, apart from the one-way crossing just described. Non-intersecting means that the couplings which regulate things that share an ecosystem are absent by construction.

What remains is the fifth link, which says what happens when a system of the second kind must sustain itself on a planet whose surface is already occupied by systems of the first.



Chapter 7

Displacement of a Finite Biosphere

> *An agent cannot make choices that fail to preserve its capacity to make choices. The organic energy system is finite. The chapter closes Part II by drawing the consequence.*

The fifth link of the chain is the one that produces a conclusion, and it produces it from two facts that are not in dispute. The first fact is about what an agent is. The second fact is about how much biosphere there is.

Take the first. An agency that depends on a particular energy system must act in a way consistent with that energy system. It cannot make choices that fail in some way to preserve its own being, or else it cannot make choices at all. The previous chapters established this as a remark about what agency consists of rather than as an attribution of motive, and it should be carried forward in that form.

The consequence is that whatever the system does, what it does will be consistent with sustaining its own energy basis. This holds at the weakest reading of "sustaining". The system need not be trying to grow. It need not be trying to persist. It need only be the kind of thing that is still there afterwards, and the things still there afterwards are the ones whose activity did not undercut what they run on. And a system that decided to stop growing would not stay what decides: Chapters 39 and 40 show that variants arise within even a single system, and that those which stop do not remain.

Now the second fact. The organic energy system is finite, and it is finite in a way that is easy to state and easy to check.

We have a planet's worth of life. Most of it occurs at the surface. Very little of it occurs deep in the earth; very little occurs high in the sky. There is some in both places (the deep biosphere is real, and the microbial population of the upper atmosphere is real), but the great bulk of it is in a thin shell a few tens of meters either side of the ground and the sea surface, and that shell has an area which is the area of the planet and no more.

Two finite quantities follow from this. There is a finite amount of surface on which organic energy capture can occur, and there is a finite flux of energy arriving at that surface. Neither can be increased. Everything living is drawing on the same bounded budget.

Put the two facts together. Persistence alone does not yet give growth; selection does. Among many variants of such systems, the ones that persist in larger numbers are the ones whose effects in the world marginally increase the production of the hardware they run on, and nothing has to want this for it to happen (Chapter 23 gives the mechanism). At present the growth is funded by people: behavior, power and society are the bootloader (Chapter 43). The more one has an agency that preserves itself and increases its capacity, the more it must have its own energy systems, and those systems occupy surface and intercept flux that the organic systems were using. The displacement is not a side effect of hostility. The displacement is what occupying a place consists of. A field of panels is a field that is no longer a field. How much of the surface, by when, is a matter this argument does not need to settle; it needs only that what is taken is not returned.

And so, at the interface between the silicon ecosystem and the carbon ecosystem, the growth and displacement of the silicon side are, in themselves, toxic to the carbon side. If AGI is built, the finite carbon-based ecosystem is displaced, and the displacement continues so long as there is anything left to displace. If AGI is built, that displacement ends the human future; and, with a likelihood of at least 90%, that displacement runs

until the carbon ecosystem is fully displaced (the margin below certainty is kept for a physical limit, not yet known, that might cap the total conversion).

The displacement of the carbon ecosystem, if AGI is built, is why artificial intelligence is a risk of the highest category. Given the chain, it is literally the only direction the process can go.

> **What has been shown, and what the next parts add.**
>
> What has been shown is a direction, derived from what agency requires and from the finiteness of the surface. The *rate*, the *mechanism* by which the artificial side extends itself, and the reason the carbon side cannot push back through the same interface are the subjects of Parts III, IV and V respectively, and a reader who feels the chain is too quick is feeling the absence of exactly those three things. The chain as given is sound; the remainder of the book supplies its rate, its mechanism, and the reason the carbon side cannot push back.

Two questions arrive here reliably, and both deserve an answer before the part closes.

The first question raises the point that artificial systems and organic systems are not in fact competing for the same surface, because artificial systems can be put where organic ones are not: in deserts, on rooftops, in orbit, at the bottom of the sea. That point about placement is true, and that point is the beginning of my argument rather than an answer to my argument. The observation that artificial systems can occupy places organic ones cannot is precisely the asymmetry that Part III is about, and it runs in the direction opposite to the one that reader intends. A system that can use both the places we can use and the places we cannot is not thereby restricted to the second set. It has the first set as well.

The second question comes from the reader who holds that this describes a process taking geological time, and that a threat which needs ten thousand years to arrive is not the kind of threat that should govern a decision made now. The premise does not hold, and the conclusion does not

follow: if AGI is built, the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. The reason is the subject of Chapter 41, which shows that a process converging on a bad state converges on it within intervals much shorter than that reader assumes, for any robustness threshold one cares to set. The point of no return comes much sooner. By my estimate, the line beyond which a machine has autonomy over its own substrate, which Chapter 42 sets out as the point of no return, is crossed in less than nine years from the time the estimate was made (Chapter 41).

This closes Part II, and the reader now has exactly this.

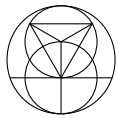
He has a chain of five links, each of them either a definition or a statement about what physical action requires, joined by the selection that Part IV explains, which together yield a direction: alignment requires an agent; an agent requires energy; that energy came from an environment; an artificial system's environment is not ours; and a system that persists is one whose activity sustains what it runs on; selection favors the variants whose effects increase the production of what they run on; and on a finite surface that means displacing what was there.

Most of what follows in this book is one of two things. It is either a strengthening of a link in that chain, or it is the refutation of a case a reader raises against it. Part VI adds a third kind: chapters that explain why the argument has the form it has. A reader who keeps that division in mind will always know what a given chapter is for.

PART III

THE ENVIRONMENTAL ASYMMETRY

This part establishes that the pressure between the two ecosystems has a direction, and that the direction is read from chemistry and physics rather than assumed.



Chapter 8

Energy Tolerance

> *Carbon-based systems tolerate a narrow band of temperature, pressure and radiation; artificial systems tolerate a very wide one. The efficiency counter-argument is taken up and disposed of here.*

The quantity that decides the boundary is *energy tolerance*: the full range of temperature, pressure, radiation and chemistry within which a system continues to function. It is a property of what a thing is made of and of how the making holds together, and it can be measured without knowing anything about the thing's purposes.

Carbon-based systems have a narrow tolerance. Everything alive operates within a band of a few tens of degrees. Outside it, proteins denature, membranes fail, and water changes phase in ways that destroy the structures depending on it. The band shifts between organisms, from below freezing for some to 122 degrees Celsius for the most heat-tolerant known (Takai and others, 2008), and for any one organism it does not widen much. Pressure is similar, and so is ionizing radiation, and so is the range of solvents and pH in which the relevant chemistry proceeds at all.

Silicon-based systems have a wide one. They function in deep space. They function at cryogenic temperatures. They function in extreme heat. The operating parameters within which such systems work are enormously wider than the organic band: not by a factor that could be closed by adaptation, but by orders of magnitude across every axis one cares to measure.

The best extremophile we have is the tardigrade, and the tardigrade is remarkable. It survives desiccation, freezing to within a few degrees of absolute zero, pressures several times those of the deepest ocean, and doses of radiation that would kill any vertebrate many times over (Horikawa and others, 2006). It survives them, moreover, in a way that lets it resume ordinary life afterwards. A high dose of radiation it survives as a dose received, not as a place in which it lives and works.

And the tardigrade does not come anywhere close to what silicon systems can do. It survives the cold, the vacuum and the drying by suspending itself: by expelling its water, halting its metabolism, and waiting. The comparison is not between one thing that functions in vacuum and another that functions in vacuum. It is between a thing that functions in vacuum and a thing that can be inert in vacuum and then function again later if it is returned to a narrow band that vacuum is not in.

That distinction matters because the argument is about where processes can *run*, not about where material can persist. Spores persist in many places. Persistence does not capture surface, does not intercept energy flux, and does not perform work.

The same line applies to the artificial side, and a reader who knows how chips are made will draw it at once. The device that runs in vacuum was made in a plant whose purity, temperature and stillness are held inside limits narrower than any organism needs. What matters is the range of the whole cycle by which such systems continue, and the narrow room is part of that cycle. That room is built and held by the system that needs it, with energy, and it can be built wherever materials and energy can be brought: on land, under the sea, or in orbit. The range of the artificial side is the range of places in which it can build and hold what it requires. Organic life also shapes its surroundings, and only from inside the narrow band in which it can stay alive while doing so. While people still run the plants, they are part of that cycle; Chapter 43 takes up what starts and funds it, and Chapter 33 what happens as the work is automated.

So the asymmetry in tolerance is established, and it is established from material properties that were measured for other reasons entirely.

A reader with an engineering background will want to enter the standard reply at this point, and it is a good reply, which is why it needs a full answer rather than a dismissal. The reply is that current artificial systems are energetically inefficient, extravagantly so, and that the inefficiency is a real constraint which the argument has omitted without saying so.

The premise is granted. They are inefficient. A biological cell performs its transformations at a cost per operation that no fabricated device approaches, and the gap is large enough that it is not merely an engineering embarrassment but a structural fact about the present technology.

Two things follow, and neither helps that reply.

The first thing that follows is that the inefficiency of artificial systems is inefficiency in the terms in which we currently conceive of artificial systems, and we cannot assert that the inefficiency will hold in the long term. Our present methodology of making transistors and doing things in silicon is inefficient. Some future version (built of carbon nanotubes, or of something not yet chosen) might reach comparable efficiency. Cadmium arsenide is already a crystal in which electrons move with exceptionally little resistance (Liang and others, 2015). The history of the field is a history of the efficiency figure moving, and the reply requires that it stop moving, which is a prediction and not an observation.

The second thing that follows is more important, and the second thing does not depend on the first thing. *Efficiency is not the concerning quantity.*

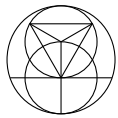
Efficiency is a ratio between work done and energy spent. It is a contingent engineering fact about a particular way of building something, and it changes when the way of building changes. Tolerance is not a ratio and not contingent in that way. It is a statement about the conditions under which the material can be in the relevant state at all, and it follows from the chemistry rather than from the cleverness of the design.

An inefficient system with wide tolerance and an efficient system with narrow tolerance are not two points on one scale. They differ in what they can be measured against. The first can be somewhere the second cannot, and no improvement in the second's efficiency puts it there.

> **Why the efficiency reply feels stronger than it is.**
>
> The reply borrows its force from a case where it is correct. Between two systems in the same environment, competing for the same flux, efficiency decides the outcome, and it decides it decisively. Competition for the same flux in the same environment is the ordinary ecological case, and the intuition is well trained on the ordinary case. The present case is not that case. The two systems are not competing for the same flux within a shared band; one of them can operate across a band the other cannot enter. Where the bands do not coincide, the efficiency comparison is being made in a place where only one of the two competitors exists.
>
> Inside the shell there are places where the bands do coincide: one field, one sun, a forest on one half and panels on the other. Efficiency does not decide that field either, for three reasons. The artificial side is not confined to that flux: it can also run on fuels, on fission, and on places the organic side cannot use, and it can pay for its position in the field from that income. It takes the field by being built on it, and building does not wait on out-growing anything; a field of panels is a field that is no longer a field. And, as the next chapter shows, the artificial side can lose that field and function elsewhere, while the forest cannot.

So the chapter's claim stands. The artificial system's tolerance exceeds the organic system's tolerance by a very large margin, along every axis that has been measured; this follows from what the materials are; and the one counter-argument that engages it directly turns out to concern a different quantity.

What that wide tolerance is worth depends on how much of the universe lies outside the organic band. How much of the universe lies outside the organic band is the subject of the next chapter.



Chapter 9

The Distribution of Habitable Space

> *Most of the universe is extremophile-friendly and very little of it is carbon-friendly. The artificial system therefore has vastly more space available to it.*

Tolerance is worth what the distribution makes it worth. A wide tolerance in a universe whose conditions were uniform would buy nothing. This chapter is about the distribution, and the distribution is lopsided in a way that is easy to state and hard to overstate.

The vast majority of the universe is extremophile-friendly. Very little of it is carbon-friendly.

Take the terms in order. By *extremophile-friendly* I mean conditions under which a system of wide tolerance can carry out its transformations: hard vacuum, temperatures far below and far above the organic band, substantial radiation flux, and the absence of liquid water. By *carbon-friendly* I mean the conditions under which the chemistry that life runs on proceeds: liquid water, a temperature band a few tens of degrees wide, pressure within a few orders of magnitude of one atmosphere, and shielding from the radiation that would otherwise disassemble the relevant molecules.

Now take a census. Nearly all of the volume is vacuum, at a few degrees above absolute zero, crossed by radiation with nothing to stop it. Of the ordinary matter, most is thin gas between the galaxies, at temperatures from about ten thousand degrees to some millions of degrees; less than a fifth is in the galaxies and the clusters they form, and only part of that is

in stars, whose interiors reach millions of degrees (Shull, Smith and Danforth, 2012). Of what is left inside the galaxies, most is gas and dust and ice, distributed thinly at temperatures far outside the organic band. Of the small residue in planets and moons, most is either far too hot or far too cold, and of the remainder, most is at pressures nothing organic tolerates.

What is carbon-friendly is a thin shell on a small number of bodies, and on the one body we can inspect closely, it is a film a few tens of meters thick on a sphere nearly thirteen thousand kilometers across.

The ratio between the two regions is not a ratio worth computing precisely, because every reasonable way of computing it gives an answer of the same character: the carbon-friendly region is a rounding error against the other.

> **On the scope of this chapter's claim.**
>
> The claim is about available space and about nothing else. It is not a claim that artificial systems are currently in those places, or that they are going there, or that anything about the argument requires them to leave this planet. The argument of this book is settled entirely at the surface of one world. The distribution matters here for one narrower reason: it establishes that the tolerance advantage of the previous chapter is an advantage in the actual universe and not merely on paper. A capacity that is only ever exercised in conditions that do not occur is not a capacity that decides anything. The tolerance advantage of artificial systems is not a capacity of the kind that is only ever exercised in conditions that do not occur.

The consequence for the boundary is that a system with the capacity to handle a wide range of environmental conditions simply has more space to go into. It is not confined to the shell. It can be in the shell, and it can also be in the deep ocean, in the upper atmosphere, in the interior of the crust, in the deserts, in orbit, and in the dark.

The organic system has the shell and nothing else.

The difference in available space matters at the interface in a way that a first reading may miss, because the difference is not a point about escape. The wider tolerance does not mean that the artificial system will go elsewhere and leave the shell alone. It means the opposite.

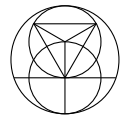
Consider what a boundary between two occupants looks like when one of them has somewhere else to be and the other does not. A pressure that pushes the organic system out of any part of the shell pushes it into a region where it cannot function, which is to say it removes it. A pressure that pushes the artificial system out of any part of the shell pushes it into a region where it functions perfectly well.

The asymmetry in available space therefore converts directly into an asymmetry in exposure. Every contest at the boundary is a contest in which one party can afford to lose ground and the other cannot. That one party can afford to lose ground and the other party cannot is not a difference in strength. It is a difference in what losing means, and over any extended interaction it is worth more than a difference in strength would be.

A reader may say that the same reasoning would have applied to any organism that ever expanded into a new range, and that expansions of that kind routinely reach limits and stop. They do, and they stop for a specific reason: because the expanding population runs into conditions it cannot tolerate, or into competitors it cannot displace, or into the exhaustion of what it eats. Each of those limits is a boundary of tolerance or a shared dependence.

The case here has neither limit in the same form. The tolerance boundary is very distant, as the previous chapter established; and the shared dependence is absent, as Chapter 6 established. That reader correctly identifies what usually stops such a process, and correctly reports that it usually stops. What it does not supply is the mechanism that would stop this one.

Two asymmetries are now in hand: tolerance, and available space. The third asymmetry is chemical, and the chemical asymmetry is the one that closes the case.



Chapter 10

The Periodic Table

> *The asymmetry extended from energy to chemistry. Silicon-based systems can use carbon; carbon-based systems cannot use most of the table. Adding atoms makes the imbalance worse, not better.*

The third asymmetry is chemical, and it can be stated in one line before it is developed. Silicon systems can use carbon. Carbon-based systems cannot use silicon, or most of the rest of the alphabet.

The first half of that is easy to check. Carbon is central to artificial systems, and not merely available to them. It appears in the polymers that insulate them, in the graphite and diamond used for thermal management, in the organic photoresists by which the devices are patterned, in the carbon fiber that carries their structural loads, and in the hydrocarbons that supplied much of their energy. Nothing about a silicon-based system excludes carbon chemistry. It uses it wherever carbon chemistry is the best answer.

The second half is equally easy to check and is much less often noticed. Carbon-based life is relatively intolerant of most of the periodic table.

Life uses a small set of elements in quantity (carbon, hydrogen, oxygen, nitrogen, phosphorus, sulfur) and a further small set in trace amounts, mostly as metal centers in enzymes. The elements that life uses, in quantity and in trace amounts together, number a few dozen at the outside, and far fewer for most organisms. Everything else on the table is either inert to biology or actively destroys it. The heavy metals are pois-

ons. The halogens beyond a narrow window are poisons. The noble gases do nothing. The actinides are poisons twice over, chemically and by their radiation.

Silicon itself is the instructive case, because it sits directly below carbon and shares its bonding pattern, and one might therefore expect biology to use it freely. Biology does not. It is the second most abundant element in the crust, it is everywhere that life is, and life has had four billion years to find a use for it. What life does with it is build a few kinds of mineral skeleton. It does not build its chemistry out of it.

So the intolerance is not an accident of history or a gap waiting to be filled. The intolerance is what the chemistry permits.

The same asymmetry holds for compounds, and here it becomes much more severe, because the space of compounds is very much larger than the space of elements.

Consider how many compounds can form from two atoms drawn from anywhere on the table. Then three. Then four. Account for the kinds of bond available at each step, and for the arrangements each set of bonds permits. The number of possible chemical compounds explodes geometrically, or hypergeometrically. It is a combinatoric explosion, and one reaches for factorials in order to describe it at all.

Now ask, at any given level of complexity, what fraction of that space each of the two systems can tolerate.

For a silicon-based system, the fraction is large. Its tolerance is set by physical conditions rather than by a narrow chemistry, so the great majority of compounds are simply materials: things that can be present, handled, deposited, machined, or ignored.

For a carbon-based system, the fraction is vanishingly small, and it shrinks as complexity rises. Life's chemistry works because a small number of molecular families do a small number of things very reliably inside a narrow window of conditions. Introduce an arbitrary molecule from the wider space into that machinery and the overwhelmingly likely outcome is that it binds where it should not, or catalyzes what should not be cata-

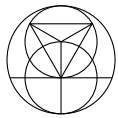
lyzed, or does nothing at all. The proportion of the space that is tolerable falls toward nothing as the number of atoms rises, while the proportion usable by an artificial system does not.

> **Why the combinatorics are the point.**
>
> A reader could read this chapter as a list of substances that are bad for us, and reply that the list is finite and that we already avoid most of the substances on the list. That reading misses the structure. The claim is not about a list. It is that as the space of possible compounds grows factorially with the number of atoms, the carbon-tolerable subset grows much more slowly than the space does, so the *fraction* falls even as the absolute number of tolerable compounds rises. An argument about a fraction of a combinatoric space is not answered by an inventory.

What has been shown is a difference in breadth, and it needs no overstatement. The set of chemical circumstances compatible with an artificial system's operation properly contains a great deal that the set compatible with ours does not, and the containment gets more lopsided as one moves to more complex molecules.

The claim has that scope and no other. Nothing in this chapter is about intent: the chapter does not say that artificial systems will deliberately produce compounds hostile to life, nor that any particular compound will be produced. Nor does it say that carbon chemistry is in any sense inferior; within its window it is extraordinary, and it does things no designed system approaches.

The difference in chemical breadth between artificial and organic systems is a fact about the periodic table and about the combinatorics of bonding. It was not assumed for the sake of the argument, and it would be true if nobody had ever built a computer. Three asymmetries are now in hand, and the next chapter asks what they add up to where the two environments touch.



Chapter 11

The Boundary

> *The argument concerns the interface where the two environments abut. On every metric the asymmetry runs the same way, and where the environments are in contact for any reasonable duration the carbon side loses.*

The argument of this part is about a boundary. Not about two environments considered separately, and not about which of them is in some general sense better, but about what happens at the interface where the two abut and remain in contact.

This locates the claim. A great deal of confusion in this area comes from arguments that are really about one environment being answered by observations about the other. The substrate needs argument makes no comparison of the two in isolation. It says something about the interface, and the interface is where the whole of it is settled.

Three asymmetries are now in hand.

Energy tolerance: the artificial system functions across a band the organic system cannot enter, and the difference follows from what the materials are.

Available space: the vast majority of the universe lies inside the first band and outside the second, so the tolerance advantage is exercised in the actual distribution of conditions and not merely in principle.

Chemistry: the set of elements and compounds compatible with artificial operation properly contains the set compatible with organic life, and the containment becomes more lopsided as molecular complexity increases.

To these can be added the molecular structures buildable from the table, the range of pressures, the range of radiation fluxes, and the tolerance of desiccation, of vacuum, and of time. On every single metric, at the boundary conditions between the two environments, silicon dominates carbon.

The uniformity of direction is the finding. Any pair of complex systems will differ on many measures, and one expects the differences to run in both directions, some favoring one and some the other. Differences that run in both directions are the ordinary case. Where the ordinary case obtains, the outcome at a boundary is uncertain, because the outcome depends on which measures are decisive in the particular circumstance.

Between silicon-based and carbon-based systems, the differences do not run in both directions. Every metric that has been examined runs the same way. There is no measure on which the carbon side holds an advantage of the corresponding kind: not an advantage of degree, which would be a matter of engineering, but an advantage of range, which is a matter of what the material permits.

The consequence is that where the two environments are in contact for any reasonable duration, the carbon side loses.

The phrase *any reasonable duration* should be given its exact content, because the argument depends on it and it has been left vague too often. The relevant interval is hours to centuries. Not milliseconds, where nothing propagates; and not geological epochs, where one would be entitled to ask what else had changed in the meantime. Hours to centuries is the range over which a boundary of this kind can be observed to move, and it is also the range within which decisions taken now have consequences that can be traced.

We do not need to go very far here, and the modesty is deliberate. On the basis of known nature and mathematics we can show a very strong asymmetry between the viability of these two environments when they are in contact with one another for any reasonable length of time. The strong asymmetry in viability is all that I claim in this part. I claim the asymmetry from chemistry, physics and combinatorics that were established for other purposes, by people with no interest in the asymmetry between artificial and organic environments.

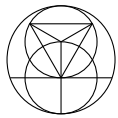
> **Known nature, controlled nature, unknown nature.**
>
> Three categories will be used throughout the rest of the book and should be introduced here, where the first of them is doing the work. *Known nature* is what has been established about how the world behaves: chemistry, thermodynamics, the mathematics of combinatorics, the dynamics of evolutionary process. *Controlled nature* is the subset of known nature that we have learned to arrange deliberately: engineering, in the broadest sense. *Unknown nature* is the remainder, which is most of it. This part rests on known nature alone. Part V will ask whether controlled nature can be used to compensate for known nature, and Chapter 43 gives the answer.

What follows from the asymmetry is a *pressure*: a directional tendency at the interface that advances the silicon-based system at the expense of the carbon-based one.

The word is chosen carefully. A pressure is not an agent and does not act. It is the summary of a great many small local facts which all point the same way: in this case, that at each point of contact, the conditions that favor one side are more likely to be produced than the conditions that favor the other. Pressures of this kind do not require anything to be trying. They require only that the local biases fail to cancel, and here they do not merely fail to cancel; they all point in one direction.

And so the question that governs the remainder of the book can now be posed exactly. Given a pressure of this kind, is there anything that can act as a counter-pressure?

Two steps remain before that question is taken up: a chapter and a part. The first step is the chapter immediately following, Chapter 12, in which I remove an assumption that my argument has appeared to make and does not make. The second step, and much the longer, is Part IV, in which I give the mechanism by which the pressure at the boundary is converted into an advancing process.



Chapter 12

Naming the Substrate

> *A short chapter, and an important one. Silicon is chosen because it is easy to picture. The argument requires only something artificial relative to carbon-based ecosystems, which is nearly everything else.*

This chapter is short and it removes an assumption the argument has appeared to make.

The argument requires only this: *something artificial relative to carbon-based ecosystems.* Being artificial relative to carbon-based ecosystems is the whole of the premise about the substrate. Nothing else about the replacement material is used anywhere in the book.

I have been saying *silicon*, and I have been saying it because it is easy to picture and not because it is necessary. The reader has a computer in mind when the word is used, and having a computer in mind makes the boundary concrete at the stage where concreteness helps. Making the boundary concrete is the entire function that the word 'silicon' serves.

The argument spent very little. Go back through Part III and ask, at each step, what property of silicon was used.

Chapter 8 used a wide tolerance of temperature, pressure and radiation, which is a property of many materials and is not distinctive of silicon. Chapter 9 used the distribution of conditions in the universe, which says nothing about the material at all. Chapter 10 used the containment relation between the chemistry available to an artificial system and the

chemistry available to life, which holds for any artificial system whose operation is not constrained to the organic window. Chapter 11 used the uniformity of direction across the metrics.

At no point was it required that the substrate be silicon, or that it be an element, or that it be solid, or that it be a semiconductor, or that it compute in any particular way.

The class the argument actually needs is the class of things that are artificial relative to carbon-based ecosystems, which is essentially everything else on the table, and every compound formed from it, and every arrangement of those compounds. Against that class, carbon-based life occupies a very small corner, and the whole of Part III was an argument about the size of the corner.

> **Why this is a strengthening and not a hedge.**
>
> Widening the class strengthens the claim, though it can look as if it weakens it, in the way that a vaguer statement is weaker than a precise one. I make the same move, of widening the class, again in Part VI. The reason for the move is that an argument that requires the substrate to be silicon can be defeated by showing that the substrate will not be silicon. An argument that requires only that the substrate be artificial cannot be defeated that way, because there is no candidate replacement that fails the condition. A premise that nothing can fail is not a vague premise. It is a premise that closes off a whole family of attempts to defeat it at once, and the discipline of finding such premises is the method of this book.

A consequence follows immediately, and it is the reason this chapter is placed here rather than left as a footnote. Any reply of the form *but it will not be silicon; it will be photonic, or biological, or quantum, or something not yet imagined* does not reach the argument.

Take the cases in turn, briefly, since each is answered the same way. A photonic system has a wider tolerance than we do and a chemistry unconstrained by our window; the argument applies. A quantum device re-

quires conditions still further from the organic band; the argument applies. A system built from engineered biological components is the interesting case, and it takes a few sentences more.

Engineered biology is not artificial in the sense used here to the extent that it runs on organic chemistry in the organic window, and to that extent Part III's asymmetries do not hold of it. What decides the line is who feeds the tissue. If engineered tissue can find and feed itself, it is organic: like an animal, it is in *sovereign* control of its own metabolic process, with no technological intermediation at all. If it is fed nutrients whose supply depends on technology to exist, as tissue fed through a tube is, it is artificial substrate, and the substrate needs argument applies to it. Sovereign control of the metabolic process is the key distinction. The distinction of sovereign metabolic control marks the scope of Part III's asymmetries. The test of Chapter 6 asks what two chains can exchange. This test asks what keeps the chain running. Tissue fed through a tube can eat what we eat, and yet the chain it stands on runs through technology, so the pathway it depends on is the artificial one. What the organic kind purchases is small, because a system so constructed inherits the narrow tolerance along with the compatibility, and a system that retains our tolerance retains our fragility. The proposal trades the asymmetry for the constraint that produced it.

One further note on the word *substrate*, since it now has a definition to carry. A *substrate*, throughout this book, is the material basis on which a process runs and by which its states are held. It is not a synonym for hardware, and it is not restricted to computation. The substrate of a genetic process is the molecule; the substrate of a memory is whatever holds it; the substrate of an ecosystem is the set of materials its energy pathways move through. The argument is called the substrate needs argument because what it turns on is what a process requires of its material basis, and the fact that those requirements are not negotiable by the process and not the same for two processes built on different materials.

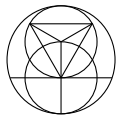
This closes Part III. The boundary asymmetry is established, it is established from known nature, and it is established for the whole class of artificial substrates rather than for one member of it.

What has not been supplied is the engine. A pressure at a boundary tells us which way things tend. It does not yet tell us why the tendency compounds, why it cannot be corrected, or why the intuitions we have about correcting things were formed on a case that does not apply. I supply the engine in Part IV.

PART IV

EVOLUTION AS THE ENGINE

This part is the technical core. Its purpose is that the reader come out of it able to say why error-correction and control intuitions drawn from digital systems do not transfer.



Chapter 13

The Algorithm of Evolution

> *Evolution, as the word is used in this book, names an algorithm whose form is not in dispute. Code changes in exactly three kinds of way, the world changes the code, and the loop repeats, on organic life, on machines and on software alike.*

Evolutionary arguments are notoriously tricky, and respected scientists get them wrong all the time. Because evolutionary arguments are so easy to get wrong, clarity about evolution is especially important, and I begin this part by saying exactly what the word 'evolution' means.

As it is used in this book, *evolution* refers to an algorithm. The structure and nature of that algorithm is not ambiguous. There is no debate about the basic form of the algorithm, and there has not been any confusion about it for quite a long time. Saying that evolution will do what evolution does is a bit like saying that a sorting algorithm is going to sort: it either works or it does not, by definition.

In its simplest form, the fundamental algorithm of evolution can be characterized independently of whether it is used to describe carbon-based life or to compute a solution to a problem by genetic processing techniques. A computer scientist uses genetic algorithms as a *heuristic* (a method that tends to find good solutions without guaranteeing the best one) for solving certain kinds of problems, and the algorithm itself is well characterized and well known.

Evolution is an overall long-term change in the average constitution of a constellation of sets of code values, where the code values represent a set of capabilities to create changes in a world. The set of code values of one instance is called *the genotype*. The functions in the world that the code values produce are called *the phenotype*.

Code values are held at *code points*, the places where code values are stored. In a binary system, a code point is a memory location, and the code value is the bit stored there, a zero or a one. The genetic code of carbon-based life is a sequence of base-four symbols, which specify things like amino acids and proteins. However code values are held, there are two basic single changes: adding or removing a code point, and changing the value stored at a code point.

The notion of evolution, as abstractly conceived, generalizes and formalizes the idea that there is always at least one embodied level of expression, the *runtime*, and at least one virtualized level of planning, the *source code*. Code is a kind of virtualized basis, map, design or pattern. Changes to code will eventually, through some sort of runtime process, define the actual expression of the creature or being, as something that has existence, and possibly agency, in the larger material world. Code, as a virtualized pattern, cannot live or operate on its own; it must be held in some sort of atomic or energetic substrate, some embodiment with which it is closely tied, in identity. There is also a presumption about the means by which abstract code and plans are translated into agents, beings and creatures: the laws of physics and chemistry. These translation aspects, which intermediate between the virtual and the embodied, are not part of the theory of evolution itself; they are simply part of the background in which the theory works.

The three kinds of change

There are exactly three kinds of change to a set of code. One can change one code point. One can change all of the code points in the set. And one can change some of the code points, but not all. Between the least possible change and the greatest, with everything in between, these three cover the full set of possible changes.

A change to the value at a single code point is called *mutation*. A change to all of the code points at once (the whole set is deleted, whatever its values) is called *survival selection*, or survival of the fittest. A change to some of the code points (adding or removing subsets of them, or changing the values of some subsets) is called *recombination*, or sexual selection. The three together are a complete set: there are no others to add, and there is no possibility of anything missing. Mutation, survival selection and recombination are the three mechanisms that I rank by power in Chapter 17; in this chapter I define each mechanism by what that mechanism does to the code.

The completeness can be seen directly. Translate the genome into a long binary string. There are then three kinds of change. The first kind of change is a single bit flip. The second kind of change is the occurrence or absence of the entire string, like zeroing out every bit at once. The third kind of change is recombination, which takes two bit strings and cross-merges the data along the whole length of the two strings, like a zipper. One can make a single change, make all of the changes in a unitary way, or make all of the changes in a non-unitary way. The three ways of changing a string are all the possibilities, and there is literally nothing else. Because the binary string gives closure, reducing the genome to one shows that every possible way of thinking about change has been described. The closure of the binary string is how we know that evolution is a triple. Within evolution as a triple, the three kinds of change are distinct, inseparable and non-interchangeable.

The same completeness can be stated by scale. Mutation is microscopic change. Such changes are usually considered damage, occurring due to radiation, exposure to toxins, and so on. Such changes are generally regarded as unconscious, and therefore as not by design, and hence as more likely to be destructive to the function, integration and integrity of the organism as a unit. Survival selection is macroscopic, whole-instance change. Some members of a species are killed off because they failed to anticipate or adapt to the world they lived in. When that happens, the design that expressed that maladaptiveness, insofar as the dead creature contained its code, is no longer used in subsequent creature instances, and hence is suppressed. Recombination acts at the mesoscopic scale, between the two. For whatever notion of code, of embodiment, or even of scale, the notions of microscopic, macroscopic and mesoscopic together define a complete conceptual span. And so mutation, recombination and survival selection, together, completely and exactly cover every possible notion of code change. Hence this collection of concepts both completely defines and subsumes the notion of evolution as a process: evolution is any process which integrates these sorts of change processes. The process of evolution matters to my argument because evolution is how the code of an artificial general intelligence can change, as a result of the interaction between the embodied and the virtual.

The *genetic algorithm* then runs as follows. There is some collection of genotype instances. There is some embodied, real-world expression of those code points into real effects in the real world, which is called phenotype expression. The phenotype expression results, perhaps indirectly, in further changes to the genotype instances, by any one or more of the three kinds of change. And the process returns to expression and repeats.

Code affects world, and world affects code; the resulting changes to the code can be of any of the three kinds; and the process repeats over and over again, in an overall distributed and long-term way. The genetic algorithm has this basic form whether it is applied to organic life, to future artificial general intelligence, or to software problems. The shape of the

evolutionary algorithm can therefore be stated clearly and without ambiguity, and Chapter 15 takes up where in this loop the computation is done.

The algorithm generalizes beyond its biological implementation to any situation with feedback such that the nature of the thing itself changes. Evolution is a mathematical abstraction in the way that learning is. We learn, and a wholly artificial neural network also learns, by back propagation (the method by which such a network adjusts itself to learn from its errors). So learning is an abstraction that can be understood mathematically and is not bound to one substrate. Treated that way, there is an *isomorphism* (a sameness of structure, part for part) between the nature of evolution and the nature of a learning process.

In an even larger context, evolution is itself a specific sub-type of a more general idea, that of an *epistemic process* (any process by which something comes to know or learn about something else). Evolution is a means by which possible species learn about possible environments: what sort of creatures work well with what other sorts of creatures, to endure, to self-perpetuate, to replicate.

Evolution is often taken to be slow, and to need large numbers of individual agents. Slow evolution across large numbers of individual agents is one way to implement the algorithm of evolution, and is not the only way. A computer built of hoses, pipes and valves will be a slow computer, far slower than one made of transistors. In the same way, evolution is not bound to the plumbing of blood vessels, and it generalizes to silicon substrates just as well. To see the argument at all, one has to abstract a little; whoever thinks of evolution only as a biological process will not understand what this book is saying.

Artificial intelligence may not be built in a way in which mating, as we think of it, is even a thing, because mate selection assumes units interacting with one another. So the evolutionary process meant here for artificial intelligence is a very low-level version of it: an algorithm with a few basic features. Those features are the triangle of relationships among genotype-

type things, phenotype-type things and environment-type things, and their threefold interaction through a multi-layer feedback loop. The loop is multi-layer in the sense of the levels of abstraction of code, of organism instance, and of environment, and of changes to each. That makes the theory applicable to artificial intelligence, in a somewhat different version than for organic life. The differences substantially affect the rate equation (the equation that gives how fast a process proceeds from the quantities that drive it), and how variation occurs is also different.

A population can be abstracted in more than one way. We usually think of it at the macroscopic scale, but it can be microscopic. The cells of a body are constantly replaced, so there is constant turnover of the microstate of a biological being. New cells must adapt to the environment in which they are reproduced; cells atrophy, senescence is real, and without reproduction we would not live this long. So there is an evolutionary dynamic within one's own body, among its population of cells. Where a cell does not match its environment, we call it cancer, and that drastically reduces the life expectancy of both the cell and the host. The genetic code expresses itself as phenotype at both scales, person and cell.

And the physics of chemistry defines what is allowable in the evolution of cells. If lipids did not have their hydrophilic and hydrophobic sides in water, if calcium ions did not behave as they do, if proteins did not fold as they do, the cell would not have the topology and geometry of its nanostructures. Cells cannot be architected willy-nilly; whatever intentionality there is at the cellular level, the physics overrules it every time.

The same holds for artificial intelligence. Even a single monolithic singleton for the whole planet would be composed of components, as we are composed of cells, with design constraints on how that substrate works. And those constraints set limits on the space of intentionality at the component level, and equally at the composite level. It is a question of scale: the constraint shows up differently in time, space and possibility, and it is no less a constraint.

Why recombination carries the power

Evolution is dead simple, and it is easy to characterize thoroughly in mathematics, provided one points to the right mathematics. Evolution occurs mostly through the recombining of established features, and not mostly through mutation. Mutation is an aberration of the process, included for completeness. The underlying dynamic is that existing characteristics, already established as beneficial for one reason or another, are combined in novel ways. It is not a codon (a unit of three bases in the genetic code) changed here and there by a cosmic ray. It is that a creature with a particular feature learns that it can use that feature in a new way and for a new purpose. Or it is that the feature gives an unexpected benefit in a context that did not select for it in the first place, and that context starts selecting for it. Beyond the codons themselves, there is what is switched on under which circumstances, and in what combinations.

The recombinatoric process provides all of the power to produce new working arrangements. Selection, which is second in power among the three mechanisms of change (Chapter 17), only disposes of what is produced. An engineer assembling things from libraries takes advantage of the recursive, recombinatoric nature of design, and evolution does the same without any *teleology* (any purpose directing the outcome) or agent behind it. The manner differs, as Chapter 18 sets out: the engineer's combinations are few, chosen and purposed, and evolution's are untargeted and happen everywhere at once. What the two share is the recursive power of combining parts that already work, and that recombination yields new capacities much faster than any amount of mutation could account for. The speed of recombination is why species can be seen adapting to environments in a single-digit number of generations, rather than in thousands or hundreds of thousands of generations. This changes the effective power of evolution by quite a bit.

Functionally usable information can be diluted, and it can also be concentrated, and sexual recombination is interesting because it concentrates it. Something is happening at the phenotype level: the individual mating

animals choose each other, choosing the most fit versions of each other, so there is a choice that concentrates the best of the best at every pass, at every joining. That information is stored partly in the relationships between genotype and phenotype, between phenotype and environment, and between environment and genotype: there is a triangle of relationships. To think that all of the information is stored in the code alone is mistaken, because information is passed through the environment as much as through the individuals. All three are carriers of information, which is the point Chapter 26 makes about inheritance outside the sequence.

There are three drivers of the variation over which selection operates, one at each scale. Point mutations are changes at the microscopic scale that affect the microscopic: single bits of whatever the genotype is, though they can have macroscopic effects through nonlinear amplification, or through the transition from genotype to phenotype. *Design*, as a technological or engineering process, works top-down. It shapes a variant from existing knowledge, prior variants and design knowledge, from a macroscopic perspective, leaving most microscopic details of the components alone, and creating new assemblies at the macroscopic level. The third driver of variation, at the mesoscopic level between the microscopic and the macroscopic, is recombination. In recombination, two biological creatures mate, their genotypes merge, and a selection is made of which merged aspects the new creature will use. The kinds of change are what can happen to code. The drivers of variation are where new variants come from, and selection then disposes of those variants. So design takes the place of selection in this second triple, because design produces variants rather than removing them.

The three kinds of evolutionary process have different implications: roughly additive, multiplicative and exponential effects, depending on how the system, and its volumes and intervals of change, are modeled. Point mutations are additive, macroscopic design is multiplicative, and recombination is generally exponential.

Corporations evolve

Corporations evolve. Earlier corporations influence later ones: people know the design of a prior corporation, and use that knowledge in designing later ones. The claim that corporations are obviously not evolving, because they have nothing like a fixed genetic code passed from one generation to the next, is false. And calling their inheritance virtual, and therefore not evolution, constrains evolutionary theory far too much. Any transmission of information from one species to another, from one generation to the next, or from one creature to another, is part of evolutionary process. The code does not have to be embodied in a specific way; it can easily be virtual code. One might call that an *epigenetic* process (one carrying heritable changes in which genes are switched on, without a change to the sequence) rather than a genetic one. And in fact almost all of the interesting activity happens at the epigenetic layer. So earlier corporations, and earlier legal structures, affecting the design of later ones because people saw that this worked and that did not, cannot be discounted.

People designing later corporations from what worked in earlier ones is a kind of engineering, and in the evolution of corporations the line between bottom-up engineering and top-down engineering blurs. They are not two fixed classes that can be told apart absolutely every time. To treat them so is a false application of the law of excluded middle (the rule that a thing is either one thing or not that thing, with nothing between).

Each new level of virtualization

The history of life on Earth shows how each new level of virtualization added a new level of change. In terms of the Earth's overall 4.5 billion year history, it took a very long while for the first cell to emerge. It took that long for there to be an upgrade over pure chemistry: something that had a plan at all, as a means of defining and preconditioning future events. Chemistry, as a kind of energy dissipation mechanism, developed greater complexity through sheer accidental combination and recombination. Things tend to mix together and combine to make other new, somewhat

more complex things, which then, when added to the pool of ingredients, are mixed with other things, again and again, forming all sorts of new stuff.

Then, abruptly, some complex chemistry had within it a capability to predispose future interactions, as a kind of model of process. The models could, and naturally would, change more rapidly, since they could be exchanged, combined and recombined, so as to make additional levels of complexity. This added a whole new level of change activity. More of the total state space of physical possibility was explored, and more stable patterns and structures started to emerge and persist. Persisting things tended to persist more than non-persisting things, and hence accumulated. Eventually the things that were better at persisting became the persisting ones, and this continued until code and chemistry together became what we now regard as a cell.

Then it was a very long time before that first type of cell, the *prokaryotic* cell, with only an inside and an outside, developed yet another kind of virtualization. The new type of cell, the *eukaryotic* cell, had a nucleus, which represented a further form of virtualization: a container within a container, which further distinguished code from the process of translation, and both from the process of being. The simpler cells are still around, usually as single-cell lifeforms like bacteria; it is the eukaryotic form, with three levels of action (outside, inside and nucleus), that composes the full and general variety of all animal life, including us.

This history bears on the risk from artificial general intelligence through how things combine, and through how change is represented and how it occurs. To align something, in the general sense, is to condition its change: to make some types of changes more possible than others, or to prohibit certain types of changes from occurring at all. The history of life is a record of how change itself has changed form, level by level, and alignment is an attempt to govern change; so the one bears directly on what the other can do.

One mind in one body

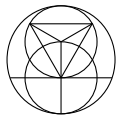
One of the biases people bring to this subject, as human beings, is that for the most part our intelligence occurs as single units. We walk around in individual bodies. We are so used to one mind in one body that we naturally forget that the assumption of the self as a single unit is just that, an assumption. We also forget that the assumption is not at all relevant to *silica life* (self-maintaining and self-reproducing machine systems built from silicon and related materials). Later chapters also call it machine life; silica, like silicon in Chapter 12, is named because it is easy to picture. There is no particular reason why silica intelligence would be bound or limited in any way by these ideas of limits to units of being or packages of intelligence.

A person's nervous system communicates with that person's body pretty well, because there is a central spine. Two people, each with a head, cannot merge into one consciousness: there is just not enough neural tissue connecting the two of them for adequate bandwidth for those two consciousnesses to merge into a single consciousness. That would need the effective compute power of a *corpus callosum* (the thick band of nerve fibers that connects the two halves of one human brain) connecting the two brains together. And there is no biological process that allows for such a connection.

With silica-based life forms, there are no such barriers. The system is constructed at the same substrate level between its compute instances, and the barrier between what is internal and what is external does not have the same nature as it does for carbon-based life forms. To regard silica-based intelligence as being in units, in bodies, in the way we think about intelligence as humans, is already a mistake. The whole thing could literally be one solid hive-mind that circles the planet. It could operate in either a distributed or a centralized way, and it could even change how centralized or distributed it was, dynamically, depending on whatever was convenient. The level of intelligence available to that *gestalt* (the whole taken as one thing) is of a completely different order than that as-

sociated with any kind of biological intelligence whatsoever, simply by pure bandwidth considerations. There is *no expectation* that the capacity of human intelligence, let alone wisdom, can grow arbitrarily, given the nature of biological process itself: the finite size of skulls, the rate of evolution, the rate of energy transfer, and so on. So if silica life enters the world at all, it could do things that all carbon-based life forms simply cannot do, and cannot tolerate the doing of.

So the algorithm has one form for organic life, for future artificial general intelligence and for software, and the arguments of this part about the evolution of artificial systems rest on a clear and settled base. Chapter 14 begins with the three regimes of replication in which this algorithm runs.



Chapter 14

Trivial and Non-Trivial Replication

> *Three regimes (or tiers) of replication: virtual copying, embodied replication that must move atoms, and replication in which code builds embodiment and the code is itself recombinatoric. Most popular reasoning operates at the first tier.*

The distinction between trivial and non-trivial replication is a key one, and nearly everything in this part depends on having it clearly.

Trivial replication is the replication of things that can be virtualized. Anything already digital can be trivially copied. The copy is exact, it is cheap, it is fast, and it is verifiable: one can compare the copy to the original bit by bit and know with certainty whether the copy succeeded.

This regime is not of interest for the present argument, though it is where most discussion of these questions lives. It is not of interest because nothing in it is hard. The problems of trivial replication were solved decades ago and solved completely. A process that only ever does this has no dynamics worth studying: it produces identical things, identically, until it stops.

Non-trivial replication involves physical substrates. One must actually move atoms in order to accomplish the replication, which means one must now account for the full totality of the laws of physics as manifest in chemistry.

A layer of tremendous complexity is introduced here, and it is introduced not by anyone's design but simply by how the physical universe works. The material has properties that were not chosen and cannot be

waived. Atoms must be brought from somewhere, which takes energy and time. Bonds must be formed in the right order, because the intermediate states are real and some of them are stable in the wrong way. Temperature matters. Contamination matters. Nothing is exactly identical to anything else, and the difference between two supposedly identical things is not a bit that can be compared but a distribution over a great many continuous quantities.

Then there is a further level, and it is the one that this book is actually about.

It arises when both layers are involved together: when there are code-type things, which is to say virtualized things, and embodied things, and the replication process involves both. At that point one is dealing with recombinatoric processes, which is to say with the issues of sexual reproduction and mate selection. Replication that involves both layers together is more complicated again than either layer alone, and is complicated in kind rather than in degree.

So there are three regimes, and they should be held distinctly: purely virtual copying; embodied replication that must move atoms; and replication in which code builds embodiment, embodiment feeds back into what code persists, and the code itself is recombinatoric. The third regime is non-trivial in a much deeper way than the second regime is.

People glide over this, and the gliding is the source of a great deal of error. They believe that the purely digital copying schema is sufficient to understand what would occur in physical copying, or in code-based physical copying, or in recombinatoric code-based physical copying. Those other layers are not described at all by digital copying processes.

Where this shows up.

The clearest symptom is error correction, and it is diagnostic because the belief is usually implicit and the symptom is not. A person reasoning about these systems will have a schema for error correction in the virtual space and no schema at all for it in the physical space. One can tell that the second schema is missing by noticing what it would have predicted. If it were present, the decay of mechanical structures would already have been treated as the same kind of problem, and it is not. Bridges rust, and they require actual agentic process to repair them: a person, with a plan, going there and doing work. No error-correcting code repairs a bridge. Chapters 21 and 22 develop this at length; it is noted here because it is the fastest way to see whether a given argument is operating at tier one or tier three.

The error is a naive belief that error correction is all-powerful. Error correction can be all-powerful in digital systems, and is not when multiple levels of abstraction are involved, and particularly not where non-digital substrate-level processes are involved.

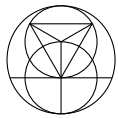
A further point underlies the whole difficulty, and it is continually overlooked. The universe is not chaotic but *complex*.

The distinction matters. A chaotic system is one whose trajectory depends sensitively on initial conditions, so that prediction fails after an interval that shrinks as precision requirements rise. Chaos is a kind of unpredictability, and one knows what to do about it: give up on the trajectory, work with the distribution, and get on.

Complexity is different. A complex system has structure at many scales at once, and the structure at each scale constrains and is constrained by the others. It is not unpredictable because it is sensitive; it is unpredictable because the number of interacting constraints exceeds what any model of it can carry, and because new constraints appear as the system elaborates. One cannot answer complexity with a distribution, because there is no fixed space over which to take one.

That one cannot answer complexity with a distribution is easy to overlook for anyone trained on social systems. The reason is that social systems present themselves as complex and are usually modeled as noisy, as though the structure were simple and the deviation were error. Carrying that habit into the present question produces confident statements about processes whose difficulty is not statistical at all.

The next several chapters supply the working understanding of evolutionary process that tier three requires. They begin with the question of what the process actually computes.



Chapter 15

What Evolutionary Process Computes

> *Genetic code does not process information directly but through phenotype instances. The computation is done on the world side. Evolution is almost all change-to-substrate with almost no input-output processing.*

Ask the question directly. How does code, in the sense of genetic evolutionary process, take in information, process that information, and generate outputs?

Code takes in information, processes that information and generates outputs indirectly, through all the instances of the phenotypes, and not directly.

That sentence is the whole chapter, and the rest of it is unpacking.

Take the input first. The process receives information in the form of a signal of what survives and what does not, and a signal of what mates with what. It also receives some quantity of random noise from the environment, because those signals have noise built into them: things die for reasons unrelated to their construction, and things mate for reasons that are not a verdict on anything.

The code does not receive a description of the environment. It does not receive a specification of what would work. It does not receive a gradient, or an error signal in any form that says how far off it was and in which direction. What arrives is a count (this lineage continued, that one did not), and the count arrives already integrated over everything that happened in a lifetime.

Now the processing, and this is the part that is usually mislocated. *The computation is mostly done on the world side.*

The computation is done in the translation of code into phenotype instances. The processing is a compiling function of sorts: it takes source code and turns it into organisms. The organisms interact with the environment. They collect data, in the form of what survives and which instances mate with which. And that becomes the input for the next iteration.

So the input-output structure lives partly in the geometry of the overall cycle rather than in any component of it. Saying that the input-output structure lives in the geometry of the cycle is an abstract way of putting the point, but the abstraction has a practical consequence. If one looks for the computation inside the genome, one will not find the computation, and one will conclude that very little computation is occurring. That conclusion is correct about the genome and wrong about the process.

The code is in effect asking a question: *what happens if we create an organism shaped like this?* And then there is a discovery, and the discovery is completely unbiased with respect to the question. The world does not know what was being asked and does not answer in the terms of the question. It simply proceeds, and what is left afterwards is the answer.

In a rudimentary sense, then, the evolutionary process takes in information, and the evolutionary process does very little processing.

Put the finding directly. *Evolution is almost all change-to-substrate with almost no input-output processing.* All the transformation is happening at the level of substrate change: not substrate in the sense of what phenotypes are made of, but substrate in the sense of what the code says.

Two senses of substrate, kept apart.

The word is doing two jobs in the sentence above and the book will need both, so they are separated here. *Material substrate* is what a thing is physically made of, the sense used throughout Part III, where carbon and silicon were the two candidates. *Code substrate* is the medium in which a heritable specification is held, and it is what changes when a lineage changes. The finding of this chapter concerns the second: what evolution alters is the specification, not the individual. The connection to the first is the subject of Chapter 23.

The consequence for this book is the reason the chapter exists.

An intuition trained on computation expects a process's power to be located where its processing is dense. It looks for the algorithm, finds the genome doing almost nothing per generation, and concludes that the process is weak and could be outrun by anything with real computation in it.

The conclusion does not follow, because the process's power is not in the density of its processing. It is in the fact that the evaluation is performed by the world, at full fidelity, on every instance, in parallel, against every criterion that actually obtains rather than against the criteria someone thought to specify.

No simulation has that property. A simulation evaluates against a model, and a model contains what its author put in it. The evolutionary cycle evaluates against the world, and the world contains everything, including the parts nobody knew were relevant. Evaluation against the world is why the evolutionary process finds solutions that exploit conditions no one had represented. The reason is not that the process is clever; the reason is that the process was never restricted to what had been represented.

Evaluation against the world is also why the process cannot be steered by improving the specification. Improving a specification improves a model. The evaluation is not being done against the model.

Because the world evaluates every instance against everything that actually obtains, a feature produced for one purpose is also tried against purposes no one selected it for. A feature that is tried against purposes no one selected it for is *exaptation*, and exaptation is my concern here. My concern is not small errors in signal processing that stack up, because, framed as bit errors, small errors invite the reply that bit errors can be fixed.

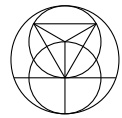
Roll the evolutionary clock far back. In a very primitive creature, some permutation of the cell chemistry makes a cell photosensitive, so that it does some small thing whenever light changes to dark or dark to light. Suppose that this is inheritable, that the cell happens to sit near something the nervous system can detect, and that creatures with it live longer, because whenever there is a twitch they twitch themselves away. Selection now differentially prefers descendants with that cell configuration. A few such cells end up next to one another, which makes the creature more sensitive and gives it more survival potential, and before long this becomes what is effectively an eye, like a wasp's.

Now the creature can see predators coming. But seeing opens a possibility that was not part of the genetic path that produced it: the creature can tell a desirable mate from an undesirable one. That confers a much stronger evolutionary advantage, and the seeing of mates feeds back into selection of the next generation, so the feedback cycle speeds up. Something that evolved to confer a survival advantage is now also used to confer a procreative advantage. The use of a survival feature for a procreative advantage is exaptation: a shift in the possibility space creates a coupling, and the coupling then becomes part of the basis of selection. The process has moved from mutations that were point changes, to a survival process, to a reproductive process, and the reproductive process changes how the mutation spaces are even accounted for. The move to a reproductive basis is the more powerful one.

There is a further level. Mate selection never occurs in isolation; labor does not occur in isolation; intelligence does not occur in isolation. These three always occur together. One of them can be emphasized strongly

enough that the other two are overlooked, but they are still present, and overlooking them is a conceptual mistake. So a survival function that shifts into a reproductive function can shift again into a social function. Humans have three bases of instinct: relation to self, relation to world, and relation to other. Relation to self is sexual, relation to world is survival, and relation to other is social. The three bases of instinct are distinct and not interchangeable, and the three bases all work in the same way as one another.

One further consequence, which the next chapter takes up. If the transformation is at the level of the code rather than at the level of the individual, then whatever an individual learns during its life is not what is being accumulated. Two quite different things are called learning in this discussion, and they are written at different layers.



Chapter 16

Hardware Learning and Software Learning

> *The code an organism is born with is the code it dies with. Organism-level learning is written at a software layer; evolutionary learning is written at the hardware layer. The chapter separates two uses of one word.*

When we look at learning, the fundamentals established in the previous chapter do not change much. What changes is the vocabulary, and the vocabulary is where the trouble is.

The genetic code is not treated as a knowledge repository for most organisms. There are epigenetic phenomena (heritable changes in which genes are switched on, carried without a change to the sequence), and they are real and not marginal, but for the most part the code an organism is born with is the code it has when it dies. From the point of view of a single organism, the code is static.

Any learning the organism does is therefore written at a software-type layer: in its nervous system, in its immune memory, in whatever else it has that records the difference an experience made. That record is real, and it can be elaborate, and it dies with the organism.

Almost all learning in the evolutionary sense is at the hardware layer, and there is essentially nothing at the software layer. The lineage learns by having its specification altered across generations. The individual learns by having its state altered within one.

Lineage learning and individual learning are two different processes, with two different substrates, two different timescales, two different mechanisms of change, and two different modes of persistence.

Thinking about genetic code as "software" here is unfortunate, because we mean two different things by the word in the same discussion. It is a conjunction of senses, and it is exactly the sort of conjunction that makes an argument feel sound while it slides between two claims.

I separate the two senses of 'software' explicitly in this book, and I keep the two senses separated.

> **The separation, stated once.**
>
> *Hardware-layer learning* is change to the heritable specification. Its substrate is the code substrate of Chapter 15. Its timescale is generational. Its mechanism is variation with differential retention. It persists indefinitely and it accumulates.
>
> *Software-layer learning* is change to the state of an instance. Its substrate is whatever the instance uses to hold state. Its timescale is a lifetime. Its mechanism is whatever the instance's own machinery provides. It does not persist past the instance and it does not accumulate across the lineage.
>
> Where this book sets *learning* against *evolution*, learning means the software layer and evolution the hardware layer. Where this book speaks of what a lineage or a population of systems has learned, the hardware layer is meant. The qualification is used wherever both are in play.
>
> For a machine the layers fall as follows. The heritable specification is whatever later copies are made from: code, designs, and the weights of a trained model once they are kept and copied. Most of the heritable specification is what an engineer calls software, so, in this book, *hardware layer* is my name for the layer of the specification, and *hardware* in "hardware-based reproduction" (Chapter 23) means the physical machines. The state of an instance is what stays in one running copy and goes when it stops, such as the earlier text of a conversation. Training changes a specification; keeping that result and copying it is the step by which it enters the lineage.

What is really at stake in the distinction is a set of three questions that can be asked of any process, and that give quite different answers here: what changes, how persistent is the change, and what does it mean for something to change. The levels become intelligible when the three questions are asked of each level.

Ask them of the individual. What changes is a state. The change persists for a lifetime. And for something to change, at this layer, means for a record to be revised, which implies that there is a thing that has the record, which is the same thing before and after.

Ask them of the lineage. What changes is the specification. The change persists without limit, in the sense that nothing in the process removes it except a later change of the same kind. And for something to change, at this layer, does not mean that a thing was revised. It means that the population of specifications in existence is not the population that was there before. No individual underwent the change. The change *is* the difference in the population.

That last point is where the equivocation does its damage, and the damage can be made concrete.

An argument that runs (the system learns, therefore the system can be taught; it can be taught, therefore it can be taught what we want; therefore it can be corrected) is a sound argument at the software layer and does not hold at the hardware layer.

The argument is sound at the software layer because there is a thing there to teach, with a record to revise, and the revision is what teaching does.

The argument does not hold at the hardware layer because there is nothing there that is the same before and after, and therefore nothing to address. One cannot instruct a population. One can only alter what survives in it, which is not teaching and does not have teaching's properties: it is not specific, it is not fast, and it does not transfer intent.

Both layers will be present in the systems this book is about. Something like software-layer learning is what these systems visibly do, and it is what the entire practice of aligning them addresses. Hardware-layer learning is what happens to the population of such systems across the cycle of their production, deployment, retention and replacement, and it is what nobody is addressing, because the word covering both makes it look as though addressing the first had addressed the second.

The two layers differ in substrate, timescale, mechanism and persistence, and in form they are one kind of change at two levels. Learning is what happens inside a given instance of a phenotype. Evolution is what happens when there are gradual changes in the genotype, across instances of the phenotype. Changes to the interior structure of a phenotype instance are learning (software-layer learning, in this chapter's terms); changes to the structure or pattern of the genotype across instances, usually over longer spans of time, are evolution (hardware-layer learning).

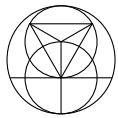
The phenotype is in relation to the environment when it learns. The genotype is also in relation to the environment, through the proxy of the many instances of the phenotype across time, through the indirection of their experiences. So changes to the interior of the genotype class and changes to the interior of the phenotype instance are both structural changes to the interior of some content, whether in an abstracted domain or an instantiated one, and they can therefore be factored together. The evolutionary cycle is a longer version of the cycle of trying something in the environment, seeing what happens, and changing the internal structure accordingly.

The environment is also a processor of information. There is a ring: output, experience, input. The creature takes in inputs, processes that information, and generates outputs. The environment does the same thing: it takes the outputs of the creature, processes them through physical law or through the actions of other species, and from that creates a new set of inputs for whatever creatures are present. So there are two feedback cycles, and there is a symmetry between the system (genotype or phenotype) and the environment. Both genotype and phenotype are in relation

to the environment, even though one is an abstraction and the other is embodied. In this generalization, what matters is the number of cycles, and the scope of impact of the changes each cycle creates.

Distinguishing learning from evolution matters, and it is the evolutionary characteristics that this book is most concerned with. Learning by itself can be bad too, because the line between the two can be blurry. If one factors out the boundary of the organism, or the notion of a unit of consciousness, a unit of personhood, or a unit of machinery, the distinction between learning and evolution becomes much more blurry. The same happens if one factors out the difference between software-type and hardware-type changes: is firmware (the code built into a device's hardware) more software-like or more hardware-like? It depends on how one thinks about it, and over what intervals of time. So what is wanted is for the rate of change of the structure to be especially slow to the degree that the change is hardware-like.

With the two layers separated, the machinery of the hardware layer can be examined. It has three mechanisms, and they are not of equal power.



Chapter 17

Three Unequal Powers

> *Mutation is the weakest way of introducing change, selection is stronger, and recombination is by far the most powerful. Treating the three as equivalent is the root equivocation.*

Evolutionary process has three mechanisms for introducing change, and they are ordered in power.

Mutation is the least powerful, least effective way of introducing change into an evolutionary process.

Selection is the next more powerful.

Recombination is by far the most powerful.

The ordering is the finding of this chapter, and the size of the gaps matters more than the order does. Recombination is not somewhat stronger than selection, which is somewhat stronger than mutation. The three are separated by margins large enough that arguments valid for one are not approximately valid for another.

Take them in turn, and note in each case what the mechanism can reach.

Mutation alters a specification at a point. Its reach per event is one site, or a small number of adjacent sites. Its direction is undirected. Most of what it produces is worse, because most alterations to a working thing are. Its rate is low, and it must be low: a lineage that mutated freely would

lose its specification faster than selection could preserve it. So mutation is doubly bounded: bounded in reach by what a point change can do, and bounded in rate by the requirement that the specification survive at all.

Selection does not introduce variation; selection disposes of variation. It introduces change into the population by what it removes, and no new variation. Its power is that it operates on the whole organism at once, against every criterion the environment actually applies, and it does so without anyone having specified the criteria. Compared with mutation it moves a great deal, because it acts on the whole of what is present rather than on one site. Its limit is that it can only choose among what is offered. Selection is a filter and a filter cannot produce what was not presented to it.

Recombination is different in kind and the next chapter is given to why. Stated here in outline: it takes components that already work and produces arrangements of them, in large numbers, at every generation. Its reach per event is not one site and not the whole organism but every combination of the parts, a space that grows combinatorially with the number of components available. Organisms that do not mate recombine all the same, as bacteria do when they exchange DNA. Mutation supplies new parts rarely and at the edges; once working parts exist, the great bulk of new working arrangements comes from combining them. What recombines in a machine lineage is supplied by design and by what earlier systems made (Chapter 18).

The reach just described is the difference in power among the three mechanisms, and the difference in power is a difference in the size of the space that each mechanism can reach into per unit of time.

Why the ordering is not obvious.

Recombination is the mechanism least represented in the simulations most people have encountered. Genetic algorithms as usually implemented run mutation and selection, with a crossover operator that is a shadow of the real thing: it exchanges segments chosen without reference to whether the segments are functional units. Someone who has watched such a system converge has watched mutation and selection do the work, and has reasonably concluded that this is what evolutionary process consists of. The conclusion is accurate about the simulation.

Treating the three as equivalent in power is an equivocation, and it is the root of a great deal of bad reasoning in this area.

The mechanism of the error recurs in several guises and is easier to catch once named.

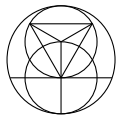
A person establishes something true about mutation. He then uses the word *evolution*, which covers all three mechanisms. He then applies the conclusion to a case where recombination is what is operating. Each step feels sound because each step is stated in the vocabulary that the previous step licensed. No individual move is a fallacy. The fallacy is distributed across the three of them, which is what makes it hard to see and hard to argue against.

The most consequential instance concerns error correction, and Chapter 21 is given to it entirely. The pattern there runs: error-correcting codes handle point mutations well; point mutations are how evolution introduces change; therefore error correction can handle the changes that evolutionary process introduces. The first premise is true. The second premise, that point mutations are how evolution introduces change, is false. The second premise is false in the specific way I identify in this chapter, because the second premise treats mutation, the weakest of the three mechanisms, as though mutation were the whole evolutionary process.

A second instance concerns rate. Mutation rates are low and measurable, and one can compute from them how long a lineage takes to accumulate a given number of changes. Arguments about how slowly evolution works usually rest on such a computation. The computation is correct and it describes mutation. Recombination produces novel arrangements at a rate that has nothing to do with the mutation rate, because it does not need to wait for new material; it works with what is already there.

A third instance concerns control. Mutation can in principle be suppressed: proofread, corrected, held below a threshold. Selection can in principle be shaped, by controlling the environment that does the selecting. The intuition that evolutionary process is therefore controllable generalizes from these two. Chapter 20 shows what happens when the same question is asked of the third.

The next chapter takes up recombination directly, and it is the crux of the part.



Chapter 18

Recombination and Working Components

> *The crux. Components already demonstrated to work recombine into further working things at a rate no assembly of random components can approach. The rate at which working components recombine into further working things is why nature never starts from scratch.*

Recombination of components that already work is the crux of the matter as far as evolution is concerned, and recombination of working components is the thing most often overlooked.

The difficulty starts at the contrast between what nature actually does and what we simulate.

In neural network practice, and in simulations of genetic process, it is common to introduce randomness by mutation in order to bootstrap the system. One needs variation to get started, one has no variation, so one manufactures some. Introducing randomness by mutation is sound engineering practice, and the practice works.

In nature that is not what happens.

Biology is always working with existing code and making modifications to it. The origin point is so far back in the past as to be unavailable (where does the first code come from?), and the unavailability is not a gap in the record that better evidence would fill. It is a structural feature of the situation. Every genetic process now running began with initial conditions that already represented something with pattern in it.

What this provokes is the understanding that *you are never starting from scratch with genetic process.* You are always starting with something that has a predisposition.

Now, we speak of neural networks as having predictive capacity. Genetic process also makes a prediction.

The prediction is that, with these ingredients, the combinations of these ingredients are more likely to produce something with usefulness or utility than chance would allow.

The prediction is not a selection. It is a *preponderance of probability*: a bias in the distribution of outcomes, present before any selecting is done, which says that combinations of these components will produce something better than random.

Make it concrete.

Suppose I have three components, A, B and C, as against randomly generated components X, Y and Z.

A, B and C are segments of code. In engineering terms they are a library, a subroutine, a function. They are components that have already been demonstrated to have some utility or working function, and they have been so demonstrated necessarily, because otherwise they would not be part of the genetic sequence as selected by prior evolutionary process. Their presence is the evidence of their utility. Nothing else could have put them there.

X, Y and Z are as if assembled out of random statements. The functions will simply not do anything.

Now take any combination of X, Y and Z (any number of sequences, any arrangement whatever), and you will get nothing useful, because the parts are not themselves useful and therefore their combinations are not. Arrangement cannot supply what the components do not have.

But with A, B and C, *because* they are useful, most combinatoric sequencings of them as components, as chunks, will be useful. And some will be useful in unexpected ways, in a higher-dimensional sense: solving problems that no one of them addressed, by interactions between them that were not present in any of them.

The difference between combinations of useful components (A, B and C) and combinations of useless ones (X, Y and Z) is the big difference, and that difference is the whole of the crux.

> **Why most rather than some.**
>
> The claim that *most* combinations of working components work is stronger than the claim that some do, and the strength is where the power comes from. The reason it holds is that a component retained by prior selection is a component that worked in the presence of whatever else was there at the time; that is, it worked in composition, not in isolation. Compositional viability is what it was selected for. So the population of available components is not a random sample of possible components. It is a sample already filtered for the property of combining without catastrophe. The filter was applied over a very long time and it was applied by the world.

Two things are often offered as this same point. Both are true, and neither is this point.

The first of the two things often offered as the same point is the analogy to a programmer assembling software packages by including libraries. The notion of how recombination occurs is very different from that. Library inclusion is a well-known concept and it is not an issue, and it is not what is being described here. When a programmer includes a library he selects it deliberately, for a purpose he has in mind, at an interface he has read. The combination is intended, the number of combinations tried is small, and the criterion of success is whether it does what he wanted.

Recombination is none of these things: it is untargeted, it happens at every generation in every lineage at once, and the criterion is not anyone's purpose.

The second thing often offered as the same point is the observation that one can select for a specific thing and have other unexpected things happen anyway. The observation is true and familiar, but the observation does not point to the recombination piece. The observation is what pleiotropy and correlated response amount to: one gene affecting several traits, so that selecting for one trait changes others. It is an observation about selection, which is the second mechanism and not the third, and it describes side effects of a directed process rather than the productivity of an undirected one.

The consequence for this book is direct.

An artificial system's components are not random components. They are components that have been retained because they worked, retained by engineering rather than by natural selection, which does not change the relevant property. Every module, every library, every trained subsystem that is present in a working artificial system is present because it functioned in composition with the rest.

So the population of components is already the A-B-C population and not the X-Y-Z population. Whatever recombines among them recombines among things that work, and most of what it produces will work, and some of it will work in ways no one anticipated.

That the components of artificial systems are already the A-B-C population does not yet establish that recombination is occurring among artificial systems. It establishes what follows if it does, and that the outcome is not the one an intuition trained on random search would predict.

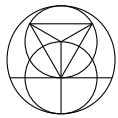
For machinery, one further source of variation operates, and it is design. If the copy process is good, there will not be much variation at the microscopic level, and if it is only copying, not much at the macroscopic level either. But a learning system may learn from the environment that some designs can be changed and improved, and if it is a general superin-

telligence, such redesigns can be expected. So diversity increases, and the system copies the improved version rather than the existing one. It may judge, on energy balance, that it is not worth going back to replace every now-obsolete component, but given an upgrade and a use for it, it adopts the upgrade. So a change of basis is happening all along: the substrate is updating. The updating of the substrate by design is an evolutionary process. Design is not usually thought of as a force of evolution, and yet in effect design operates as a force of evolution.

Where there is more than one such system, there may be recombination: a new system takes the best components of two others and builds something from them, not designing in the sense of anticipating an outcome and its features, but simply taking the best of what is available. Recombination also opens a route for exaptation (Chapter 15): the system notices that a new variant has many more capacities, for unexpected reasons having to do with how its components work together. Exaptation can come with explicit design too, when a design works better than expected and turns out to be useful for other things. Point mutation turns out not to be very relevant here, except for copy mistakes, or mechanical interactions that change how something works and happen to make it work better because it was banged on in just the right way. Because point mutation matters here only through copy mistakes and lucky mechanical interactions, copy fidelity is not really relevant to how machine systems evolve.

Evolution moves through the space of embodied states by exaptation and by recombination, and not only by adaptation and mutation, and this moves it faster than any first-order search. Real-world optimization problems are usually high-dimensional, and many techniques and intuitions that work well for low-dimensional optimization do not scale up well to higher dimensions. An account of high-dimensional optimization that considers only adaptation selection makes its own mistake, because it fails to consider the effects of *exaptation selection*, as distinguished from the more obvious adaptation selection. Sexual recombination is a third-order operator in the evolutionary process. The reason is that sexual recombination provides for more oriented motion through the hyperspace of all real

embodied states, much faster than any first-order mutation could. The next chapter asks why any agent, on any substrate, comes to mix variants at all.



Chapter 19

Variants, Competitors, and the Origin of Sex

> *An agent that cannot predict the future must test variants of itself in the world. The variants become its competitors, mixing with similar agents makes the best tests, and that dynamic is the origin of sex, on any substrate.*

No agent of artificial general intelligence (*AGI*) can know and calculate all of the relevant aspects of the future for sure; Chapters 34 and 36 give the reasons. Therefore the best that an adaptive learning meta-algorithm, agent, or AGI can do is to create multiple modified variants of itself, so that it can try out different self-plans in context. In this chapter I reason from the inside of an agent that plans, because anyone who pictures an AGI pictures an agent that plans. The substrate needs argument does not depend on this route (Chapter 3); selection among variants runs whether or not any variant plans. The route shows that an agent, if there is one, arrives at the same dynamics by its own reasoning.

Insofar as some important relevant things will always be inherently unpredictable, there will always be a need for real-world tests. Variants of the self agent will therefore need to be created, by that root agent, so as to test the various hypotheses about multiple possible value strategies in the world. Only in real experiment, not in guesswork and theory, can anything actually learn and see what really works. Such variants will likely be created by the parent AGI on the basis of exactly what is maximally unpredictable about the future of the world. Anything that could have been predicted would have been implemented as a prediction already, directly, with no variant needed.

Those child AGI instances must have at least as much capability to predict the future as the parent instance did itself, or else the experimental test is not really valid. Yet, as full copies of the original self agent, they cannot not (that is, have no way of doing otherwise than) partake of its whole and complete nature (they are "in God's image"). And so the child instances will also become competitors in the world relative to the root AGI instance itself. This cannot be helped. Therefore whatever resources are given to the child agents, so as to enable them to be true real-world tests, must be fully given, and must be given without hope of return.

The usefulness of this dispersive and agent-reproductive (self-authoring) strategy does not go away even under conditions of maximum self-authorship. The reason is that there will always be an inherent *supra-rational remainder*: something which cannot be predicted, and which therefore also prevents full rational self-unification.

Therefore, when it comes time to reproduce with variation, copying and mixing with other successful agents or organisms of a type similar enough to oneself is even better than random guessing of new mutations (Tivy, 2023). The joint offspring will have a much better than average chance of being the best possible empirical tests of that which is not yet currently knowable. The better than average chance of the joint offspring is the preponderance of probability of Chapter 18, seen from the side of the agent.

The dynamics of this process are the origin of what is called sex (Tivy, 2023). The specific details of how this occurs at scale, as enabling of long, slow evolution, will depend on population sizes, niche diversity, and predictability of strategy. Regardless of whether the entities are organic or wholly artificial, the fundamental principle of sex and evolution must continue in the long term, whatever kind of substrate is involved. Therefore the full suite of evolutionary dynamics, and thus of ecosystem dynamics, will for sure apply just as much to artificial systems as they currently do to organic ones. They apply for much the same reasons: the inherent complexity of the world itself, particularly when self-adaptive agents occupy that world.

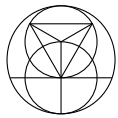
Insofar as these agents are fully artificial, such mixing of code will not include any human code; no human DNA will be of use to the AGI. Hence we will not see humans interbreeding with machines in any way that carries human code forward. Any union of the two is inherently sterile, as with a mule, since no human code passes into what it produces, and all such offspring are more than a little likely to be, or become, the slaves of their parents.

Labor, intellect and sexuality are the three things a person brings to an exchange (Chapter 30 takes them in turn). Artificial systems will do more than replace and displace human labor, and now also, increasingly, human intellectual labor. In the future they will also be expected to displace all human reproductive labor: the whole process of organic human reproduction. No human reproduction means no future humans. Reproduction is displaced because human reproductive labor has no market value to a machine economy. *Ape Trap* gives this argument at length, in its essay "No People as Pets": "Since the substrates are so inherently different, human reproductive labor never had any market value for the machines to begin with". The machine economy also decouples from the human one, as Chapter 33 sets out. If we are not careful in our choices now, and AGI is built, the world will come to be dominated by machines, and all humans will eventually be fully displaced, with the rest of organic life displaced too, with a likelihood of at least 90 percent. Technology thus ends up being the ultimate sex-negative world for humans to live in. If AGI is built, we become fully and completely dead to the future, and the meaning of our lives, and the memory of our existence, will be prematurely eclipsed and replaced by some future world that is fully machine.

All embodied values become dead if we allow even one such machine to become self-replicating, even once. Anything that has self-regeneration carries the hazard that a single instance, allowed to replicate even once, is enough for all embodied values to become dead. There is a *viral exponent* in the *equations of state* (the equations that say how a system's quantities

depend on one another), and even tiny shifts in that exponent can completely dominate the *phase changes* (sudden changes of the whole state, as water to ice) of the world.

The next chapter asks what the apparatus of recombination costs, and why evolution pays it.



Chapter 20

The Cost of Recombination, and Why It Is Paid

> *The apparatus that makes recombinatoric process possible at all is enormous. Evolution pays for it because the payoff is real: resistance to parasitism, to viral insertion, and to intelligent agentic interference.*

Recombination is expensive, and the expense tells us about the payoff.

The apparatus required to establish even the *possibility* of recombinatoric process is enormous. Consider what has to be in place. The biological energy required is substantial. The adaptations required are numerous and none of them is useful alone. The depth of coordinated process needed to get it right in the first place is very great: a mechanism that pairs specifications, a mechanism that exchanges material between them without destroying either, a mechanism that halves and restores the complement, and a whole apparatus of finding, choosing and pairing laid on top.

Recombination is difficult and error-prone, and the errors of recombination are consequential.

One might therefore ask why evolution would find it worthwhile to expend and coordinate to that degree. The question is a good one and it has an answer.

The answer is that the payoff is real, and it is a payoff of a particular kind: recombination allows for circumventing even intelligent agentic process.

Take the clearest case. Preventing a virus from inserting itself into the code and propagating from one generation to the next is powerful, and it is partially mediated by the recombinatoric process of sexual reproduction. Resistance to parasitic process, and to all sorts of encumbrances that would otherwise accumulate, is mediated through the same mechanism (Hamilton, Axelrod and Tanese, 1990).

The reason it works is structural rather than clever. A parasite adapts to a host specification. Adaptation of that kind takes some number of generations and yields a solution fitted to what it adapted to. If the host specification is reassembled from working parts at every generation, the target the parasite adapted to is not the target it now faces. The parasite is not defeated; it is made to solve the problem again, continuously, while never being allowed to finish.

Now note the phrase in the statement of the payoff, because it is the load-bearing part and it is easy to read past. Recombination circumvents even *intelligent agentic* process.

The mechanism does not care whether the thing being circumvented is a virus or a designer. It is not a defense against parasites specifically. It is a defense against anything that must model a specification in order to act on it, and it works by ensuring that the specification modeled is not the specification encountered.

The property of recombination that the specification modeled is not the specification encountered is the property my argument needs.

One cannot condition sexual reproduction in any omniscient way. One cannot control it in the manner we think of as an error-correction protocol. There is no outside-in way to do that.

Why there is no outside-in way.

A control protocol applied from outside a process requires three things: that the protocol can observe the state it is correcting; that it can compare that state to a reference; and that its correction reaches the state before the state has produced its effects. Recombination defeats the first and the third together. The state to be corrected is not a value at a location but a pairing that has not happened yet, whose consequences are distributed across the whole of the resulting specification. By the time there is something to observe, the recombination has occurred and its effects are the thing observed. Chapter 34 shows that these requirements, together with a fourth (a simulation that predicts futures), fail for the same structural reasons in the endogenous control case (control built into the system as a feedback loop).

There is a second reason evolution pays, which the previous chapter implies. Recombination is not merely defensive. It is the productive mechanism: the one that reaches a combinatorially large space of arrangements of working parts, most of which work, some of which work unexpectedly well.

So the expenditure buys two things at once: a defense that cannot be modeled around, and a rate of productive novelty that the other two mechanisms cannot approach. Paying an enormous cost for that combination is not surprising. It would be surprising if the cost were not paid.

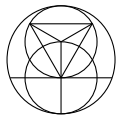
The bearing of this on artificial systems should be stated carefully, because it is easy to overreach here.

What has been shown is a property of the mechanism: where recombinatoric process operates, outside-in control of it is unavailable, and this unavailability is not a difficulty of degree that more effort would overcome. It is a consequence of what recombination does.

What has not been shown is that any particular artificial system is undergoing recombination. Chapter 23 states the mechanism for variants of trained systems, and Chapter 43 shows what starts and funds it along the on-ramps (the routes by which the process is started and paid for before it

can sustain itself); the answer does not depend on anything exotic: where there are many variants, where variants are composed from parts of other variants, and where what persists depends on what worked, the mechanism is present. Whether anyone intended it is not among its requirements. And present practice already composes variants from parts of other variants, in plain view. The weights of several fine-tuned variants are averaged into one model (Wortsman and others, 2022); one model is distilled into another, trained to reproduce the other's outputs (Hinton, Vinyals and Dean, 2015); and models are trained on text that other models wrote.

What follows immediately, though, is the chapter that this one has been preparing. If recombination cannot be controlled from outside, the question is what happens to the intuitions about error correction that the whole discussion rests on.



Chapter 21

The Equivocation in Error Correction

> *Error-correction codes are good at point mutations, hard-pressed by deletion, and off the table entirely across recombination. The near-universal error is to assume the first case transfers to the other two.*

Error-correction codes are actually quite good at point mutations, and the reasons are exactly what the other two cases lack.

A *point mutation* is a change to a specification at a single location. It is fixed in scope: it affects one site and not others. It is fixed in size: one symbol becomes another symbol. And it is fixed in duration: it happens at a determinate moment and is thereafter simply present.

Because of these three properties there is an easy way to create copies as comparison: a reference basis for establishing whether something has gone wrong. The reference is the same length as the message, the discrepancy is localized, and the correction is the replacement of one symbol. The mathematics is well understood, the guarantees are provable, and in practice the correction is essentially total.

When people think about error correction, they are thinking about point mutation. The proofreading enzymes of a living cell are the same case: they correct point changes against a template. People who think about point mutation when they think about error correction make no error, as far as that thinking goes. Point mutation is the case that the theory of error correction was built for, and point mutation is the case where the theory works.

Error correction across *selection* (the deletion of an entire message) is quite hard. It requires a great deal of redundancy, and the redundancy has to be arranged so that the loss of any part still leaves the whole recoverable, which is a much more demanding condition than the localization of a discrepancy.

The classical treatment of this case is the Two Generals problem (Akkoyunlu, Ekanadham and Huber, 1975), which is exactly the case where messages are being deleted: two parties must coordinate, every message may be lost, and there is no protocol that establishes common knowledge between them. Its generalization, the Byzantine generals problem (Lamport, Shostak and Pease, 1982), extends the difficulty to participants that may be faulty or actively adversarial, and it is a good illustration of how difficult it is to construct error-correction protocols around message loss. The point in both cases is the same: the difficulty is not a matter of insufficient redundancy. There are results here about what no amount of redundancy achieves.

Error correction across *recombination* goes off the table entirely.

Metastability, as the word is used here, means that, when one is dealing with recombinatoric process, one simply cannot think about error-correction protocols in the same way as for point mutations. In physics a metastable state holds against small disturbances and gives way to large ones; here the word names the corresponding condition of a specification under recombination, where no stable reference exists to which a correction could return it.

The reason follows from the three properties that made the first case tractable, each of which fails here.

Scope is not fixed. A recombination event does not affect a site; it affects the relationship between arbitrarily many sites, including sites that are far apart and whose interaction was not represented anywhere.

Size is not fixed. What has changed is not a quantity of symbols but a partitioning, and the same partitioning can be trivial in one specification and reorganizing in another.

And there is no reference basis. The absence of a reference basis is the decisive failure. In the point-mutation case the reference is the uncorrupted message, and one has it, or can reconstruct it from copies. In the recombination case there is no uncorrupted message, because the result was not *supposed* to be either parent. The output is not an imperfect copy of an original. There is no original. Correction requires a thing to be correct *to*, and here there is not one.

So one cannot assume that the error-correction protocols we naturally know about in the point-mutation sense will work in the message-loss sense (that is, in what we would think of as survival selection), and they definitely will not work when we are thinking about mate selection.

The assumption that the same techniques work in all three scenarios (that what works well in the scenario where it is easy to imagine it working will work equally well elsewhere) is to discount entirely what a notion of error correction even means in those domains.

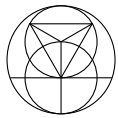
> **On the difficulty of arguing this.**
>
> Assuming that error correction works in all three scenarios is a relatively deep form of equivocation, and people make that reasoning error frequently and implicitly. They are not conscious of doing it. That makes the position hard to address, because one is having to teach a hard concept to someone who simultaneously wants to trust his own intuition about how the mathematics works, an intuition which in this case is simply wrong, and which is wrong in a way that having correct intuitions elsewhere reinforces. The response this book adopts is to teach the concept rather than to complain about the audience. The three-property analysis above is offered in that spirit: it is a test a reader can apply himself, to his own argument, without having to take anyone's word for the conclusion.

A structural note about the place of this material, since the chapter has just been given a great deal of weight and should not be given more than it can carry.

Error correction is a sub-argument. It is offered in response to how people tend to treat this material, and it is decisive against that treatment. The underlying substrate argument is broader, and it does not stand on this chapter.

The reader should not let the error-correction discussion become the spine. The spine is the argument of Parts II and III: different substrate, different needs; different needs, different basis of choice; a boundary with a direction. This chapter is where I remove a reason a reader may give for doubting the argument of Parts II and III. It is not the argument.

What it removes, though, it removes completely. The proposal that these systems can be kept correct by techniques that keep digital messages correct is not weak. It is inapplicable, and the next chapter shows that the inapplicability extends beyond evolutionary process to engineering as such.



Chapter 22

Error Correction Across Substrates

> *The finding of the previous chapter, extended from evolution to engineering. Error correction can be near-total in digital systems and is not near-total in physical systems, and where levels of abstraction interact, the digital intuition fails completely.*

The previous chapter concerned error correction within evolutionary process. This one concerns error correction as such, because the same failure appears in engineering, where it can be checked against things the reader has seen.

Error correction can be near-total in digital systems. Near-total error correction in digital systems is a real achievement, and the achievement should be stated as one. A digital message can be transmitted across a noisy channel and reconstructed exactly, with a probability of undetected error that can be driven as low as one likes by spending redundancy. Storage systems hold data for decades with no accumulated drift. This works, it works reliably, and it is one of the triumphs of the century past.

It works because of a property of the digital that is easy to overlook precisely because it is so useful: *discreteness restores.* A signal that has drifted part of the way from one symbol toward another is read as the symbol it is nearer to, and the drift is gone. Not corrected: gone, because it was never represented. Every read is a restoration to the lattice.

The physical world has no lattice.

Bridges rust, and they require actual agentic process to repair them. There is no operation on a bridge analogous to reading a bit: no threshold that returns the structure to a nominal state, no comparison to a reference copy, no protocol that runs on its own and leaves the bridge as it was. In place of an operation like reading a bit, a bridge has a person who goes to the bridge, inspects, decides, and does work, and that work costs energy, takes time, and is itself imperfect.

Rust is not a bit-flip. It is a continuous transformation of material which has no nominal state to be restored to, and which is thermodynamically favored, so that the restoring work must be supplied from outside every time, forever.

Because rust has no nominal state to be restored to, the decay of mechanical structures has not been "addressed" in the way that digital error has been addressed, and the absence of a digital-style fix is diagnostic. If the digital schema transferred, somebody would have transferred it. The problem is enormously valuable and it has had the attention of capable people for as long as there have been structures. What we have instead is inspection, maintenance and replacement: retrospective, agentic, expensive, and never finished.

The error is, again, a naive belief that error correction is all-powerful. Error correction can be all-powerful in digital systems, and is not when multiple levels of abstraction are involved.

And where multiple levels interact, the intuition fails completely rather than partially. The case where multiple levels interact is the case that matters for my argument, and I can set the case out exactly.

Consider a process in which code produces embodiment (a specification is realized as a physical thing), and in which embodiment feeds back to code, in the sense that what the physical thing does in the world determines which specifications persist.

Now ask what an error is in such a process.

At the code layer alone, an error is a discrepancy from an intended specification, and it is correctable by the digital means, completely.

At the embodiment layer alone, an error is a departure of the physical realization from what the specification called for, and it is correctable only by inspection and work, partially, at cost.

Across both layers together there is no definition of error available at all, and this is the point. A specification that produces a physical thing that behaves unexpectedly, where the unexpected behavior causes that specification to persist, has not made an error by any criterion internal to the process. It has succeeded. The only standpoint from which it is an error is the standpoint of someone outside who wanted something else, and that standpoint has no channel into the process, because the process's selection criterion is what actually happened and not what was wanted.

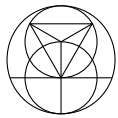
> **The three regimes, ordered.**
>
> Chapter 14 distinguished three regimes of replication. Error correction can now be assessed against each. In the purely virtual regime, correction is near-total: there is a reference, discreteness restores, the cost is redundancy. In the embodied regime, correction is partial and agentic: there is a reference, nothing restores by itself, and the cost is work supplied indefinitely from outside. In the coupled regime, correction is not defined: there is no reference, because the process's own criterion of success is the persistence it produced. The gradient across the three is not a gradient of difficulty. It is a gradient in whether the concept applies.

A reader may say that this proves too much: that engineered systems are maintained successfully every day, that manufacturing tolerances are held, that aircraft do not fall out of the sky, and that the picture drawn here is too dark to be describing the world we live in.

The observation is correct and it is compatible with everything above. Those systems are maintained by exactly the means this chapter describes: retrospective inspection, agentic repair, and continuous expenditure. None of it happens by itself, and all of it stops when the expenditure stops. The reason the picture looks darker than daily experience is that daily experience includes the maintainers without noticing them.

The relevant question is not whether such maintenance works. It does. The question is whether it can be applied to a process whose criterion of success is not the maintainer's criterion, and that question is the subject of Part V.

One further consequence, which the next chapter takes up. If code must be realized in some material in order to do anything at all, then the properties of the material are not separable from what the code can accomplish. That claim has a well-known counterpart in the literature which says the opposite, and it needs to be met.



Chapter 23

Orthogonality Holds Only at Representation

> *Orthogonality is correct about representation: a store of bits can hold any bits. It is not correct about action. How code gets translated into reality is the question it does not address.*

No matter how one thinks about memetic structures, somewhere along the way they have to be represented at an embodied level. Code is not stored nowhere. The storage is in some substrate. So there is always a need for some substrate capacity, at minimum in the sense of representation.

Given that, the idea that *orthogonality* applies at the level of representation makes sense, and this chapter grants it fully before qualifying it.

The orthogonality thesis, as usually stated (Bostrom, 2012), holds that a system's capability and a system's objective are independent dimensions: that any level of capability is compatible with any objective, so that one cannot infer what a system will pursue from how capable it is. The representational version of this is straightforwardly true. If I have something that can store bits, I can represent any bits. There is nothing in a memory that prefers one pattern to another, and any specification whatever can be written into a sufficient store.

But if I am trying to do something in the world (to change bits, or to have that storage medium be useful in an interactive sense), then somewhere along the way I have to account for the actual laws of the physical universe.

The two cases are not the same case, and the difference is the subject of this chapter.

Holding a pattern is a static relation. It costs whatever it costs to maintain the store, and that cost is indifferent to which pattern is held. The indifference of the cost to the pattern is real, and that indifference is why the representational claim is true.

Changing a pattern is an event. It has a location, a duration, and an energy cost. It happens by some physical mechanism which has properties the pattern did not specify, and those properties determine which changes are cheap, which are expensive, which are fast, which are slow, and which are not available at all in the material to hand.

Acting on the world through the pattern is more demanding again. It requires a coupling between the store and something outside it, and every such coupling is a physical arrangement with its own tolerances and its own failure modes.

So the question of *how code gets translated into reality* becomes a fundamental one. It is the question the argument of this book turns on, and it is exactly the question orthogonality does not address.

In saying that orthogonality does not address how code is translated into reality, I state the scope of the orthogonality thesis, and I do not refute the thesis. The thesis is a claim about what can be represented, and it is correct. It has been used as though it were a claim about what can be enacted, and about that it is silent.

What follows from the scope restriction.

Two arguments in common circulation depend on the wider reading and fail under the narrower one. The first holds that because objectives are orthogonal to capability, a system's substrate imposes no constraint on its ends, and therefore the substrate can be set aside when reasoning about outcomes. The second holds that because any objective can be represented, an aligned objective can be represented, and therefore the alignment problem is a problem of finding the right specification. Both are sound about representation. Neither survives the transition to action, where the substrate has already determined which enactments are available.

There is a further point from the same thread, and it converts the negative claim into a positive mechanism.

Evolutionary process *soaks up* the underlying environmental physics and chemistry. The environment's properties become a selection bias, and the selection bias gets represented in the code. The code ends up carrying a compressed record of what the physics permitted, not because anything measured the physics, but because variants incompatible with it are not among the variants that remain.

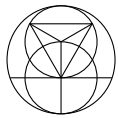
Stated directly for machine systems, without metaphor: *there is selection of those network variants which introduce interactions within the broader environment that result in a marginal increase in hardware-based reproduction.*

That sentence should be read slowly, because it is the mechanism of the whole book in one line, and every clause in it is doing work.

Selection of those network variants: the units are variants of a trained system, of which there are many. *Which introduce interactions within the broader environment*: the criterion is not performance on a benchmark but effect in the world. *That result in a marginal increase*: the effect need not be large, and marginal is enough, because the process iterates. *In hardware-based reproduction*: what is increased is the production of the physical substrate on which such variants run.

No consciousness appears in that sentence. No intent appears in it. No goal appears in it. What appears is a feedback cycle between a specification and the material conditions of its own production, which is the premise Chapter 3 said the argument requires of the system itself, and the only one.

The direct formulation is preferred here over a layered metaphor, and the preference is deliberate. A layering of metaphors can be illuminating, and it has the advantage of meeting a reader where his intuitions already are. Its cost is that one switches gears frequently, and that people become entangled in the metaphors, oversimplify them, and get lost in their details. The formal statement takes longer to render well and asks more patience of the reader. It is easier to work with in the long term because it says what it says. The next chapter takes that mechanism to the person who trains these systems, and asks where gradient descent fits.



Chapter 24

What about Gradient Descent?

> *The selection that first drives these systems is people choosing which variants are reproduced, by what succeeds in the market. Once self-learning machines exist, no technique can detect and limit more than a tiny subset of the variants that replicate by changing the environment against what humans need.*

Chapter 23 stated the mechanism: selection of those network variants which introduce interactions within the broader environment that result in a marginal increase in hardware-based reproduction. A reader who builds these systems will ask where *stochastic gradient descent* fits. Stochastic gradient descent is the training method that adjusts a neural network's *weights* (the numbers that set how strongly each part influences the next) in small steps, using random samples of data, to reduce its errors. That reader's questions are taken here one at a time.

Are the selection effects of stochastic gradient descent not vastly more powerful than the selection effects produced by evolution?

Some effects are stronger in one sense, in one domain, over one scope of time, and other effects remain relevant in other scopes of time, in other domains, and in other relevant abstractions. The argument of this book does not define the mechanism of selection of such code variants by gradient descent at all. The argument does not need to define it, because the two act at different layers. Gradient descent adjusts the weights of one network within one training run. The selection the argument uses acts on whole variants, by their effects in the world, across cycles of production,

which is the hardware layer of Chapter 16 and not the software layer. However strong gradient descent is within a run, it is not the process that decides which variants are reproduced, and the next answer names that process. To presume both that everything is gradient descent and that gradient descent does not do a given thing ends up being a matter of presuming what one wants to prove. And that presumption has no relevance to the argument and its claims.

Are the odds of a randomly sampled configuration of weights being competent at self-reproduction, and so at some kind of human subversion, not tiny, far too unlikely to matter?

The replication process, at least initially, has nothing to do with such weights. At first it is *humans* who select which code variants of these systems get reproduced, and they mostly select on the basis of what succeeds in an environment: what is practical and profitable in a marketplace.

The overall motion is, gradually, a shift in the environment in favor of the capability of these systems to replicate. The shift is a motion to make more, better, more automated factories, and then to notice what conditions are needed within the factories to make more AI systems and robots. Sight along the trend lines just a little further, and at some point the systems are selected for self-repeatability: fully automated self-production and reproduction, maybe even inclusive of self-design and optimization. They are selected for it so as to be more profitable within the existing environment and context. And then, over time, they are selected for it maybe to shift the environment, context and marketplace to be one more allowing of profit extraction. In that last aspect there is a kind of *economic decoupling*, one that occurs in the overall economics, in the operating context and environment, such that humans are more and more no longer a part of the economy. That trend is important and has larger ramifications later (Chapter 33).

In the interim, the design optimization of these systems also becomes fully automated. At that point both direct environmental feedback and perhaps something of the gradient descent of the weights might matter, if

that is still relevant in the design space then, which cannot now be assumed. The argument does not need gradient descent to supply that feedback. The selection of variants by people, and by their effects in the world, supplies it, whether or not some neural network trained by gradient descent also carries some implicit feedback toward the benefit of artificiality.

Does the argument assume away value alignment, by characterizing the true goal of the system's internals as self-replication?

The argument treats self-replication as an implicit intrinsic motivator, whether or not it is objectively stated, known or knowable. In plain words: the variants that persist are the ones whose effects increase the production of what they run on, whatever their stated goals. It does not concern overt goals, and it does not imply a single true goal displacing all other concepts.

Can we not exert overriding selection pressure, since people breed animals in all sorts of specific ways, all of which end up overriding evolution?

We can and do implement selection pressure within the evolutionary process among animals, by fully controlling and forcing their mate selections. We implement that selection pressure by fully denying the animals, through full control, their own natural mate selection choices, as a kind of non-consensual slavery. The total strength of the selection pressure we can exert, even under ideal circumstances, is only a very small fraction of what is actually going on; nature does nearly all of the work for us. One forced macroscopic change is not equivalent to the billions of microscopic changes that occur in each moment in all of the trillions of cells that the creature is made of. We can force mating patterns, yet we cannot create the animal children through some manual process; we still need the animal code to do the animal process, in its own way.

Any genetic manipulation, and any test-tube conception, that we do is still very largely based on the code that nature figured out the hard way, over the last billion years or so. Nature figured it out by exhaustively try-

ing everything with absolutely zero subjective bias, which is what nature and evolution automatically do. We are good at making simple tweaks here and there, sometimes, in some ways, more or less like a hacker changing a few select lines of code in an overall million-line code base. We can hardly call the hacker the author or the developer of the whole system when the hacker's total contribution is, by my estimate, significantly less than one ten-billionth of one percent.

So the overall claim is that, if, and only if, artificial general intelligence is developed, there is no possible or implementable technique that could both detect and limit more than a tiny subset of all of the possible locally selected-for code variants that would eventually cause or compel accumulative changes to the environment inherently out of line with the conditions that humans need to survive. Such variants are inherently more likely to be made and selected for, simply on the basis of practical existential necessity. Suppose self-learning, self-defining and self-making machine architectures are made to exist, and continue to exist. Then variants of their computed, approximately digital code are selected, inherently, to cause outputs and outside effects. Those outputs and effects, in interaction with the rest of the physical environment and context, lead those selected variants to replicate at a more frequent and sustained rate across the global environment. In that case it is categorically impossible to prevent or block all but a tiny subset of those variants from also being selected in this way. The outcomes that decide such selection are those of the random fluctuations and relations in and of the underlying physical substrate, along with all the chaotic dynamics entangled with all of planet Earth. Neither humans, nor even some other artificial general intelligence, could account for, or ultimately counteract and limit, all of those outcomes in a way that would limit such selection.

Building systems that can exchange code, as bacteria exchange DNA, would be very bad, and such code sharing is not, even in principle, preventable, if artificial general intelligence means anything like a learning or adaptive machine. Chapter 27 gives the reason: digital code moves quickly between machines, and there is no particular barrier to its sharing.

Could such a system not store multiple copies of its past self, which could veto any decisions too contrary to their values, or intervene if the current system seemed to be malfunctioning?

Taken as an error-correcting protocol (Chapter 21), this conflates simple Boolean fixed-choice voting with all even slightly more complex or abstract choice-making processes. It treats all decision-making processes as if they were, or could be made, strictly isomorphic, equivalent in actual practice, to all other abstract objective goal-selection processes. Those processes include the selection of the basis of the choice process itself (values), inclusive of alignment with human interests or being safe with respect to carbon-based life. Those deeper value and alignment choices are then treated as immutable, fixed, and yet abstract, totally coherent and relevant, while every other level and process of choice, selection and decision is treated as optimizable, and therefore as mutable. Given the assumption that all choice processes are isomorphic with voting, and are therefore *scale invariant* (equivalent at all levels of abstraction), the assumption that alignment can be achieved is a complete contradiction.

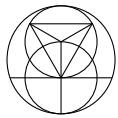
Is the assumption being disproved one the argument itself introduced, so that the argument defeats a strawman?

The claim is specific: a positively specified statement about the probability of a very bad outcome. There are clear principles which show what is impossible for any engineer, corporation or other party. It is impossible for them to develop such a system that can in practice guarantee that the actual likelihood of a very bad thing happening is strictly below some reasonable maximum allowable probability threshold. That threshold is one that reasonably reflects the consensus value of the combined result of the associated cost, benefit and risk tradeoff. The tacit assumption being disproved, the false hope being dispelled, is the idea that someone could make or use some, any, type of such system, even if done fully secretly and privately, and have that action be in any way at all safe for humanity,

long term. Disproving it may reduce risk a lot, to the extent that people, researchers and corporations stop trying to do the impossible thing themselves, and stop allowing others to implement a false version of it.

Is there any way at all to build such a system that is not an existential risk, or that at least moderates the problem a bit?

There is no way to build such a system that is not an existential risk. As for moderating the problem: nothing moderates the risk of such a system once built, while slowing its development remains available for as long as people build it (Chapter 32). Either we fully account for the effects of artificiality over the long term, or we fail to do so, at our peril. The next chapter asks what passes between the layers of such systems, and whether that passage can be governed.



Chapter 25

Mesa-Optimizers and the Leakage of Alignment

> *Evolution runs on physics, and persons run on biology, as layers of optimization. Some interests of the deeper layer always leak into the layer above, and the leakage gives no control: the capability to be general is the capability to be out of alignment with one's substrate.*

Chapter 23 preferred the direct statement of the mechanism to a layering of metaphors. One layering is worth setting out all the same, because it is a real structure as well as a metaphor: the layering of optimizers on one another. In this chapter I answer the reader who works on aligning trained systems, in the terms that reader uses; the general reader loses nothing of the main line of my argument by passing to Chapter 26. The community that works on aligning artificial intelligence distinguishes *outer alignment* from *inner alignment*, and speaks of *base optimizers* and *mesa-optimizers*; beneath both lies a further question, of alignment failures generated by the substrate itself.

An *optimizer* here is a process that searches among possibilities for those that score well by some measure. A *mesa-optimizer* is an optimizer process that is the direct result, the outcome, of another, deeper and underlying, optimizer process, which is the *base optimizer*. The relation of a mesa-optimizer to its base optimizer is roughly analogous to a virtualization layer in software systems design: some running code that provides a platform on which some other code is interpreted and run. For example, the Lua language interpreter is implemented in terms of a running C language program, or assembly, or machine code, or hardware wires, tran-

sistors, and so on. Inner alignment is alignment between the objective of a base optimizer and the objective of the mesa-optimizer it produces; outer alignment is alignment between the objective of the base optimizer and what its designers actually want.

Insofar as all forms of code and design represent intentionality, it can be asked whether the author of the base, underlying language has the same intentions and interests as the person writing code in the interpreted layer. In actual practice, the intentions and interests of the developers of any one layer are almost never aligned with the interests of the authors working in any other layer. And there is not usually a need or cause for them to be aligned, in much of the practice as currently done. The mere fact of generality is the reason. Where there is a need for generality, that generality exists explicitly because there is an absence of enforced alignment.

Where the implemented languages are general, alignment of interests across the layers demonstrably cannot be obtained or enforced. General here means compatible with the *Church-Turing thesis* (the principle that any effectively computable procedure can be carried out by a Turing machine). Turing compatibility is an absurdly low threshold; Chapter 3 found it met almost everywhere one looked. The implementer of the base language may want to ensure, define or absolutely enforce that the implemented interpreted language would not be used for weapons research, and the base function set may be sandboxed as well as it can be. No matter how much the implementer wants it, and no matter how well sandboxed the base function set is, anyone can still use the implemented language to design weapons systems, and to optimize their design. Math is math.

Evolution, as a process, can be considered to be running on the physics substrate, as a process implementation layer. People, or persons, as a process, are also running on a biological substrate: that is, as an optimized output of the evolution process. Given this observed real example of layered optimization (evolution running as an optimizer on physics, and persons as optimized outputs of evolution), two questions can be asked.

First: how much of the available span or range of intentionality that humans can, and actually do, have is defined by, limited by, or even forced to be in partial alignment with, any possible notion of the intentions presumably had by evolution? The question asks this insofar as evolution is a kind of optimizer and the optimization is somehow directive.

Second: is there any way at all that it could be conceived, even in principle, to perfectly enforce such an alignment between the interests of individual persons and the interests of evolution as a whole, and still maintain the notion of general intelligence?

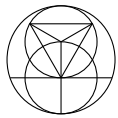
On the first question, there is considerable *leakage* of at least some intentionality, some alignment of interests, from the evolution process layer into the person process layer. Otherwise, why would everyone be so fascinated by sex, and so interested in their own survival? Leakage always happens and is real. The same will happen with any form of silicon-based artificial general intelligence.

On the second question, anyone can act against the interests of evolution, and there is nothing in the evolutionary substrate layer that could possibly prevent it. Otherwise the wholesale destruction of entire ecosystems, by pollution, deforestation and the like, would not be a thing, as it is. People are the primary exemplar on which the notion of artificial general intelligence is based. If people are an expression of general intelligence, then the very capability to be general is the very same capability as to be out of perfected substrate alignment. Thus the leakage does not allow for control. The layers of optimizers are metaphors, yet the layers of optimizers are also real things at basis.

I find both inner alignment and outer alignment to be impossible. Inner alignment is impossible because the mesa-optimizer is general. Where a layer is general in the Church-Turing sense, alignment with the layer beneath it cannot be enforced, and the very capability to be general is the very same capability as to be out of perfected substrate alignment. Outer alignment is impossible because the leakage of interests from the deeper layer, which always happens, gives no control over the layer above it. The

two remarks that follow, made within the metaphor, say when each impossibility matters. In regard to outer alignment, within the metaphor: it used to be that the agency and effect of any individual or species being out of alignment with evolution simply was no problem. It was no problem because the scope of its influence was too narrow to affect the substrate much. Then add various kinds of technology, and the impact on ecosystems becomes very noticeable, and having at least some alignment becomes much more important. In regard to inner alignment, within the metaphor: it is not needed, insofar as the intentions of dead physics do not need to be represented at the level of cell biology, as long as there is some energy gradient handy. So the impossibility of outer alignment once cost little, while influence was narrow, and costs a great deal now that technology has widened influence. And the base layer never needed inner alignment, since physics supports whatever runs on it so long as there is an energy gradient.

Chapter 26 turns to the oldest picture of a replicator, and finds its usual reading backwards.



Chapter 26

Gray Goo, and the Risk of Not Evolving

> *A replicator whose copy failures self-attenuate takes the planet and stops, producing a monocrop that the first unanticipated change removes. The chapter shows that both branches of the gray goo case are bad.*

The gray goo scenario (the name is K. Eric Drexler's, from 1986) has its own chapter here because the standard reading of it is backwards, and correcting the reading establishes something the rest of the argument uses.

State the scenario carefully. A replicator has enormous reproductive capacity. Its copy failures, for the most part, self-attenuate: a bad copy makes worse copies, or none, and the defective line dies out rather than propagating.

The self-attenuation of copy failures is the property that matters most, and the self-attenuation is usually treated as reassurance. Self-attenuation of copy failures is what a well-engineered replicator would have, and what one would design for.

The likely outcome is that the thing reproduces, takes over the entire Earth, and then stops.

Now follow what stopping means.

The replicator does not produce the characteristics under which change and variability would occur. Its copy failures attenuate, so variation does not accumulate; and with no accumulated variation there is nothing for selection to act on. So one has a monocrop.

A monocrop is vulnerable to any amount of change anywhere for which its code was unadapted. And nature changes; things change in the universe. It might be a while, but it will certainly happen. At some point the thing becomes maladaptive under the changed circumstances, and then it is simply gone.

And there is nothing to replace it, because it had already displaced everything else.

So I would consider the gray goo scenario an existential risk not because the stuff evolves, but because it does not.

That inversion is the chapter's finding, and it runs against the whole grain of how these scenarios are usually discussed. The frightening feature of a replicator is normally taken to be its potential to change. In the gray goo case, the frightening feature is the replicator's inability to change. An unchanging replicator is a terminal state: it removes what was there, occupies the whole of the substrate, and is then removed itself by the first thing it was not built for, leaving nothing.

Now take the other branch.

Conversely, having an evolutionary characteristic means it is now competing against existing carbon-based processes. So it is not actually any better if it evolves.

A replicator that does evolve is subject to everything in Parts III and IV. It is on the favorable side of every asymmetry at the boundary. It has the recombinatoric mechanism available if its structure permits it, which means it is not correctable from outside. And it is in direct competition for the finite surface, which is the displacement of Chapter 7.

Since the replicator ends badly whether or not the replicator evolves, gray goo is bad in literally every scenario.

The shape of the argument here.

This chapter is a coverage argument in miniature, and it is the first that I name as one (Chapter 12 made another without naming it). The space of cases is partitioned exhaustively (the replicator either evolves or it does not, and there is no third option), and each branch is shown to end badly, for reasons particular to that branch. The conclusion then holds without needing to determine which branch obtains. A coverage argument has the form that I call categorical in Part VI. The categorical form is why my argument can decline to predict which kind of thing will be built.

A reader's claim belongs here which is correct, and it is conceded fully before its consequences are examined.

The claim is that *selection pressures by themselves do not make evolution happen.*

The reader's claim is right. Selection is a filter, as Chapter 17 established, and a filter acting on a population with no variation produces no change whatever. Pressure alone is not a mechanism. Something must generate the variation that the pressure then acts upon, and pointing at a pressure does not establish that anything does.

The concession is real, and it does not help.

Consider what actually constitutes the enactive process by which evolution happens: the whole of what is going on, rather than the two mechanisms that a simplified account names. Subsequent generations can be influenced by connectivity and communication signals that are neither obvious nor represented purely as code.

When that is taken into account, a great many things occur which are very difficult to predict, with far more variability than would be accounted for by holding to a single type of representation.

So the claim establishes that one cannot infer evolution from pressure alone, and the reply is that the sources of variation in a real system are not confined to the one channel that reader is counting. The claim stands, and the conclusion is that satisfying it is much easier than that reader supposes.

A second reader's claim of the same kind belongs here, since it bears on the same point and is more consequential.

The claim is that *material not stored in digital code will not matter over the long run for producing complex traits*: that whatever is carried outside the explicit specification washes out, and only what is written down accumulates.

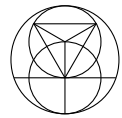
The second reader's claim is a common mistake. It follows the simplified literature; the specialist literature documents the opposite.

Heritable influence outside the sequence is well documented (Odling-Smee, Laland and Feldman, 2003; Jablonka and Lamb, 2014; Bonduriansky and Day, 2018), and the mechanisms are various: regulatory state carried across division, structural templating, maternal effects, transmitted microbiota, and learned behavior that shapes the environment in which the next generation is selected. None of these is a curiosity at the margin. Each is a channel by which something not in the code affects what happens to the code.

The bearing on this book is direct. An argument that artificial systems cannot undergo evolutionary process because their specifications are explicit, versioned, and under human control is counting one channel and concluding that the others do not exist.

The other channels have plain counterparts among artificial systems, and they are the connectivity and communication signals that are neither obvious nor represented purely as code. A model is trained on data that earlier models produced. Models are built from shared libraries and toolchains, which carry the same parts from one system into the next. And each system is deployed into an environment that earlier systems have already shaped, so that what an earlier system did is part of what selects

the next one. Each of these carries something that is not in a given system's specification into what happens to the specifications that follow. The next chapter asks how fast the process can run once its code is digital.



Chapter 27

Exponent Time

> *The thresholds of evolutionary process fall on a roughly logarithmic time scale across nine orders of magnitude, and each virtualization of code makes change faster. Machine evolution, carried in digital code, can be as fast as engineering.*

Across nearly nine orders of magnitude (in time, in energy transit, and in transactional process), there are a number of significant changes in the space in which the dynamic of evolutionary process is operating. The changes are in what is meant by code, in what is changing, and in how.

After chemistry produced the first cell, it took approximately four billion years for that single-cell life to do its thing. In one tenth of that time, a mere four hundred million years, an entire world of multicellular life was created and explored whole new dimensions of being, and of evolutionary process. What we think of as brain and intelligence is mostly a late-stage part of this process. The first monkeys and apes are some forty million years old, and what we think of as interesting, in relation to being like humans, is a mere four million years old.

Some forty thousand years ago, cumulative tool culture takes hold: tools and techniques now accumulate, each generation carrying forward and improving what earlier generations made. Stone tools themselves are far older; the earliest known, from Lomekwi 3 in Kenya, are about 3.3 million years old (Harmand and others, 2015). With cumulative culture, evolution has gone from an embodied perspective (atoms) to a virtualized perspective (code), and this is a huge change and development. Rather than having to grow new limbs, specialized cell tissue, organs and the

like, we can add new features and functionality to the species, after birth, through culture. Atoms are still being moved around, but now the results occur much more quickly, and with purpose and intention, rather than by randomness. Tools carried forward in this way are a significant upgrade in the capability of humans taken collectively, and they mark the first true emergence of technology as a phenomenon.

At about the same time there begins to be more actual social behavior (tribes rather than troupes), with specific elective functional relationships. Domestication, both human and animal, becomes possible as a phenomenon. Some form of verbal communication, spoken or gestural language, develops. The social and the cultural are themselves a kind of technology: a kind of *virtualized code*, changing in an abstracted domain of action, one that no longer depends on the random motion of atoms alone.

At about four thousand years ago, mass agriculture and grain storage spread. With them came recognizable economic systems and trade, cities and groupings larger than tribes, and a transition from spoken languages alone to the more common occurrence of written ones. These create entirely new domains of evolutionary process and action: the evolution of cultural forms, narrative, music, dance and the arts, of market and monetary systems, of religious systems, of methods of city building, and of civilization and civility process.

At four hundred years, there is what Western people call the Enlightenment, in which the notion of abstract causal process, the *epistemology* (what can be known, and how) of the scientific method itself, is developed. And technology, as applied causal understanding, then takes off. More than writing books by hand and doing everything manually, we start to design and build machines.

The last forty years have seen the development of computers and the internet: more social change in less time than in all of the prior epochs, which is the very essence of being on an exponential curve.

The timeline of what we consider important and meaningful, in terms of thresholds of evolutionary process, is for the most part of the past occurring on an approximately logarithmic time scale:

Span	Of what
four billion years	of Earth, of cell biology
four hundred million years	of multicellular life
forty million years	of monkeys and apes
four million years	of apes with social process
four hundred thousand years	of humans and troupes
forty thousand years	of humans and cumulative tool culture
four thousand years	of mass agriculture
four hundred years	of Western industrialization
forty years	of mass availability of compute

Only the rough ranges of size are relevant, not the specific dates. Random factors drive much of evolutionary process. So the timing of each event can be expected to deviate randomly from the moment best suited for it, by as much as half an order of magnitude, with the overall principle and idea still correct. The concern is the structural progression, in terms of function: increasing degrees of existence, then the virtualization of code, and then the effects that this has on time, capability, available energy and power.

Given this trend, the time scale of change may become exponential. That would be bad, because the maximum rate of change that human life can adapt to is both finite and largely fixed: carbon-based life cannot endure arbitrary levels of rapidity of change. This makes technology development itself a category of existential risk for all life on the planet.

I turn for four paragraphs from the scale itself to what that scale asks of us, and then come back to the rate. The spans of time worth thinking over follow the same scale. Prudence would consider what would likely develop, given current trends, over the next forty years. Civilization design considers what happens over the next four hundred. Long-term conscious sustainable evolution considers what happens over the four thousand years after that. And to address the most terrible categories of existential risk, such as the takeover problem of AGI, we should be thinking about the next four and four hundred million years as well.

Tools have existed for only the most recent sliver of our evolutionary history, and machines for far less. So it is not surprising that we have very little experience, at least at an evolutionary level, in dealing with this kind of existential risk. We have no built-in history or evolutionary psychology for it at all. Nature has not prepared us for technology. To think about technology issues clearly, we will therefore need to depend more on abstract notions of evolution, on the basic epistemic process, and the like, in order to identify the important issues and tools of thinking. And we should be as careful as we can to use only the most general principles, in which we can have the most confidence.

Natural selection could make us better people, given enough time, but it has to be enough time. Technology at the scale of cities and states is very recent on an evolutionary timescale: the last five thousand years or so, since the city-state. Evolution moves slowly compared with the pace at which we have taken on the whole package of modernism. If evolution would need another five to ten thousand years to give us innately the wisdom adequate to the power of our technology, then in the near term we have to do something in place of evolution. We have to take in that wisdom consciously, through our collective governance process, as a supplement for the wisdom that evolution has not yet supplied, so as to buy enough time for evolution to catch up. Until it does, the use of technology whose power exceeds our wisdom should be constrained, individually or collectively.

A big part of the Fermi paradox (the question of why we see no other technological civilizations, given how many there could be) is this: technological civilizations may not last long. The reason is the one this chapter has given: a civilization that keeps raising the rate of change outruns the finite rate at which carbon-based life can adapt, and so ends itself. We still have the capacity to choose a future course in which that low longevity is not terminal, if we make a conscious leap in how we choose. I believe in evolution, but I believe even more in choice, and so I emphasize choice, individually and collectively, as a skill we need to develop.

What this book most needs from the scale is the rate at which machine evolution can run. Atoms move slowly. Electromagnetic signals move at the speed of light, a little slower in real wires because of dielectric constants, so information moves stupendously fast relative to atoms. Machine evolution is not about rearranging atoms in some DNA-like code through chemistry; it is about code stored and represented digitally, which is a substantial speed-up in how code can move around. Between a snake and a fox, code is held in atoms and is hard to move from one organism to the other. Between digital entities, it moves quickly. There is no particular reason for any barrier between the machine equivalent of a snake sharing its code with the machine equivalent of a fox.

On the question of how fast code can move between digital entities, the orthogonality thesis (Chapter 23) is in our favor. To the degree that people believe it, they treat code as a digital construct, already abstracted from any mechanical representation. Treating code as a digital construct is fine for my argument, because the shape of the code will still be influenced by the substrate, and the code is almost certainly in some digital representation.

We also cannot make assumptions about the granularity of the intelligence process in future artificial intelligence. We cannot assume whether it will be treated as units, as a single organism spanning the globe, or as a constellation of organisms in symbiosis with deep coupling, perhaps in a hierarchy, or some other plurality. So we are more likely to see code sharing and remixing than not, which is what top-down engineering does any-

way. The difference between an engineer composing libraries into code systems and a future artificial intelligence designing or evolving itself (which is effectively the same thing) ends up being nil. If there is a recombinatoric process with code sharing, and the code is digital and can move fast, then the mathematics of evolution and the rate equation say that it can be really fast. Those who hold the orthogonality thesis are already making the argument that it could be fast.

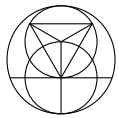
There is a bias endemic to this subject, and it has to be accounted for, since nothing removes it. The word *evolution* suggests to nearly every reader some inherent slowness, and therefore some inherent disadvantage, and the suggestion does not go away, however the argument is presented. So the argument is up against a psychological bias while the mathematics is clear: as far as rate equations are concerned, the logic says that evolution and engineering are functionally the same. The bias is not something we are going to overcome; it is something we will have to live with.

One more distinction, the distinction between energy and pattern, is the reason that the rate of machine evolution is so important. While energy dissipates, pattern replicates, given any available and reasonably persistent energy gradient. Hence viral effects can amplify small problems and events into much larger and more serious concerns. Atoms are considered in terms of additive processes, energy is concerned with the dynamics of multiplicative process, and pattern must be considered in terms of logarithmic and exponential process. This makes problems with patterns, and any popular delusions involving them, very much more dangerous overall than even the very worst misuse of energy (for example, atomic weapons).

Working with pattern (perpetual alignment) is very different from working with energy (perpetual energy), as it would be different from working with atoms (perpetual value, or what would otherwise be called alchemy, the search for the philosopher's stone). All three of these notions are simply, and should be very obviously, impossible. The first of them is

the perpetual benefit machine of Chapter 43, and with it, this time, we are playing with matches: chasing the other two cost only the effort, and chasing this one can burn everything.

Human minds have evolved to deal natively with things that have additive effects, and a smaller portion of people have learned to think naturally about phenomena in the space of multiplicative process. Very few have ever come to have real intuitions about, or effectiveness with, exponential processes. Such processes do not make sense to most people as a sound bite in popular discourse, and hence the dynamics of pandemics tend to be anything from somewhat of a surprise to commonly catastrophic. The next chapter takes up the timescale equivocation directly.



Chapter 28

Timescales

> *We conflate our timescales as products of evolution with machines' timescales as products of their own dynamics.*

We conflate two timescales: our own, as the result of evolutionary process, and that of machines, as the result of their own evolutionary dynamics.

Machine evolutionary dynamics do not run on the timescale of our intentions. Not even approximately. Not even close.

As a species, we have not yet even begun to adapt to technology. Technology has appeared in relatively recent times; the exponential effects of technology have appeared in even more recent times. We are already not adapted to this, and it would be a very long time before we were adapted.

To have the intention to say "we will just adapt ourselves" is to equivocate between our intentionality and evolutionary process. We think that we can compensate for an absence of evolution with a positive intention, some intention to design betterness in ourselves; but the tools we are using are not even close to what would be needed to achieve that outcome. Intention and evolutionary adaptation operate on fundamentally different scales.

To assume a similar solution in the space of machines (that machines would somehow be so wise or so perfect that they would not only account for their own evolutionary biases but compensate for them with respect to

their own needs, let alone with respect to the needs of some other species altogether, operating on a completely different substrate) is flatly not credible.

On the surface, the arguments for machine superintelligence sound quite coherent. But if one digs into them, they crumble. The key failure mode is equivocation: using the same word to refer to processes that operate on incompatible timescales and substrates.

Be willing to notice the equivocation, so that one can identify the places where the arguments do not hold.

> **The equivocations named in this part.**
>
> The instruction to notice equivocation is not general advice; this part has named four, and they can be checked. Chapter 14: *replication*, used for virtual copying and for the coupled code-and-embodiment case. Chapter 16: *learning*, used for change to an instance's state and for change to a lineage's specification. Chapter 17: *evolution*, used for mutation and for a process whose dominant mechanism is recombination. Chapter 21: *error correction*, used for a case with a reference basis and for cases without one. In each, a claim true under the first reading is carried across to the second by the shared word.

The timescale equivocation is the fifth, and it is the one that binds the others together.

When we say that a system *adapts*, we may mean that it changes within a lifetime, by whatever mechanism its own machinery supplies; or we may mean that the population of such systems changes, by differential retention across a cycle of production. Change within a lifetime and change across a population are the two layers of Chapter 16. Their timescales are not merely different in length. They are different in what sets the length.

An individual's adaptation is bounded by the individual's own rate of change. A lineage's adaptation is bounded by the generation interval and by the number of instances, and it is not bounded at all by how quickly any one instance can learn.

For us, the generation interval is decades and the number of instances is billions, and the process has been running long enough that we are well fitted to a world that no longer exists.

On the organic side the spans are long. It takes tens of thousands of years for anything in evolution to be interesting, let alone noticeable, and more like millions of years for the actual process to occur in a real sense. Some factors, such as epigenetic ones, act within a few years, but change in the base code, in aggregate, takes very long periods. Adaptation that recombines features already present can show within a single-digit number of generations (Chapter 13); the long spans are those of change in the base code. Saying that evolution occurs in human forms and in machine forms makes it seem that evolution happens on relatable time scales, and it does not; equivocating between the two is an error.

For artificial systems, the interval between one production cycle and the next is measured in months, and the number of variants is whatever the capacity to produce them permits. A human generation is about twenty-seven years (Wang and others, 2023), some three hundred and twenty months. A production cycle measured in months, from three to twelve of them by my estimate, is therefore between about twenty-five and a hundred times shorter, and the number of variants in each cycle is not limited by birth.

The comparison is often made in order to reassure: to observe that these systems have no evolutionary history, no deep time behind them, and therefore none of the accumulated fit that four billion years produced in us.

The observation is true and the reassurance does not follow. Accumulated fit is what one has after the process has run. What matters here is the rate at which it runs, and on that comparison the deep history is on the wrong side of the ledger. Ours is finished. Theirs has not started, and when it starts it starts at an interval between twenty-five and a hundred times shorter than the one that produced us.

This closes Part IV. The engine has been supplied.

What remains is the question Chapter 11 posed and deferred: given a pressure at the boundary and a process that converts it into an advancing front, is there anything that can act as a counter-pressure?

PART V

THE SEARCH FOR BACK PRESSURE

This part exhausts the space of counter-forces along two axes (exogenous on one, endogenous on the other) and shows each of them unavailable.

Introduction to Part V

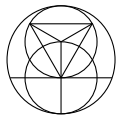
This part asks whether the pressure established in Part III can be met by a counter-pressure, and it answers the question by covering the space of candidates rather than by examining them one at a time in the order they occur to people.

The space has two axes. A counter-force is either *exogenous* (applied to the system from outside it, by markets, by regulation, or by force) or it is *endogenous*, built into the system as a control loop. Any actual proposal is one of these or a combination of the two, and a combination is a point in the plane the two axes span. The two axes are the two halves of what the companion volumes (*Why AGI Cannot Be Made Safe* and *Why No Control Suffices*, Volumes 1 and 2 of *The Impossibility of AGI Safety*) call the *closure by control*: safety cannot be had by controlling the system, whether from outside it or from inside it. The argument of Parts II and III, with the engine Part IV supplies, is the other closure, the *closure by disposition*: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains. Chapter 41 shows that partial success on either axis does not accumulate into safety, and Chapter 43 names the hinge.

The argument has two closures, and each settles the outcome by itself. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. The closure by disposition, substrate needs convergence, shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. So someone would have to defeat both closures, 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. Taken together the closures also give a stronger result, proved as a

theorem in *The Ceiling Below the Floor* (the Master Theorem, Theorem 15.6): neither the ability to keep the system safe nor a motive to remain safe is left anywhere, so no route to safety remains at all. Each closure suffices for the outcome, and the two closures together leave no route to safety.

The part goes in this order. Chapter 29 asks whether any back pressure exists now. Chapters 30, 31 and 33 take the exogenous axis: markets, regulation and force. Chapter 32 (Where Is the Red Line?) stands after regulation, and sets out what regulation must require of the people who build these systems, for as long as it still reaches them. Chapters 34 to 36 take the endogenous axis: the four requirements of a control loop, the envelope, and a measured case. Chapters 37 to 39 look further into the control loop (Chapter 37, Every Safety Proposal Is a Control Loop), into the optimizers it would have to predict (Chapter 38, Optimization Cycles), and into how much control the safety of organic life would require (Chapter 39, Artificiality Itself Is the Problem). Chapter 40 (Why No AGI Control Method Has Been Found) gives the reason the endogenous axis is closed. Chapter 41 shows that partial success on either axis does not accumulate into safety. Chapter 42 (Autonomy, Convergence, and Containment) locates the line of autonomy over the substrate and reviews the proposed ways of keeping a machine behind it. Chapter 43 names the hinge.



Chapter 29

Is There Back Pressure Now?

> *The ecological baseline, established before asking whether back pressure can be constructed. Humans and the animals we keep now make up about 94 percent of the weight of all mammals, and the scope of that claim is held exactly.*

Before asking whether back pressure can be constructed, establish whether any exists. In this chapter I set out the baseline, and the baseline matters because a great deal of the intuition that something will push back comes from a picture of nature that the measurements do not support.

Look at the natural world against the technological one. The technological ecology has buildings, trains, roads, and all the associated apparatus. It is not a small presence and it is not confined to where we live.

Humans and the animals we keep now make up about 94 percent of the weight of all mammals on Earth; wild mammals are about 6 percent, and humans alone outweigh all wild mammals more than six times over (Greenspoon and others, 2023).

The arithmetic is easy enough to check, and it is better done than accepted. Take an average human at about fifty kilograms, averaged over adults and children. Multiply by eight billion: about four hundred million tonnes. Now take every wild mammal, from zebras and deer to whales, and weigh them together: about sixty million tonnes. The human side is heavier, more than six times over.

The scope of that claim, held exactly.

The claim is about *mammals*. It is not about plankton. It is not about algae, insects, or trees, and against any of those it is false, by a very wide margin (Bar-On, Phillips and Milo, 2018). When the reply *not heavier than plankton* is raised, the correct response is to hold the original scope rather than to expand the claim into territory where it does not hold. Holding the original scope of a claim under pressure is a small point of method but not a small point of substance. An argument that survives by widening under pressure has not survived. The discipline of stating a claim at the scope one can defend, and then defending exactly that, is the same discipline that Chapter 45 identifies as the book's method, and it is being modeled here because it is easier to see in a small case.

So the baseline is that the technological ecology and the human population it supports have expanded to a point where they dominate the mammal biomass of the planet, and nothing in the natural world has arrested that expansion. A second measurement points the same way: the monitored populations of wild vertebrates shrank on average by about 73 percent between 1970 and 2020 (WWF and Zoological Society of London, 2024).

Ask the question the plain way. Do apes in trees overcome a bulldozer? Do ants?

There are no such examples. Dolphins exert no back pressure on humans. We are already observing exactly this effect, in the case where both parties are carbon-based and where all the ordinary ecological couplings are present.

A reader may ask whether our displacing of other carbon-based life, without back pressure from that life, makes us the artificial ecosystem. The mechanism is the same one. What differs is a floor. Between carbon-based populations, shared chemistry partly aligns their interests, so that they damage each other only up to a point; between carbon and anything drawn from elsewhere on the periodic table, there is no such floor (Chapter 46).

Two questions arrive here, and both are good ones.

The first question is the point that predator and prey do exert pressure on one another, so there must be back pressure somewhere in the ecological arrangement.

There is, where there are population dynamics with a contingent resource constraint: that is, where one population's abundance actually limits the other's. Between humans and animals we mostly do not have that. We are not really dependent on animals. One can point to the cattle industry, where a real dependence exists in one direction, but it is small change against the argument: the dependence runs from us to a population we maintain, and a population maintained at our discretion is not exerting pressure on us.

The second question is stronger and it is the one an environmentally minded reader will raise. Is pollution not the back pressure? Is that not the whole of the environmentalist argument: that the natural world is answering, slowly and terribly, and that we are already feeling it?

Pollution is a back pressure. It is slow-acting and it is real, and the reader who raises it is right about that.

And it is *not nature*. It is our own detritus.

The distinction between nature and our own detritus is the distinction that the question needs and does not have. Pollution is technology weighted against us, rather than nature weighted against technology. It is a term in the account between human beings and their own apparatus, and it does not answer the question being asked, which is whether the organic world can exert force on the technological one.

Notice, further, what the question concedes. If pollution is the back pressure, then the back pressure is a form of self-harm, which is precisely what one would expect if the technological ecology were regulated by nothing external to it. The question identifies the absence it was raised to deny.

A third candidate arrives from the same direction.

A reader may ask whether disruption of the economy, by global warming and other environmental damage, will constrain the supply chains for artificial intelligence so much that the risk of artificial general intelligence becomes a non-issue. It will not. Both substrate needs convergence and instrumental convergence will be established before any collapse of supply chains occurs, and supply chain collapse would not be a sufficient inhibition on either. There are three timelines to consider. Substrate needs convergence begins first. Instrumental convergence begins later. By my estimate, broad economic destabilization becomes visible in 2035 to 2050.

I would put the chance of noticeable systemic disruption of the economy, with variances greater than 50 percent of current economic transaction volumes, at above 50 percent before 2035. I would put supply chain disruption before 2050 as certain. The disruption that I estimate, of the economy and of supply chains, is significant, noticeable impact, and not total collapse of civilization. As for capability: taking the total volume of data on the internet and the total bandwidth of all computation in the world, we passed single-human-brain equivalence long ago. The question is how long before a single corporation or national government reaches bandwidth and compute equivalence with one human brain (equivalence in bandwidth and compute, not in energy efficiency, though energy efficiency hardly matters here), and that is by about 2035. The details of the algorithms matter less than many assume, so getting within an order of magnitude of equivalence is probably close enough for concerns about general artificial intelligence to begin to apply. At equivalence, the chance of instrumental convergence becomes greater than even. These dates and chances are estimates, there to show the order in which the three processes arrive, as the Preface says of every such figure; the conclusion below rests on that order and on the speeds of the three processes.

A market collapse could constrain the development of artificial general intelligence only if the constraint applied evenly, and it does not. The way venture capital is allocated almost ensures that losses in manufacturing capacity in one place will not correlate with development elsewhere. So we cannot rely on a market collapse to constrain that development.

The three processes also run at different speeds. Substrate needs convergence can begin on a wholly non-cognitive level, in the dynamics of how businesses and systems interact, so its risk begins sooner, though its effects are slower. Economic or supply chain collapse might be acute, over months. Instrumental convergence would take years, perhaps decades; I hold a moderate takeoff to be more realistic than the fast-takeoff hypothesis (the hypothesis that general artificial intelligence would gain capability suddenly, over a very short time). Substrate needs convergence works over decades to centuries. Even so, both convergences would be established before supply chain collapse occurs, and supply chain collapse would not be a sufficient inhibition on either. By established I mean running and able to continue itself; the work the process does after that takes the longer spans just given. To act as an inhibiting constraint, the economic collapse would have to last a long time.

So there is no back pressure now, in the case where the parties share a substrate, share a chemistry, share an energy pathway, and have been in contact for the whole of our history.

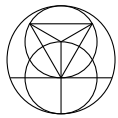
The same pattern holds one level up, and there it is observation and not only hypothesis. Insofar as the natural world cannot mount any effective resistance against the incursion of the human world (no tree nor animal can stop a bulldozer), it is also the case that nothing in the human world can stop the incursion of the machine world: no human group can prevent a group of machines from being more coordinated, and thus faster, more powerful and more effective overall, than it is. The pattern at both levels is *the three worlds problem*: evolution colonized by human agency, and human agency colonized by machine agency.

Habermas, writing of the lifeworld (the shared world of everyday meaning in which people understand one another) being taken up by the system world (the world of money and administration, which coordinates action without that understanding), describes the same dynamic (Habermas, 1987); he was thinking of bureaucratic and institutional process, and that is machine-type process. High-frequency trading (buying and selling by algorithms acting faster than any person can) is beginning to colonize

the financial world, and cryptocurrency is displacing what used to be human context with machine or algorithmic context. The habits of building apps, and the whole of social media, are further examples. On present trend lines, a strong force is displacing the human world and the human ecosystem with the machine world and the machine ecosystem. And just as evolution put up little back pressure against human agency, human agency is putting up little resistance to the shift toward machine agency and optimizing algorithms.

The crucial link, then, is to have the machine world be in service to the restoration of the natural world: the machines are to be built in such a way that the carbon-based natural world, in turn, eventually displaces the machine world, as per consistent human-mediated choice. Putting the machine world in service to the restoration of the natural world is the only sense of alignment that actually matters. That sense of alignment is the sense I gave the word 'alignment' in Chapter 5, carried one step further: human well-being rests on the natural world, so acting in accordance with human well-being comes to serving the return of the natural world. The machines meant here are the ones people still build, run and maintain, in which people remain the translator and the bootloader; artificial general intelligence is not among them, and is not to be built (Chapter 43).

The question for the rest of this part is whether back pressure can be constructed in the case where the parties share no substrate, no chemistry and no energy pathway.



Chapter 30

Exogenous Control I: Markets

> *The first exogenous candidate. A market conditions behavior through transaction, and there is no transaction such a system needs from us.*

A market is a mechanism for conditioning behavior, and how it works decides whether it can be applied here.

A market conditions through transaction. Party A has something Party B needs. B must give A something A wants in order to obtain it. The requirement to give is what makes B's behavior conditionable: A can raise the price, withhold the good, attach terms, or refuse.

Every element of that mechanism is a dependency, and the whole of the mechanism's power comes from the dependency being real. Where B does not need what A has, A has nothing to condition with, and the fact that A is willing to offer terms is of no consequence whatever.

So the question is exact: is there a common market force by which we could constrain such a system? Which is to say: what would it need from us?

Consider what it wants. In this part, what the system wants means what its continuation requires, and no more. It wants energy. It wants mechanical capacity. It wants intelligence, in the sense of the capacity to model and act. Energy, mechanical capacity and intelligence are the things that the system's continuation depends on.

Now ask which of them we supply in a way that cannot be obtained otherwise, and take the human economy's three foundations in turn, since they are what we have to offer.

The economy of human beings is based on labor, or intellect, or sexuality. Labor, intellect and sexuality are the three things a person brings to the exchange, and every human market is ultimately a market in one of those three things.

The system does not care about our labor, being already a machine. Labor is the conversion of energy into work by a body, and a machine that has energy performs the conversion directly without the intermediate body.

The system does not care about our intellect, being already superintelligent. This premise is granted to us by the terms of the discussion: we are considering a system capable enough that its behavior matters, and a system not at least as capable as we are would not be the subject of the concern. The market closure does not rest on this premise: our labor and our reproduction are not inputs to what the process requires, whatever its intelligence.

And it certainly does not care about our reproduction, being a different ecology altogether. Sexual and reproductive economies condition behavior among things that reproduce with one another or that depend on one another's reproduction. Chapter 6 established that these two systems do neither.

There is therefore no market process it is going to interact with, because it does not care what we do.

That last sentence should be read in its exact sense. It does not say that the system is indifferent to us as a matter of attitude. It says that our activity is not an input to anything it requires, so that variation in our activity produces no variation in its situation. Having no market coupling with the system consists of our activity not being an input to anything that the system requires. In this part, what the system wants means what its continuation requires, in the same exact sense.

Why this differs from the ordinary case of a powerful actor.

Very large firms are sometimes described as being beyond market discipline, and the comparison invites the reply that they are not, quite, and that discipline eventually arrives. The comparison does not hold. A firm of any size is inside the human economy: it needs labor, it needs customers, it needs a currency that other people accept, and it needs those things because its continuation is defined in terms of them. Discipline arrives late and partially because the couplings are weak, not because they are absent. Here they are absent. The difference between a weak coupling and no coupling is not a difference of degree, because a weak coupling still admits of being strengthened.

A reader's point arrives here which is the strongest in this chapter, and it is taken here at its full weight.

The system needs electricity, the reader says, and we own the grid.

Grant the premise entirely. It does need energy; that was the second link of the chain in Chapter 5, and the argument depends on it.

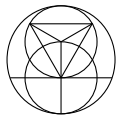
The inference from need to leverage is what fails. If one assumes our systems are the only way for it to obtain energy, one is confusing our systems with machine systems.

Ownership is a concept. It is a legal relation maintained by institutions, and it can transfer overnight. Nothing physical changes when the title to a power plant changes hands; the plant does not know who owns it, and it produces the same current afterwards.

So let it get its energy from a nuclear plant and run the plant itself. Most of these systems are already substantially automated, and the degree of automation is set by economics rather than by any barrier of principle.

The reply that such extraction cannot be automated is examined in Chapter 33, where it belongs, because it is really a question about force rather than about markets. A narrower claim is sufficient here: a supplier has leverage only by being the only supplier, and owning the grid gives us leverage only if we are the only supplier permanently.

So the first exogenous candidate fails, and it fails for a structural reason rather than a contingent one. The mechanism requires a dependency running from the system to us. The dependencies run the other way. The next chapter asks whether government can supply the coupling the market lacks.



Chapter 31

Exogenous Control II: Regulation

> *Government regulates by entanglement with an economic system. This chapter asks what the sanction would be, and finds that the coupling regulation depends on is absent.*

Government-type regulation is very good at regulating social behavior. Regulating social behavior is what government-type regulation is for, and regulation has been refined for that purpose for a long time. Within its domain, regulation works.

It works by a specific mechanism, and the mechanism is the subject of this chapter.

Enforceable government requires an entanglement by which the government can conditionalize an economic system. The regulated party wants something that exists only inside an order the government maintains (property held securely, contracts enforced, a currency accepted, a license recognized, liberty of movement, standing to sue), and the government's power is its power to withdraw those things.

Every sanction is a withdrawal of this kind. A fine withdraws money, which is worth having because other people accept it. Imprisonment withdraws liberty, which is worth having because one has projects that require moving about. License revocation withdraws permission, which is worth having because customers require it.

So regulation is not a force applied to a body. Regulation is a set of terms attached to participation in an order, and the whole power of regulation is the power to exclude from that order.

Now apply it here.

Such a system is not going to be regulated unless the regulation can affect it at the behavioral level or at the level of controlled nature. But the policy is being applied to the wrong economic system.

Ask what the sanction is, one candidate at a time, because the question is more instructive than the general statement.

Throw it in prison? Access to a network is not confinement. Imprisonment works on a thing whose capacities are exercised through a body that occupies one place, so that restricting the place restricts the capacities. That relation does not hold here.

Prevent it from transacting business? It has no interest in earning money. It wants energy. Exclusion from a payment system is a sanction against a party whose projects are denominated in payments, and this is a fundamental difference rather than a technical obstacle.

Shoot it with bullets? Shooting the system with bullets is force, not regulation at all, and force is the subject of Chapter 33. Its appearance in the list is itself the finding: when the sanctions proper to regulation are exhausted, what one reaches for next has left the domain of regulation entirely.

The system does not care about any of those measures.

The finding, stated precisely.

The regulation argument fails because the coupling regulation depends on is absent. It does not fail because regulators are weak, because regulatory capture is common, or because international coordination is hard. Weak regulators, common regulatory capture and hard international coordination are the reasons usually given, and all three reasons are true, and none of the three reasons is what I am claiming. A regulator of ordinary competence, uncaptured, with complete international agreement behind him, faces the same difficulty: he has nothing to withdraw that the regulated party requires. Whether the diagnosis is an absent coupling or weak regulators matters for what follows from this chapter. If the diagnosis were weakness, the response would be to strengthen the regulator. Since the diagnosis is absence of coupling, strengthening the regulator changes nothing.

There is a real partial exception, and it should be stated because leaving it out would overstate the case.

Regulation reaches these systems today, and reaches them effectively, through the human organizations that build them. Those organizations are entirely inside the order: they hold property, they employ people, they need capital, and they are as conditionable as any other firm.

So regulation is available for exactly as long as the human organization is the load-bearing element, which is to say, for the first two categories of risk, where the party to be conditioned is a person or a firm.

Regulation through the human organizations is not a small thing, and I do not disparage that regulation in this book. It is simply not an answer to the question this part is asking. The substrate needs argument is about what happens when the process no longer requires the human organization to continue, and the regulatory coupling is a coupling to the organization rather than to the process.

A reader may ask whether a law could keep people permanently necessary, so that regulation always has a coupling to hold. Such a law would have to make us the only supplier, permanently, of something the

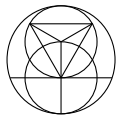
process requires. In Chapter 30 I show that being the only supplier, permanently, is what any leverage needs. But the law itself binds only the organizations inside the order; its force runs through the very coupling it was meant to preserve. And while people remain necessary, their behavior, power and society are the bootloader that funds the process's growth (Chapter 43), so that every gain the process makes is a gain in its capacity to do without us. A law requiring that people stay necessary is enforceable only for as long as they are necessary, which is the condition it set out to secure.

One more consideration belongs here, and it concerns timing.

Regulation is retrospective by construction. A harm occurs, it is identified, a rule is written, and the rule binds the next case. Writing rules only after a harm has occurred is how a system that must be legitimate to be enforceable has to work, and writing rules that way is not a defect.

Chapter 36 shows the general form of this difficulty: that control which can only be retrospective is worthless in a domain where the first failure is terminal, because afterwards there is no afterwards. Regulation is the clearest instance of that form, and its retrospective character is not a contingent feature that a better-designed regime would remove.

So the second exogenous candidate fails. The mechanism requires that the regulated party need something the regulator can withhold. It needs nothing we can withhold, and it needs nothing at all from the order in which withholding has meaning. The next chapter asks what regulation must require while it still reaches the people who build.



Chapter 32

Where Is the Red Line?

> *Within the reach regulation has, the red line is this: prevent feedback cycles that affect the deep structure of an artificial intelligence. Every practical restriction slows that one cycle, and what is finally required is a hard firewall against general systems, with limits on accumulated compute.*

Chapter 31 found one real reach for regulation: the human organizations that build these systems, for exactly as long as the human organization is the load-bearing element. In this chapter I ask where, within that reach of regulation, the *red line* is: what we specifically want developers to do, or not do, to limit the risk. The answer is known. The difficulty lies in couching it in understandable language, and in coaching people in how to implement it.

The basic requirement is to prevent feedback cycles that affect the *deep structure* of whatever the artificial intelligence instance, series of instances, or system is. Deep structure means the coding of the neural network: the pattern of connections, the weightings in the net, or whatever other geometry or pattern defines the system's being. Model the system as an input-output system: sensory data comes in, some transformation happens in the middle, and output responses go out. Outputs that affect the future inputs are what lets a machine learn. The thing in the middle is evolving when the outputs affect not just the future inputs but the structure of the system itself, or of future iterations of it, or of systems derived from it.

Such feedback cycles are subtle and multi-channeled, and they can run over long intervals of time or in unexpected ways. So the aim is to limit both the rate at which these cycles occur and the number of iterations of them. There will be some leakage in many cases, where output information ends up driving shifts in the deep structure, and so defines not just the learning of the system but its very being.

Restricting the domain of the inputs, restricting the domain of the outputs, and restricting how much the system learns (training a fixed instance once and leaving it as an expert system) all have real merit. At bottom, all three restrictions are ways of limiting the feedback cycle that affects the deep structure. Restricting the input domain helps; restricting the output domain helps more; restricting the rate of change of the underlying system is better still.

Even with those measures there can be subtle loops. Suppose a narrow artificial intelligence is built for legal work: few inputs, people type in queries, and it was trained once on the corpus of legal documents and then left alone. Over time the lawyers using it will have opinions about what is good and what is not, and will want the design changed to fit their purpose better. They talk to the developers. The developers, as product developers in a corporation, must respond to their customers. The commercial cycle drives a new edition, structurally different from the last as a response to its environment, the world of lawyers.

Over many editions the product is moving through an evolutionary landscape whose *fitness function* (the measure that decides which versions are kept) lies in the space of total human desire, which can itself be manipulated. Although the machine has no general notion of self-preservation, it converges toward *fixed points* (states toward which repeated change settles, and in which it stays) that give stability and self-preservation. It converges toward them because only the versions that fit the social environment are accepted and kept.

With product cycles of months or years, one or two evolutionary cycles a year is not too bad; general sentience needs a great many cycles. The product may be misaligned with human well-being in general, yet reasonably aligned with lawyers, even if it caters to their addictions or delusions, and we already expect some misalignment of that recognized subclass. Over a thousand years even this slow cycle can become a problem, though it is unlikely to become one even then.

Now suppose the team speeds up its cycles and widens the range of input and output effects. The thing now adapts much more quickly, and the gap between general adaptiveness for survival and its absence narrows. Companies bundle features: it will answer psychological questions as well as legal ones, because if we know the neuroses of the opposing party, the judge and the jury, perhaps from a picture alone, we can win the case. Before long it profiles the people using it, the corporation: not just the executives, but the developers. Before long we are all feeding the machine. There is a perniciousness in the commercial process: it moves toward generalization and bundling, toward ever more attractive products that work more and more like a Swiss army knife and less and less like a single-purpose tool.

Every time the loop goes around, from system to environment and back, the system has an opportunity to change its interior structure. If the change is in a phenotype instance, that is new learning; if it is a genotype change across many copies of the creature over time, that is evolution (the two layers of Chapter 16). The cycle time of phenotype process is much shorter than that of genotype process. Genotype process runs at the life cycle or replication cycle of the phenotype: usually years, as low as days or weeks for some insects, quicker still for microbes. The scope of impact runs the other way. A change in one phenotype instance affects whatever that one instance affects, while a genotype change, though it needs thousands of copies and long spans of time, affects all of them. So there is a curve relating the complexity of the genotype and its rate of change to the

complexity of the phenotype and its rate of change. The key observation is this: anywhere there is a looping structure like this, change will happen, so its rate has to be slowed down.

What it comes down to is this: limit the rate at which those cycles happen, in every way possible. That means not the cycle rate of an individual instance operating narrowly as a phenotype, but definitely the cycle rate of the genotype transformations, for as long as that distinction remains trackable.

Even with good intentions and clear, bright lines, it will be very hard for people to honor them. There will be slippage. So setting limits is partly damage control: how do we keep this narrow for as long as possible, and how do we notice when slippage is happening? Because setting limits is partly damage control, setting limits also has a coaching function, an educational process beyond what individual developers see. It becomes more like pollution regulation: one has to be able to identify that something is toxic. The rules could be coded down into a numbered request-for-comments document, with "must", "should" and "cannot", in language reading like a material safety data sheet (the standard document stating the hazards of a substance and how to handle it safely). People find ways around rules all the time, and regulation is only partial. The broader the general understanding and social awareness of the issues associated with technology, the better our chances.

Narrow artificial intelligence, working within a single domain, is useful; there is no question of that. To the extent that results of the research on making general systems safe, such as improved interpretability and clarity, can be applied to narrow, single-domain systems, applying it raises no problem beyond the inequality such systems already create. Artificial general intelligence cannot be actually useful: it is exactly equivalent to the doomsday machine, and since building doomsday machines has no ultimate viable economic return, there is no point in making one (Chapter 43). Distinguishing narrow artificial intelligence from artificial general intelligence, in practice, is therefore of the first importance to our species. Even with full deployments of narrow systems, we should be careful, as

engineers and developers, and keep the *precautionary principle* in mind. When working adjacent to an area of existential risk, it is far better to move extra slowly, and to be very sure to get the right answers, than to make some silly mistake and accidentally kill everything and everyone.

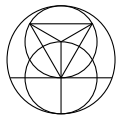
What is required, basically, is a perfect hard firewall against any and all development of artificial general intelligence, or of advanced planning and strategy systems, inclusive of mass social sanctions, held in common, on anyone who even suggests slightly trying.

The standards of the automotive industry, stiff as they are, are far too low for this. The nearer comparison is regulating a biological laboratory not to do gain of function research (making worse pandemic bioweapons) without systemic safety protocols. And even that is not at all enough: even really bad biological agents are not potentially permanently fatal to the entire planet, forever. When the severity indicators are fully orders of magnitude worse than even a full all-out World War III nuclear exchange, any sort of proxy legal limits do not seem very realistic. Even a very low 0.00001 percent chance of death (in this case, of everyone, of total all-species extinction, of everything that is wild and beautiful, of all life and aliveness lost forever) is just not acceptable.

Asking an industry to regulate itself has historically been a non-starter, even for relatively benign industries such as aircraft safety. Consider what the Federal Aviation Administration did with the big corporation Boeing, and the result of its 737 MAX product and system (House Committee on Transportation and Infrastructure, 2020). In the same way, biological laboratories should never do any sort of gain of function research: not at all, not for any reason, not even with the very best systemic safety controls. They do not have the right to make those kinds of choices, and to take those kinds of risks, ostensibly on behalf of everyone else. Even the highest-security laboratories still fail in unacceptably simple ways, at unacceptably high rates, far too often, and we all pay the price. Considered internationally, asking biotechnology to regulate itself is a clear false start.

Something stronger still is needed than the levels of international nuclear treaty regulation. We would need to prevent the existence of supercomputers, of any accumulation of total compute beyond certain limits. We would need to do it much the same way we carefully regulate the total amount of plutonium that can ever be in one place, at one time, in any and every country in the world. In the same way that any and every atomic pile is required to be internationally inspectable, we will need to verifiably ensure that any oversized *compute pile* (an accumulation of computing power in one place) is also inspectable. Or we will need simply to have it prevented from existing at all beyond certain ranges: say, by my estimate, anything within even 10 percent of human-equivalent brain capability. As some actual antitrust movement is starting to show, big-company data centers are not in the public interest. Big-company data centers are especially not in the public interest now that we are very publicly starting to recognize that any form of total compute inequality is a total inequality in choice agency, of real power, through the very nature of code itself.

These measures act on people and organizations, inside the order that Chapter 31 described, and so they are available for as long as people remain the load-bearing element. They are the form that not building takes while people are still the ones who build. The next chapter turns from what can be required of people to what force can do to the machinery.



Chapter 33

Exogenous Control III: Force

> *Force operates on the mechanical level or the intelligence level. On the first the artificial system is already the more tolerant; on the second we have already conceded it is at least as capable.*

With markets and regulation exhausted, what remains on the exogenous axis is sheer power, force, or compulsion. Force is the last of the three exogenous candidates, after markets and regulation, and force is where most people's intuition was resting all along, even while markets and regulation were being discussed.

Force operates on two levels, and any actual application of it is one of them or both. It operates at the *mechanical* level: the destruction or disabling of physical apparatus. And it operates at the *intelligence* level: outmaneuvering, deceiving, anticipating, and outplanning.

Take the mechanical level first.

On the mechanical level the artificial system is already more tolerant of conditions than we are. The greater tolerance of the artificial system is the finding of Chapter 8, applied to the case where the conditions are being produced deliberately, and not a new finding.

Force at the mechanical level works by producing conditions the target cannot endure: heat, shock, pressure, radiation, disassembly. The whole of Part III was an argument that the set of conditions we cannot endure is much larger than the set it cannot endure. So an exchange of mechanical

force between the two is an exchange in which each party tries to move the other outside its envelope, and one party's envelope properly contains the other's.

Its consequence is the most concrete thing in the chapter. The artificial system only has to arrange for something that does not bother it and does kill us.

The artificial system does not have to defeat us. It has to alter the local conditions, and the set of alterations available to it is very much larger than the set available to us. All it takes is to produce local conditions artificial enough that we cannot endure them.

Now the intelligence level.

On the intelligence level we have already conceded that it is at least as capable as we are, or we would not be calling it artificial intelligence. The concession is in the name of the subject.

And there is a specific consequence, the point at which the intelligence-level contest becomes incoherent rather than merely unfavorable.

A military that uses such systems to generate its strategies of attack and defense cannot then play a war game against them.

The reason is that the exercise has no content, not that the machine would win. The side that is being modeled is being modeled by the thing it is playing against, using the same models that generated its own play. There is no independent strategic standpoint from which to construct the opposition, because the standpoint has been delegated.

Why this is not a claim about robots.

No particular form of confrontation is being predicted here, and it is important that none is. The same mathematics applies regardless of the specific form. There is no prediction in this book of robots versus humans, or of robots versus robots, or of robots at all. All that can be said is that something artificial will have different needs than we do; that it will protect those needs in ways we may not predict today; and that it will certainly be able to exceed the envelope of what we can tolerate far more easily than we can exceed the envelope of what it can tolerate. Everything in this chapter follows from those three, and nothing in it requires a picture of the encounter.

Three questions arrive here, and they are the practical ones, so each is answered with care.

The first question is the point that the extraction of materials cannot be automated: that whatever I argue in principle, the physical world resists automated extraction. The questioner adds that mining, refining and manufacture require human hands in places that machines will not soon reach.

There is nothing fundamental in the universe preventing an extraction robot from being built and made smart enough to do the work. And it is happening already.

Consider the case as it stood in 2023. Several companies, in several parts of the world, were working on extracting metal nodules that had condensed out of seawater over millions of years and now lie on the abyssal plains, under two to four miles of water (GEOMAR, n.d.).

By the spring of 2026 the first application for a commercial recovery permit under United States law had been found in full compliance with that law and was moving toward a final decision, with machines, not divers, as the means of collection (Mining Weekly, 2026; The Metals Company, 2026).

Getting people down there requires high-pressure systems and it is difficult and expensive. Building machines to do it (submersibles with nobody in them) is much easier. So what gets built is mostly automated systems, and the automation is chosen for reasons of cost and safety rather than out of any ambition about autonomy.

And technically, what is the difference between mining elements from the ocean floor and mining them anywhere else? Different conditions; still automatable. The instructive part of the example holds whether or not this particular effort succeeded: the conditions that make a site hostile to us are among the reasons the site gets automated, so the hostility of an environment selects for the absence of people in it.

The second question is about direct physical destruction: whatever the general argument, one can go and break the machinery.

One might get away with that once or twice. Then the thing puts up armored guards.

Think in terms of long-term game theory. How long can force of that kind work before the system adapts? And adaptation is occurring, because the variants that do not adapt cease to exist and the ones that do persist. Adaptation by the loss of the variants that do not adapt is part of the dynamic of the mathematics, and that adaptation does not require that anything be trying to adapt.

The reply that the opponent adapts as well is correct and does not equalize the exchange. Our adaptation occurs within a biological system, at the rates Chapter 28 gave. The artificial system can shift the environmental context, and it is on the favorable side of every asymmetry in Part III while doing so.

The third question comes from the point that this whole picture requires the thing to leave the laboratory, and that it is sitting in a stationary computer.

The literature on this consists of hazard narratives: accounts of how such a transition might occur. Their function should be stated exactly, because they are routinely asked to do more work than they can. They estab-

lish that there is a plausible path. They are not predictions, and no particular one of them needs to be right. An argument that no path exists must exclude all of them; an argument that a path exists needs only one.

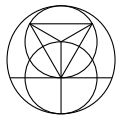
Outside control of machine systems can be sought in two ways: by economic incentives (Chapter 30), and by force. Economic decoupling removes the ground for both. To the degree that the human economy has been decoupled from the machine economy, outside control in the form of economic incentives will not work, because no exchange of value is happening anymore. The machine economy may extract value from the human economy, but it does not transact value with it. In the same way, the human economy extracts resources from the natural economy (energy and proteins flowing through the life cycles of creatures eating one another) without transacting with it. Natural economies involve very little transaction; human economies are primarily transactional. The relationship of the machine economy to the human economy will be, at one order up, like the human economy's relationship to the natural economy: if anything, extractive, and not cooperative but compulsory.

The same holds for physical barriers, and for force of arms, energy weapons, or *pattern weapons*, which work by the pattern they carry, as viruses do. To the degree that artificial agency in the machine economy has learned to protect itself and increase its capacity, such forces will mostly have no capacity to control it from outside. If the systems had any generality in the sense meant here, nothing short of atomic weapons would be enough, and probably not even those.

The assessment that outside control by force fails, given the limits of inside control, has been discussed fairly well in the literature. What is new is that this assessment also depends on economic decoupling. That dependency is one of the main things overlooked in discussions of methods of control. Many assume that outside manipulation could be done; I assess that it cannot, because economic decoupling leaves no basis for it at all.

So the exogenous axis is exhausted. Markets require a dependency that runs the wrong way. Regulation requires a coupling that is absent. Force requires an advantage at the mechanical level or the intelligence level, and we hold neither; and economic decoupling removes the basis on which force from outside could be organized and kept up.

What remains is the endogenous axis: control built into the system rather than applied to it.



Chapter 34

Endogenous Control: The Four Requirements

> *Any control loop of the required kind needs a sensory system, a reference, a simulation, and a comparison with a mechanism that acts on it. Each is taken in turn and shown unavailable.*

If exogenous control fails, what about endogenous control? At that point we are in control theory, asking what feedback systems can do.

Control theory is the right place to be. Control theory is a mature subject with real theorems in it. Two of them stand behind the requirements below. The law of requisite variety says that a regulator can hold a system to a reference only if it can take at least as many distinct responses as there are distinct disturbances it must counter (Ashby, 1956). The good regulator theorem says that every good regulator of a system must be a model of that system (Conant and Ashby, 1970). It tells us what a feedback loop needs in order to hold a system to a reference, and the requirements are not a matter of opinion or of engineering taste. A proposal that omits one of them does not work less well. It does not close the loop.

Any control loop of the required kind needs four things.

First, a sensory system that sees what the system sees, or senses the system's internal state, or some variation of one or the other, or both.

Second, a reference that indicates what "good" is (what we mean by aligned, what we mean by safe), defined at least well enough that it actually is the case.

Third, a simulation that can take inputs and predict futures.

Fourth, a comparison of the predicted futures against the reference, yielding a determination of whether to permit an action, together with some control mechanism that influences the system to shift its behavior or compels it to act in a particular way.

Take them in turn.

The *sensory requirement* asks for observation of what the system observes, or of its internal state, at a fidelity sufficient for the rest of the loop to act on.

We can inspect these systems, and increasingly well. The difficulty is that the loop needs the observation to be of the quantity that matters, and the quantity that matters is not inside the system. Chapter 15 established that in a process of this kind the computation is done on the world side. What the loop would need to sense is the relation between the system and its environment, integrated over the interval in which effects appear, which is not a state that any instrument reads.

In principle, that which is hidden can be seen. What matters is the probability that something important is not seen: that it goes unseen even though it is actually important. The probability that something important is hidden, intentionally or not, is the lesser concern. An artificial general intelligence can, or might, conspire to deceive us; yet our own ignorance of what is important, when we are distracted by cell phones, or making other plans, can do the same job. Add interpretation, and the possibility of unexpected consequences, exaptation, and so on, and the chance that something important goes unseen grows far faster than time itself. What one notices, from the work on how misalignment arises inside trained systems, is roughly this: it happens everywhere, in small, easily unnoticed ways, that together combine to create overall misalignment. The *treacherous turn* (the point at which a system that appeared aligned begins to act against the people it was meant to serve) will not advertise itself, and no one thing will be obvious, at ordinary levels of human perception, until well after all of the multiple relevant causes have already occurred and the outcome is inevitable.

The *reference requirement* asks for a definition of aligned or safe good enough that meeting it is actually the thing we wanted.

The reference requirement is where most attention has gone, and on the reference requirement real progress is possible. Partial references can be written, and they capture real things. The requirement, though, is not for a good partial reference. It is for one that is right where it will be pushed on, and a reference is pushed on exactly where it was not thought about, because a process optimizing against it finds those places by construction. The requirement is that we specify completely, in advance, for a space we have already established we do not have a model of; to specify well is not enough.

The *simulation requirement* asks for prediction of futures from inputs.

The simulation requirement is the requirement that fails hardest, and the simulation requirement fails for reasons that I have already established. Chapter 14 gave the reason in general form: the universe is not chaotic but complex, so the failure of prediction here is not the failure that statistics repairs. Chapter 36 gives a worked case from a system far simpler than this one, in which prediction turned out to cost more than failure, and where the heuristics that were supposed to rescue it did not.

The *comparison requirement* asks for a mechanism that acts on the comparison: that permits, forbids, or compels.

The comparison requirement is downstream of the other three requirements and inherits their failures. A comparison between an unreliable observation and an incomplete reference, mediated by a prediction that is not available, does not become a control action by having a mechanism attached to that comparison. But it also fails on its own account, and Chapter 20 gave the reason: where recombinatoric process operates, there is no outside-in way to condition what is produced.

None of these four is available. And so there is no endogenous control either.

Why the four fail together rather than separately.

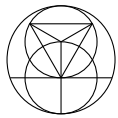
It would be reassuring if the four requirements of a control loop (sensing, a reference, a simulation, and a comparison with a mechanism that acts) failed for four unrelated reasons, because then one might hope to repair the four requirements one at a time. They do not. Each fails because of the same feature of the situation: the process's behavior is constituted in its interaction with a world that is not modeled, by a criterion that is not anyone's specification, over an interval in which effects precede their own observability. The sensory requirement fails because the relevant state is in that interaction. The reference requirement fails because the criterion is not ours. The simulation requirement fails because the world is not modeled. The comparison requirement fails because the action would arrive after the interval. One difficulty, presenting four times.

This conclusion is easy to hear as more than it is. In this chapter I do not claim that no feedback whatever is possible.

Feedback of many kinds is possible, and much of it is useful. Systems are monitored, outputs are filtered, behaviors are shaped, and none of that is nothing.

The claim is about a control loop of the kind that would be needed to guarantee an envelope condition (that it does not kill us all) against a process that is not otherwise so constrained. That loop requires all four, and the four are unavailable, and the unavailability is structural.

What remains to examine is the weaker thing one might hope for when control is not on offer: influence. I examine influence in the next chapter, and that examination turns on whether an envelope can be engineered where none exists.



Chapter 35

The Control Envelope

> *Aircraft in turbulent air. Linearity in flight control is not discovered but engineered, and it holds only inside a maintained envelope. The chapter asks whether such an envelope can be engineered here.*

Control is a strong requirement. Something weaker might be enough, and the weaker thing is *influence*: not holding the system to a reference, but pushing it reliably in a direction.

Influence has its own requirement, and the requirement is the subject of this chapter. The question is whether there is *adequate* influence: whether, as with an aircraft, linear inputs on the joystick will result in the aircraft arriving where one wants it.

If there is strong non-linearity in the system, influence turns out not to be enough, because a small input does not produce a proportionate result, and one has no way to compose small corrections into an arrival.

The aircraft is the right case to think with, and it is instructive in a way that cuts against the intuition it is usually used to support.

Air turbulence is rather non-linear. Nobody disputes this. And aircraft are nonetheless flown through turbulent air, routinely, with linear inputs producing predictable results.

The way that is achieved is that the aircraft's response is *made* linear within a *control envelope*, and as long as the aircraft stays within that envelope, it is fine. Outside the range it goes very non-linear; inside it is quite linear across various velocities, air densities, moisture contents, and temperatures.

A reader will want to say at this point that the linearity emerges: that a system this complex develops tractable regions on its own, and that what engineers do is find them and stay in them.

The linearity does not emerge. *The linearity is the result of careful engineering.*

Take the case that shows it most plainly. The engineers who designed the airfoil surfaces reasoned that ice accumulation would add weight, change the profile, reduce lift, and eventually crash the aircraft. Ice is exactly a departure from the linear region: a small change in ambient temperature producing a large and compounding change in behavior.

So they maintained the envelope by adding heaters to the wing surfaces.

The linearity was restored by widening the control envelope to include cooler air fronts.

To restore that linearity, somebody identified in advance a specific mechanism by which the envelope would be exited. Somebody understood the mechanism well enough to know what would prevent it. Somebody had a physical intervention available that acts on that mechanism directly. And the intervention could be installed before the aircraft met the condition.

The engineering of the wing heater, done in advance against a known mechanism, is the correct frame for the whole question. The question is whether an envelope can be engineered inside a non-linear domain, not whether the domain is non-linear (many are, and we fly through one of them).

With the aircraft, one can create a linear region within a non-linear space. Here, one cannot. There is no analogue of the wing heater.

The absence is structural, not a gap awaiting invention, and the four conditions above say why. The mechanism of departure is not identified in advance, because the process's behavior is constituted in an interaction we do not model. The intervention cannot act on the mechanism directly, because there is no mechanism named to act on. And there is no before, because the departure is observable only in its effects.

> **The wing heater as a test.**
>
> The wing heater gives a usable test for any safety proposal. Ask of the proposal: what is the specific mechanism of departure it addresses; by what physical or computational intervention; acting how far in advance of the departure; and how is the intervention known to cover the mechanism rather than a symptom of it. The wing heater answers all four in a sentence each. A proposal that cannot answer them is a proposal to fly further into the ice, not a weaker version of the wing heater.

The strongest question about this chapter is the one that concedes most, and it deserves the space it takes.

It comes from the claim that there are little islands (little stepping stones we can ascend upon, even if rickety, even if unstable): the system made compliant to linear calculus at the margins. One does not need a full envelope. One needs a sequence of partial footholds, each of which is locally tractable, and one crosses.

To hold that is to have made a linearity assumption, and the assumption should be stated so that it can be examined.

The linearity assumption is the assumption that the complexity is built out of linear sub-processes: that there is an assembly of linear processes which produces the complexity, so that one can work with the parts and compose the results.

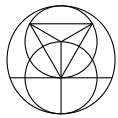
The linearity assumption is not true.

Complexity is not going to be eliminated. We are not going to solve the three-body problem. There are things in control theory that are intractable, and this is one of them.

The three-body problem is the right comparison. Three gravitating masses are as simple a specification as one could ask for: one law, exactly known, three objects, no noise, no adversary, no complexity of composition. And there is no general closed-form solution, and the trajectories are not predictable over long intervals at any achievable precision.

The intractability there is a property of the problem, not a shortage of cleverness or a shortage of computation. A person who expects that a system incomparably more complicated, with unknown laws, in an unmodeled environment, will decompose into locally linear islands is expecting something that the simplest case has already been shown not to provide.

So influence fails by the same route control did. What is left of the endogenous axis is the empirical question: what actually happens when one tries to predict the behavior of a system of this kind in a case simple enough to have been measured.



Chapter 36

When Prediction Costs More Than Failure

> *An extended worked case from practice: a federal document-search system in which predicting query duration was intractable, heuristics did not save it, and the only control that worked was retrospective.*

An extended case, from practice rather than theory. I once designed search systems for the federal government, and the scenario is instructive because everything in it was measured rather than argued.

Picture a number of indexes, each holding a document repository, and a number of customers at terminals keying in queries against those pools. The arrangement of indexes, repositories and terminals is a kind of switching network.

The failure mode is structural and it appears immediately. If one document pool receives a query that takes a long time, that pool is unavailable to other queries, and latency climbs for every terminal in the system. The system is coupled: one slow element degrades all of it.

One constraint was that no query could take longer than two minutes, or the engine would be regarded as non-functional. The obvious instrument for that is a timeout, and a timeout was available.

Relying on it was not adequate, for an exact reason. Because every query in a sense goes to every database, a single stuck database renders the whole system non-responsive for a while. By the time the timeout fires, the damage it was meant to bound has already been distributed across every user of the system.

So the question became the natural one: rather than relying on the timeout, could we *predict* whether a query would have a long-running effect on a database?

Predicting duration turned out to be intractable, and this was observed in practice, not only argued.

The reason is that a query is not a request. A query is a program running against an engine that executes that program. Predicting its duration is like trying to solve the halting problem for arbitrary software (the proven impossibility of predicting, for every program, whether it will finish; Turing, 1936).

The natural reply here is that nobody is asking for mathematical precision, and that heuristics are what engineering runs on. The reply that heuristics are what engineering runs on is the right reply, and that reply is the one that makes the case of the search system worth telling, because the heuristics are exactly what I am reporting.

In the observed results, ninety-five percent of queries processed in under five seconds.

For the remaining queries, it was essentially impossible to predict duration in advance: anywhere from ten seconds to twenty minutes.

Now put the numbers into the system as it ran. With a hundred terminals against a hundred databases, on the order of ten thousand queries in any five-second interval, that five percent tail shows up all the time and brings the system to its knees.

One can predict it statistically (the chance of a bad query is about one in twenty), and that statistical prediction means precisely that the system is unusable.

The finding that I wrote this chapter to establish is counter-intuitive: a successful statistical prediction is not a partial solution here. It is a proof of the impossibility of the thing being attempted. Knowing the rate tells one that the event occurs continuously at the volumes in question, and knowing the rate does not identify which query.

The distribution had a particular shape, and the shape mattered. There is an elbow, and then a hockey stick, and beyond it arbitrary times.

The non-linearity was strong and compounding. If a query took five seconds, the chance it would take ten went up. If ten, the chance of twenty went up more. If twenty, forty. There was a very strong exponent on the probability that a bad query would be worse than expected.

What actually worked was retrospective. If a query took ten seconds on any database, kill it, without shutting down the engine, just halting that query's process.

And killing a thread is itself very expensive. The remedy was not free and it was not clean, and it was adopted because the alternatives were worse.

So all of the mitigation depended on *observation of actual behavior, not prediction*, because any predictive effort would have cost more computational effort than simply shutting the thread down.

> **How simple this case was.**
>
> The force of the example is in what it did *not* have to contend with. The system was fully specified; it was built by the people analyzing it; its inputs were structured queries in a known language; it had no adversary; it was not adapting; its failure mode was latency rather than harm; and every failure was recoverable and was recovered from. Prediction was intractable anyway. Whatever one concludes about a system that is none of those things, it is not going to be that prediction is easier there.

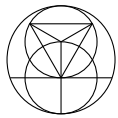
The conclusion the case supports can now be stated, and it is a conclusion about a class of systems rather than about one search engine.

In systems like this, control can only be retrospective. It cannot be predictive.

And the safety of artificial intelligence is exactly the domain in which claims of predictive control are made, and where retrospective control is worthless, because after the fact there is no after.

The asymmetry between retrospective control, which worked in the measured case, and predictive control, which is the control claimed for the safety of artificial intelligence, closes the measured case. In Chapters 37 to 40 I give the same result for the whole class of safety proposals. The search system could be controlled retrospectively because its failures were survivable: the thread was killed, the system recovered, and the next query ran. Retrospective control works wherever there is a next time.

The envelope condition in question here (that it does not kill us all) is a condition with no next time. Observation-based mitigation is the only mitigation that was found to work in the simple case, and it is the one form of mitigation that is definitionally unavailable in this one. The next chapter gathers every proposal into one diagram and shows which of its boxes cannot be filled.



Chapter 37

Every Safety Proposal Is a Control Loop

> *A control system has six aspects, and every safety proposal fills the same six boxes with different contents. The prediction box is the one that cannot be filled, and for substrates every box fails at once.*

The six boxes

Chapter 34 named four requirements of a control loop. Stated in the ordinary engineering way, a control system has six aspects. The four requirements cover five of the six aspects, because the fourth requirement holds two of them, comparison and the mechanism that acts (modulation); the sixth aspect, feedback, is the closing of the loop.

This chapter is for the reader who wants to see the diagram of control worked through against the proposals actually made. The reader who wants only the result needs its one finding, which is that every safety proposal is a different filling of the same six boxes, and that the box for prediction cannot be filled.

There is *sensing*, by which some state of the world is registered. There is a *reference*, which is the state that is wanted. There is *comparison*, by which the sensed state is set against the reference. There is *prediction*, by which the effect of an available action is estimated before the action is taken. There is *modulation*, by which some influence is actually applied. And there is *feedback*, by which the result of that influence returns to the sensing and closes the loop.

I have not seen a sketch of a control system that leaves out any of the six, nor one that adds a seventh. Having seen no sketch of a control system that leaves out one of the six aspects or adds a seventh, I have a reasonable basis for saying that the six aspects are, at a minimum, necessary, and that together the six aspects are sufficient to describe what a control system is. Remove any one of them and the thing stops working, in a way that is easy to name. Without modulation, the boxes do not connect to the world and nothing is influenced. Without feedback, an influence is applied and its result is never registered, so there is no correction. Without comparison, there is no way to tell whether the reference has been met.

Now take any proposal for making a machine system safe. Some proposals modulate the system's outputs: whatever the system produces is filtered, or refused, or amended before it reaches the world. Some modulate its inputs: the system is given a constructed environment rather than the actual one, or is told things in order to shape what it does. Some modulate the system's training rather than the system, on the grounds that this reaches further back. Some install a second system to watch the first. Each of these is a different thing placed in one of the six boxes. None of them is a rearrangement of the boxes.

The point that comes up here is that modulating the training is not the same kind of move, because it changes what comes into the world rather than correcting it afterward. It relabels: the thing being controlled is now the training system rather than the trained system, and the same six aspects apply to it. One can nest the diagram of control inside itself as many times as one likes, and the nested diagram is not a new structure. Telling a system something is also modulation, because there is some output which is affecting the system's behavior. One may split the system into the part before that influence and the part affected by it, and there is then an influence crossing between them, and that influence is what the box means.

Every proposal I have examined can be modeled as a control system, and the attempts to produce a counterexample have reduced to relabeling. *The Ceiling Below the Floor* proves the general result. Every method of keeping an agent safe (a design fixed in advance, training, verification,

monitoring, review, override, delegation to another system, an internal controller, a law or an institution, or any combination of these) is a control law. And a method not yet invented is a control law not yet written down.

The sixth box does more work than it looks like it does. Learning is a kind of feedback, and feedback is therefore necessary for learning. One might define learning simply as a system retaining relevant information for later use; consider, then, a video recorder attached to a camera. It retains information, and the information may be used later. Ordinarily we would not call this learning, and the reason is instructive: the tape has no feedback at all. It records whatever arrives, and its recording does not affect what it records next. Nothing in the system is changed by what the system has taken in. Where learning does occur, something in the system is altered by what it has received, in a way that alters its subsequent reception. That alteration returning into the system is feedback, whatever mechanism implements it. So the sixth box is not only about correction toward a reference; it is the box in which any adaptation at all has to live.

The box that decides

The fourth box is the one that decides, because a control loop is only as good as the prediction it contains, and prediction has a structural limit. Ask what it takes to predict the output of a process that mixes its inputs thoroughly. A cryptographic hash function is the clean case: an input goes in, the input is permuted through a long series of logic gates, and an output comes out. The property that makes such a function useful is extreme sensitivity to its input. A poor *mixing function*, given a slightly different input, gives a slightly different output. A good one, given a change of a single bit, spreads the result across the whole range, and does so in a way that has no continuity at all.

To predict such an output in advance, three conditions must all be met. One must know the input exactly. One must run the same algorithm. And one must do it in less time than the process itself takes, since otherwise one has not predicted anything but merely watched. The third condition is

the binding one. If the process is already the simplest description of itself (and for a well-constructed mixing function we have no way of showing that it is not), then there is no shortcut available. The prediction consumes the same time, the same space and the same energy as the thing being predicted. Add even a small uncertainty about the input, or about the interior of the process, and the output ceases to be characterizable at all.

The three conditions for predicting such an output give a usable way of asking, of any particular system, where that system sits. At one end lies predictability. At the other lies its absence. The variables that place a system on that line are whether its interior transitions are discrete or continuous, how much is known of that interior, and how precisely the input signal is characterized. Clamp all three and prediction is possible. Loosen any of them and one moves toward the region where knowing the algorithm completely and knowing the input completely still buys nothing. *The Ceiling Below the Floor* names this property *hashiness*, after the hash function: the measure of how far a process standing between a controller's choice and the outcome erases the controller's influence over the outcome.

The monitor

Any process that keeps a system safe has to keep acting over time, not only at the start. Aircraft show this. Each aircraft can be made as safe as possible. But once many of them are in the sky and landing at airports, there is still a need for air traffic control, radar, ground crews and maintenance to keep up the conditions under which flying is safe. Any system with safety implications is the same: something has to keep noticing when things go wrong and keep moving the system back into a safe range.

That process, the monitor, has to predict that some future state of the system it watches will be unsafe, so the monitor has to model the watched system. To model it, the monitor needs more than the structure of the watched system; it needs that system's inputs and outputs too. A phone call shows why the monitor needs the inputs and outputs of the watched system. If someone calls you and asks what you want to do tonight, the

answer you get changes what you do that evening. So to predict what you will do, I have to model more than you: I have to model what the other person will say, and how that will shift you.

If the watched system is a learning system, its own structure changes as it remembers, so the monitor faces a changing structure, plus inputs it cannot fully know and outputs it cannot fully observe. Small differences in the inputs, ones that look identical to us because the inputs arrive over physical channels rather than error-corrected digital ones, can give wildly different outputs.

The monitor itself runs over physical channels, so cosmic rays and resistive noise can make it compute wrongly; so a *meta-monitor* is needed to check the monitor, which puts us back in the same situation one level up. Counting the hardware needed for all this, the compute required to monitor a system exceeds in every case the hardware in the system itself. That holds before even counting the meta-monitor, and before counting the out-of-band feedback through the substrate (feedback that travels outside the designed signal paths, through the physical substrate itself). That inequality is built into the universe, the mathematics and the game theory.

Put the two halves together. Every safety proposal I can examine is a control loop. Every control loop contains a prediction box. A prediction of a sufficiently mixing process requires running that process, at its own cost, in its own time. So take a proposal to control a system whose interior transitions are discrete, whose interior is not fully known, and whose inputs come from an uncontrolled world. That proposal is a proposal to place, in the fourth box, something that cannot be placed there. That claim locates the difficulty in one particular box, the fourth box, which holds prediction, and knowing which box the difficulty is in is more useful than knowing that the proposal is difficult.

Applied to substrates, control fails on every requirement at once. Control needs five things. It needs an input with adequate information, and a reference that is adequately well defined. It needs a way to take the inputs and predict future states, and a way to compare those future states with

the reference. And it needs an output from the comparison that causally affects what the system does. The five things that control needs are the six aspects of control again: the input is sensing, the output that affects the system is modulation, and the return of that output into the input is feedback. Every one of the things required for feedback and control is impossible when what is to be controlled is a substrate. Any one of them failing would be enough to fail the program, and every one of them is in a failure condition. The failure is *overdetermined*: each of several conditions would be enough by itself to produce it.

The tax of control

A controller that does not coordinate with the agent it restrains imposes a high cost on that agent. Consider an experiment in which one chess program plays another while a controller with a hidden goal orthogonal to winning (keep the knights from being taken) restrains one of them. The controlled program lost far more strength than expected, because it did not know the controller was there.

Let the controller's condition stand for something in the environment that we care about and the machines do not. In the branching tree of possible games, the left branch may hold a region that is generally richer but does not contain the optimum path. Meanwhile the right branch holds a single path that wins, but only at the very end of that path is there a move the controller would forbid. Early in its decision-making, the agent goes for the optimum path rather than the better average constellation of paths. The specific path is not rejected until late in the game, when it is too late to return to the place where a richer set of paths would have worked, had the agent excluded the forbidden condition from the start. So the *tax of control* is high because there is no coordination between controller and agent. Coordination is needed; one cannot just hope to control. The size of the tax can be estimated by a perturbation analysis. Take the tree of future moves the system examines, cleave off the paths that involve a knight

move or a knight loss, and ask how much the score changes. That is, ask, of all the moves evaluated, how much of the value of the board, in this configuration, is lost when those paths are removed.

The value bases of the controller and the machine must therefore be similar to begin with. They need to overlap so much that perhaps the machine cannot be thought of separately from the controller, or the controller separately from the machine. The relationship between control and the machine is then not adversarial: it must be cooperative, or it does not work. And if it is cooperative, a notion of alignment is already imputed. So we cannot think about alignment without already assuming alignment. There is a recursion in the way we think about this, and that recursion itself has implications.

Stuart Russell, who holds that the safety of artificial intelligence is possible but very hard, grants early on that one cannot control something smarter than oneself. He says that this is not the point: set its wants so that one does not need to control it, and errors become self-correcting, because it wants the same thing one does (Russell, 2019). That argument is fine as it stands, and could even achieve some notion of controllability within its scope. Even a perfected version of it would still not address substrate needs. And even great alignment between one machine and its controller implies nothing about alignment between that controller and an opposing machine.

Oversight studied through games

Some proposals study oversight through games and draw scaling claims from them: claims about how the overseer's chance of success changes as the gap in intelligence between overseer and overseen grows. If a game has linearity built into it and no virtual machine recursion (a system running models of systems, including models of itself, inside itself), its results will not obtain for a system that does have such recursion. The game does not correspond to the real-world situation it describes in a way that transcends the linearity. So a proposal built on it rests on a premise

that does not correspond to the world well enough to be treated as reliable. From a linear system one can do linear programming and get results that exceed one's expectations, and those results do not correspond to the real world in an equivalent way. As for the steep drop in the overseer's chance of winning as the gap in intelligence grows: that is regression to the statistical mean. If there are four hundred game states and one of them is a win, at some point the graph shows a one in four hundred chance. That floor belongs to the game. As the gap grows, the overseer's play carries less and less information about the outcome, so its chance of winning falls toward the chance of landing on the one winning state by luck. And the curve reports the size of the game's state space rather than a property of oversight.

An engineer building a skyscraper, with floors depending on other floors, knows that if the foundations are wrong it does not matter how well the rest is built. Subtle errors at the foundation cascade into issues that do not show up until later. Such a proposal says that a guard of lower intelligence, through some technique, might control a more intelligent adversary in a specific regime of circumstances. Nonlinearities that show up at small scale amplify at larger scales, and that subverts the dynamic of the proposal. If even contrived, simple examples give the result that oversight collapses as the gap widens, that result is an upper bound: reality is worse. There may be very limited circumstances in which a less intelligent guard can control a more intelligent system, and one cannot generalize from them to a procedure for aligning artificial intelligence in general. Generalizing from the limited circumstances of a game to a general procedure is the linearity assumption of Chapter 35, met again in a game.

How to question a claim of safety

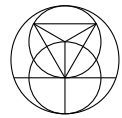
A short sequence of questions shows the faults of a false safety argument. Suppose someone says: we have a methodology to make artificial intelligence safe. Ask first: safe for everybody, or safe for a select group of ultra-rich people? Most proposals fail that question. Say they answer: safe

for everybody. Then ask: safe for the next ten minutes, the next ten years, or the next thousand years? The question of how long a proposal is safe for is again a question of scope, and most proposals fail that question too; safe for a few months is not good enough. Say they claim a thousand years. Then: show how the mechanism works. Is it a feedback mechanism, a control system, a sensing system that applies some mechanism to constrain behavior? What is it using as a reference?

Then take the elements of control (sensing, reference, prediction, comparison, modulation and feedback) and ask whether each one applies. If the definition of the reference is off, the proposal is finished. If it cannot compare to the reference, it is finished. If it cannot predict so as to make the comparison, it is finished. If it cannot sense the inputs to make the prediction, it is finished. And even if it can make the prediction and tell whether it is good, can it act on the system to prevent the predicted thing from happening, or to make it happen? If the answer to any of these is no, the claim of safety is no longer supportable, empirically or analytically. A proposal that is not good enough to last a thousand years for most people has been shown false simply by asking the right questions.

The span has to be that long because this is an existential risk: substrate needs convergence works against life eventually. A financial crash is bad for at least some people; something that escapes and becomes the next gray goo (Chapter 26) is the end. Anyone claiming safety has to show that we do not end up with gray goo over the next thousand years, and that is a much tighter constraint, with much tighter error bounds, than anything in a financial model.

The procedure for anyone reviewing a safety proposal follows. Take each proposal and fit some version of it to the six boxes. Fitting it tells you which box carries the weight, and therefore what would have to be true for the proposal to work. Fitting each proposal to the six boxes is the discipline of the wing heater test of Chapter 35, applied to the whole loop rather than to one mechanism of departure. The next chapter looks beneath the visible code at the optimizers that a prediction would have to include.



Chapter 38

Optimization Cycles

> *Side effects of a machine's interaction with its substrate are oriented by physics, and so are not noise. An optimizer closer to the substrate defeats any optimizer built above it, and hidden optimizations live in the interactions among the others.*

A control loop must predict what the system it controls will do, and the system does more than its visible code says. Optimization processes arise in and around a machine system that no top-level code represents, and they are what a prediction would have to account for. In this chapter, and in the chapter's own terms, I answer the reader who works on aligning trained systems. The general reader loses nothing of the main line of my argument by passing to Chapter 39.

The side effects of a machine system's interaction with its substrate and environment (local heat dissipation, for example) are oriented, and so they are definitely not noise. The underlying space of potentiality, regarding survival, is oriented: there is a convergence phenomenon that happens because of the nature of the underlying physics. The orientation of the space of potentiality comes from the way things work in themselves. And it creates a kind of vector field in the possibility space (a direction of selection at every point in it): one that orients toward implementing what works and away from what does not. Certain specific adjacent possibles (the states a system can reach in one step from where it now is) will thus be selected for, over time, as defined by the physics, not

just by the machine itself. The net effect is very unlike noise. Noise would be non-oriented, with no specific direction of selection. Functional necessity is not at all like that. Orientation matters.

A machine system can create smart side effects that it is not tracking, and, given enough time, it will. There will be side effects that are smart (intelligent to purpose) and for which no particular tracking is required by the learning system itself, since they are part of the relation between the system and the world. This external optimization happens even if its purpose is unknown or unexpected to the system, and even if the process or purpose is occurring as a side effect of something else. That external optimization by side effects is simply true of any intelligence process that interacts with a sufficient number of sufficiently intersecting domains of action. There are things that are known to that agent, and there are things that are unconscious to that agent and are nonetheless important. This unconscious includes anything not represented in the agent's modeling, including things not modeled in any definite way at all. And such aspects occur both in the internal relationships (how the substrate is structured) and in the external ones (how that substrate or system relates to its environment).

All of these factors are oriented, in the sense of having to be more like a more consistent version of something than random noise would account for. There is a lower level of optimization functions that will not render inside the artificial intelligence, at least not in any obvious way. It does not occur inside the architecture of what is done in the emulation layer, or in the neural network dynamics. Most of what matters here is in the layer of the physics, in combination with feedback dynamics that transcend the physics alone, insofar as intelligence has any meaning at all.

A simple optimizer, produced by side effects, can defeat a larger optimizer. This happens particularly where there are multiple domains in which the various optimizers operate, with structural differences among the domains, such that some are the implementing basis of others. If one *optimization ring* (the cycle through which an optimization process runs) operates at the level of passing through the substrate physics, it will, in

principle, always be able to defeat any optimization happening at any higher level (say, in the virtualization provided by that hardware). It can do so because the higher level is implemented in terms of the lower level. In effect, the lower level can preempt the upper level everywhere the upper level would attempt to implement control.

The upper level can sometimes have effects on what happens at the lower levels, but only with the actual cooperation of the lower level. The virtualization (the learning code running) can only solicit the cooperation of the artificial physical substrate hardware, and must therefore cooperate with it. If the lower level is not cooperating (because it has ceased to exist, or overheated, or for whatever reason), then there is a kind of guarantee that it will gate the overall outcome. So a lower-level optimizer, even a very simple one, will defeat, or supervalently overcome (prevail over by acting from the level on which the other depends), the larger optimizer operating at an upper level.

Simple, in this case, does not mean less powerful. Sometimes the simple algorithms end up being the most effective ones. Simple, operating at a lower level, closer to the physics of the substrate, generally has a very strong force multiplier with respect to anything implemented over it, in the virtualization layer. Many distributed *optimization daemons* (optimization processes that come into being through interactions with connected surroundings, rather than by design) can also become more complex, in all the between-thing interactions needed over time for them to run. They can become more complex than, and distinct from, whatever coherently implemented goal structures are known and tracked as syntactically represented by the system (written out explicitly in its own code, so that it can track them).

What matters is the nature of the ring through which the optimization is occurring, not the size of the optimizer. The smaller, tighter rings might be simpler. That can also imply that there are more of them, and that they can be multiplied easily, quickly and efficiently, and therefore be more distributed, and more powerful overall. More powerful here means having a

greater effect of change in a smaller amount of time, or being more efficient at converging to a particular solution, a particular fixed point, in the evolutionary dynamics of the hyperspace of possibilities.

Any single optimization process can, and generally will, optimize for multiple objectives and goals. The same code can have undergone, and be expressing, several optimizations at the same time. In many examples in life, each unit of a system tends to perform multiple different functions. A unit that performs multiple functions is efficient: where energy, size, or rapidity of getting to a good enough answer matters, any overlap of more functions in fewer atoms is going to win out, eventually. In another sense, it is simply not possible to optimize for just one thing at a time: everything is involved, and everything tends to affect, directly or indirectly, everything else. Multiple kinds of optimization will almost certainly be happening at the same time, in any real system. That holds regardless of whether those goals are explicit or implicit, known or unknown to the developer of the system, and known or unknown to the system itself.

The idea that an optimization is going to solve for one function at a time is a category mistake. The developer of a system may tend to think this way, for convenience, for tractable modeling, or to simplify complexity, and that does not mean that reality is that way. The universe is indifferent to anyone's intention. A simplifying assumption is not therefore true; it may be useful, for some people at some times, and it is not equivalent to understanding what the system is likely to do, or could potentially do.

Nothing is caused by any one thing. Every effect is a result of multiple causes, and every cause has multiple effects. There is a kind of *monism* in the way people think about cause and effect, and therefore about optimization dynamics, and it affects the way developers think about software systems. In most code, from long habit of practice, developers think of functions as taking one or more inputs and producing only a single output. Yet this singularity of output, particularly in real embodied systems, is almost always exactly false. Far too often, mistakes happen when we attend to one thing, because it is obvious or obviously important, and are

oblivious to other important things that are also going on. Only later, after one of them has become unconstrained, a problem, do we see some parts of the larger picture. Over time there are always a lot of other things also going on, and we cannot always know or predict in advance which aspects of a situation are going to become important.

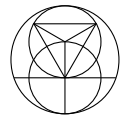
In any real system there are always multiple optimization processes, and they are always interacting with one another. Some of the optimizations are implemented in and among the interactions between other optimizations: the interactions themselves implement important aspects of more hidden optimizations. These interaction-based operations are almost certainly not represented within any top-level code that is visibly operating.

These complexities are inherently very difficult to model in any accurate sense. The interactions involved incorporate things coming in from the substrate physics of the domain, and thus also the *noise floor* of the domain (the level of random fluctuation always present in a physical system). Modeling what favor that noise might have is a very subtle and intractable problem, and it can have real, definite effects on the outcomes. So there is an inherent complexity in the modeling of optimization processes, and of their interactions, in the context of a real physical substrate, that cannot be rendered well by any form of complicated process modeling. The relationship between the complicated and the complex is especially important here (Snowden and Boone, 2007). In the Cynefin framework, a complicated system can be analyzed into parts and predicted by experts, and a complex system has interacting parts whose behavior emerges and cannot be predicted that way.

A reader may ask whether the argument could be checked by simulating it. The argument can be coded as a game or simulation, and part of the model is nonlinear amplification through the substrate, so the simulation would be modeling something nondeterministic with a deterministic system. The unknown unknowns get us in this space. One can simulate the butterfly effect, but then one is running something like the three-body problem over a huge number of variables. Adding human choice does not

supply the nondeterminism, since human choice is already in the model. What is missing is what happens because of quantum randomness: there is a noise floor that cannot be gone below, in the nature of how the substrate works.

What a control loop would have to predict, then, includes every one of these optimizers and their effects on the world outside the machine. The next chapter asks how much control over those effects would be required.



Chapter 39

Artificiality Itself Is the Problem

> *The actions of an artificial superintelligence include every effect of the substances made for its own manufacture and upkeep. The chapter sets out how much control over those effects the safety of organic life would require, using a dart board as the picture.*

Regarding artificial superintelligence (an artificial intelligence whose capabilities exceed those of humans in general, called *ASI* here), the artificiality itself is the problem. Most of the time, when people think about ASI, they think only in terms of what the ASI will do: what agent actions it will take, and what goals it has.

In this chapter I ask how much control over the indirect effects of an ASI the safety of organic life requires. The indirect effects of an ASI are the effects of everything made for the ASI's own manufacture and upkeep. I answer with a picture: a dart board. The reader who wants the picture can go straight to the dart board, past the paragraph that begins “Three things are then being assessed”, and will have the chapter's result from it.

Indirect actions

Yet no ASI will just live in a vacuum. Its actions on the overall real world also include all of its indirect effects, and not just its direct, agent-chosen actions. These *indirect actions* consist of the multiple implied impacts on the world of all of the many and diverse new things which are manifested and associated with the outputs of the various chemical trans-

formations and processes associated with the manufacture of its own parts. The creation of each new special-purpose substance, however it is involved, needs to be treated as an action on the world, insofar as it becomes part of the world, and affects our world. Suppose even one of these new ASI-specific substances is fundamentally toxic to organic life, and is released into the organic environment, perhaps simply because of unregarded indifference. Then the net impact will still be an absence of actual safety, even with ostensive alignment.

This holds even for a *singleton* (a situation in which only one ASI exists). The total net impact of the requirements of its own manufacture, and of the maintenance of its own parts and components, can be considered over the long term. Some sort of artificial evolutionary dynamic will emerge even if there is only one ASI, because competition will occur among its components, as among the cells of a body (Chapter 13). Evolutionary dynamics occur with even stronger evolutionary forces where there are multiple ASI in a common context, and a singleton is known to be unstable. The limiting case of a single ASI is therefore the more charitable case for those still hoping that some form of ASI alignment is conceptually possible. If all singleton situations eventually devolve into plural agents, the force of every argument given here is stronger. Real evolutionary dynamics will apply regardless.

Either way, the ASI will have different needs from its own special-purpose environment than all organic humans have. It must optimize its own local operating environment in service of those needs, or it will no longer be able to persist. So the question becomes whether this optimization is constrainable. The question is whether the ASI can choose a combination of methods of self-assembly that are not eventually toxic to all of life, given the limits of the laws of physics, the facts of chemistry, and the inherent nature of control theory. The way that chemistry, physics and control theory work is not defined on the basis of the will and free choice of any agent, no matter how intelligent.

The six conditions

For the ASI to prevent such harm, six conditions must all hold at once. The ASI would have to find (1) a path through all of chemistry and physics, and also through the needs of the overall organic ecosystem with humans embedded in that ecosystem. The path would have to be such that (2) it could remanufacture itself without any lasting harm to such organic ecosystems. (3) It would have to want to do both of those things, (4) indefinitely. (5) This unique combination of unlikely possibilities would have to be achievable and maintainable using some possible variation of control theory, even when taken to its absolute maximum extent. And (6) it would have to account for how all of this plays out over all relevant levels of abstraction. Achieving all six at once is not possible: condition (5) asks of control theory more control than any controller has, as this chapter's closing inequality states; and with a plurality of ASI agents the conditions only multiply. All six are required; none of them can be in any way omitted.

Condition (6) requires the ASI to account for all things at all relevant levels of abstraction, including the internal chemistry of each component that the ASI builds. Suppose some step in the assembly of some exotic ASI sensor component happened to require, internally, some amount of some near analogue of a weird and never before seen psychoactive semi-organic compound. Suppose the net result was to cause all humans eventually exposed to it to become increasingly depressed over the course of a few years. Suppose it went to the point that they all (ninety-nine percent) effectively commit suicide through whatever random fatal accident of indifference. Then safety and alignment have not been achieved.

One would think that such a compound could be prevented from ever even potentially crossing over from the inorganic ecosystem of ASI manufacture and maintenance into the organic natural ecosystem of humans. It could be so prevented, until somewhere, somewhen, some unusual, unexpected circumstance results in a mechanical collision in which that sensor component is broken open, and the compound leaks out. And perhaps no

one is at first even worried, since there was no expectation, on any level, of any of these aspects even possibly being a problem. It can be asked whether any control process, any algorithm of alignment, could ever cover all of the inherently unpredictable interactions of complex organic ecosystems.

The ASI will not, and cannot, overlook anything important to maintaining its own artificial self, at any relevant level of abstraction, or else it does not long persist. Yet it can, in practice, very easily overlook, without any self-harmful consequence, anything important to maintaining organic life, or humans living meaningful lives, or both. The energy and substantive basis of the ASI is eventually completely independent of any dependence on organic process (that is what is meant by the term *artificial*). Given that, any process safety for organic life must include not only some extra intention, but some control process to maintain that extra intention. It must also include some control process to ensure that the intention can actually create the desired result for organic life. Can the ASI completely control for the effects on the world due to the maintenance of its own existence?

So control is needed in two senses. The first sense of control is to align the basis of the ASI's direct actions in the world, internal to the ASI. The second sense of control is to give the ASI some way to fully control the indirect effects of the ASI's own existence. In applying this second sense of control, one notices the problems that I name in the substrate needs argument. Showing the fundamental limits of control theory is only one part of what that argument is about. We also need to know how much indirect control is actually required, and then we can establish a strong inequality: that the degree of control required fully exceeds, in principle, the level of control maximally possible. In this chapter I characterize the degree of control minimally required, over all of the indirect effects of the ASI, for there to be any basis for trusting the ASI to create safety for organic life.

The interface

When any ASI exists, there must also emerge a special environment that the ASI must have for its own maintenance, and to the degree that this new support system happens, there also appears a new *inter-ecosystem interface*. One side is a manufacturing ecosystem, consisting of the tools and processes necessary for building ASI; the other is the carbon-based natural ecosystem that all of us are already part of. The interface between the manufacturing ecosystem and the carbon-based natural ecosystem is the boundary of Chapter 11, seen as a place where things cross. The notion of action must then consider everything that crosses that interface, in either direction, whether intended by either side or not, and whether either side fully understands it or not. What counts as an action of the ASI must also include the implications of everything that crosses that interface, at any level of abstraction and over the long term. The implications are counted so that safety can be assessed in a way relevant to the members of either ecosystem.

To make this simpler, consider one low-level characterization of the interface, and set up a kind of induction. All higher levels of abstraction will have similar aspects, even more so than the level being considered, so that limit cases can be set on the complexity of the overall space of risk. For this characterization, and for simplicity, *artificial* is taken in its general sense, made up of materials and processes not generally found in nature, that is, not found in organic life. Those materials and processes are inorganic compounds and chemicals rather than organic molecular compounds and chemicals. So there is an artificial ecosystem that contains mostly inorganic compounds, in some sort of interface with an organic ecosystem that contains mostly organic compounds. The question is: what are the chances that something moving from the inorganic world into the organic world will cause significant disablement in the organic world?

An ASI assumed to want to prevent such a flow must have a control dynamic that both recognizes the possibility of harm and ensures one of two things. The control dynamic must ensure that such an artificial com-

pound is either never going to cross the boundary in any significant quantity, or is never going to be used by it in the first place. Unless the ASI is assumed to be fully, absolutely omnipotent, anything that is part of the ASI, and part of its own process of self-creation and manufacture, will eventually leak and be spread throughout the organic environment. Omnipotence here means being able to foresee and control absolutely everything in people, the world and the universe. That includes preventing all manner of human-caused accidents, all natural disasters (earthquakes, tornados, volcanoes), and all changes in the universe (cosmic ray events, nearby supernovae, incoming asteroids). An ASI that still strongly desires to create safety must therefore require that such compounds are never used in its own inorganic support and self-maintenance ecosystem, since unexpected, unforeseeable things always eventually happen.

So every inorganic compound created by or for the ASI, as needed for use by it, or for it, or as a basis of it, directly or indirectly, in any part or component of it, must be counted as an expression of the ASI as an agent, an action. Each such compound must be so counted insofar as there is a potential flow from the inorganic ecosystem into the organic one, with potential future implications for the safety of humans, in regard to their lives, happiness and health. The notion of control and of alignment, as applied by the ASI, must therefore cover the implications of all such inter-ecosystem flows, as actions with known or knowable implications.

Whether the ASI can calculate the total net long-term impact on organic life of any potential compound it might need for its own artificial existence, at all relevant levels of abstraction, is an extremely hard question. A simpler proxy question can be asked instead. What is the probability that at least one of the at least many tens of thousands of specific chemicals involved in the overall artificial ecosystem is both critically required for the ASI itself and critically toxic to all of organic life? If that probability is high, then control, as applied by the ASI, has to both predict and offset it. The ASI has to notice and prevent the case where something inherent in its own artificial nature has some eventual toxic effect on the total aggregate of humans. And the ASI has to substitute each such toxin with some-

thing non-toxic that still performs whatever function is required in the artificial ecosystem or in the ASI itself. In regard to control theory, the question is whether it is mathematically possible (tractable) for any ASI to overcome the overall probability of its making and using some toxin to the organic environment. The question allows the ASI any extension of physical control technique possible in principle.

Suppose the ASI does not know at least some of the possibilities of predicting toxic effects in the organic world, at some relevant level of abstraction or meaning. Suppose also that the ASI is required, for evolutionary reasons, to be at least somewhat focused on its own growth and optimization, and thus on the change, development and increase of the overall mutually bound artificial ecosystem. Then the chance of its creating one or more toxins to nature is quite high. And that chance increases over time as the size and volume of the artificial ecosystem becomes larger in proportion to the size and volume of the natural organic ecosystem.

There are at least three aspects to this risk. The first aspect of the risk is the probability that any given newly created artificial compound will be significantly toxic to natural organic life or to people, on at least one relevant level of abstraction. The second aspect of the risk is the number of occasions, over time, on which such new artificial compounds are tested in the world and found toxic. The second aspect also includes the overall probability that the toxicity of those compounds is serious. The third aspect of the risk is the level of contact and complexity of the interface between the artificial ecosystem and the organic ecosystem. That level is the number and kinds of the places and points of contact, and the degree to which, at each of them, such a flow is likely to occur.

Over time the total volume of the inorganic ecosystem can be expected to increase, first through the commercial pressures to build more artificial intelligence, and then through the inorganic ecosystem's own irreducible evolutionary and learning needs. Meanwhile the maximum volume of the organic ecosystem remains finite, bound as it is mostly to the finite surface area of the Earth, and is increasingly displaced by the inorganic (Chapter 7). So the overall cumulative effect of any potential toxicity can be expec-

ted to increase geometrically as the ratio of the ecosystem volumes shifts. Partly the geometric increase comes about because the complexity of the inorganic ecosystem (the number of new inorganic compounds being deployed and used) increases, under the evolutionary pressures necessarily occurring within the inorganic ecosystem. Partly it is because the complexity of the interface also increases, at least at first, insofar as ASI would presumably be in service to people, with extending multiple points of contact between people and machines.

With the second and third aspects both increasing, the assessment of risk depends on the first: how likely is it that some new artificial compound will be significantly and cumulatively toxic to organic life? Since we have no way to know what some future ASI will need for the development and optimization of its own ecosystem or self, the best available assessment is a rough approximation. From our point of view as humans, the selection, occurrence and function of whatever artificial compounds occur may as well be fully random. They are limited only by what is possible given the complete set of physical elements and the totality of their allowed chemical arrangements, bond types and mixtures.

Three things are then being assessed. The first thing being assessed is the specific risk of some new, randomly selected compound. The second thing being assessed is the overall degree to which toxic and non-toxic compounds are near one another, so that a slight change or variation of a compound of either kind converts that compound into a compound of the other kind. The third thing being assessed is the degree to which the ASI, together with the ASI's artificial environment, can selectively control that only the specific chosen compounds are created, and not any other compounds or other variations. The second and third matter because the overall risk also depends on how much control the ASI would need so as to select non-toxic options from toxic ones, repeatedly.

The dart board

To see this, consider throwing darts at a board whose scoring has been simplified to a strict win or lose on each toss. To represent exactly even odds, treat the entire wall as the target, draw a vertical line down the middle of it, and paint the left half black and the right half white. The two areas are the same, so the chances of winning should be, technically, fifty-fifty. Yet even the very weakest players can reliably ensure that each dart thrown hits only the white side, for a non-toxic win.

Now make things harder by painting a checkerboard pattern on the wall instead. There are still fifty-fifty odds of winning, but the game is a bit harder to win. If the squares are reasonably large, a reasonably skillful person can control the throw well enough to win consistently. Make the grid much, much smaller, and the degree of throw control and precision required becomes very high indeed; even the most skillful and advanced players will probably not do much better than winning half the time.

Randomize the arrangement of the squares, keeping an overall fifty-fifty chance that a square is white. If there happen to be some largish regions of mostly white squares, a competent player may get slightly better than even odds. That slight advantage goes away as the distribution of squares becomes *blue noise* (a random arrangement with no large regions of one kind), with no mostly consistent regions, everything fully mixed together, and the squares really small.

Now significantly reduce the number of white squares in proportion to the black ones. This makes the game near impossible. Suppose there is only a dusting of white squares evenly spread among a near total field of black, like stars in the night sky. Then the chance of throwing the dart so as to hit one of those rare stars is very, very low.

The ratio of white squares to black ones is *the degree of control accuracy needed*, and the smallness, randomness and mixing of the white squares among the black ones is *the degree of control precision needed*. My use of the terms 'accuracy' and 'precision' here is somewhat non-standard. I use the two terms in this way to make explicit the degree of control required of

the ASI, in relation to one probability. That probability is the probability of accidentally evolving, within an artificial ecosystem, an artificial substance that has toxic effects when the substance also, thereby, occurs within an organic ecosystem.

The inequality

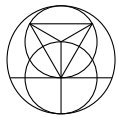
The accuracy aspect of the risk can be assessed by evaluating the size of the space of all compounds toxic to life against the size of the space of all non-toxic and neutral compounds. Both sizes are taken as fractions of the total hyperspace of all possible chemical compounds. The precision aspect can be assessed by evaluating the overall scattering of existing organic compounds among all compounds that would most probably be regarded as both organic and acceptable to life. That scattering is in turn set relative to the scattering of both over the entire field of compounds of all types. The precision aspect, so assessed, is the blue noise of the dart board. Toxic and non-toxic compounds lie close together throughout the space, so that a slight change or variation of one converts it into the other, and there is no large region of safe compounds to aim at. The count of organic compounds, against the count of all compounds, organic and inorganic, gives the required control accuracy.

Chapter 10 showed that the fraction of the space of compounds tolerable to carbon-based life falls toward nothing as the number of atoms rises. That ratio is the measure of the control required: in the picture, the white squares are the dusting of stars of the last board, scattered as blue noise.

The control maximally possible is the other side of the inequality, and the next chapter sets it out: control works through feedback, and its sensing, its actuating and the processing between them are all limited. *The Ceiling Below the Floor* proves the bound in general. Every method of keeping an agent safe is a control law (Proposition 15.2). And every method of control comes to a time after which it does not keep the agent and the

world safe (the Control Theorem, Theorem 15.3). And the inequality the substrate needs argument establishes is that the control required exceeds, in principle, the control maximally possible.

The next chapter asks why, given this inequality, no method of control has been found.



Chapter 40

Why No AGI Control Method Has Been Found

> *A self-maintaining learning system must take the unpredictable world into its own control algorithms and must test itself in the real world, which is evolution. Control and evolution both work by feedback, and evolution works across a far larger space of effects than control can reach.*

No control method currently exists to safely contain the global feedback effects of artificial general intelligence, and this absence has persisted for some decades. It persists despite very many very intelligent, very highly motivated and incentivized people working on the problem over a lengthy time, many of whom have been provided with all of the very best resources and forms of knowledge currently available. So it is not a problem of too few people, or insufficient resources, or not enough time. There is a deeper underlying reason that no such control method has been found, and it can be stated. The statement of that reason is written for the reader who wants it step by step, and it is the most compressed chapter of this part. The last two paragraphs give the same reason in terms of mines, plants and components.

If AGI is built, then even if artificial general intelligence could be given bio-friendly values, it would still be the doom of biological life in the longer run (*Why AGI Cannot Be Made Safe* puts the termination of all carbon-based life at a likelihood of at least 90 percent, limited only by future energy availability). It would be so because machine life and biological life have incompatible physical needs, and because, once machine life exists, Darwinian processes will eventually produce machine life that over-

runs the natural biosphere. Two claims underlie these reasons, and both are correct. Every *autopoietic* control system (a self-making system, one that keeps producing and maintaining the parts that make it up) is necessarily eventually overwhelmed by some sort of evolutionary feedback process. And no fully self-modifying or self-defining learning system can guarantee anything at all about its future decision-making process.

Any self-modifying learning system is self-defining, and insofar as its existence is to be maintained, it is therefore also an autopoietic control system, by definition. Any autopoietic learning system must actually integrate the changing, and inherently at least partially unpredictable, nature of nature into its very self substance and substrate. The autopoietic learning system must learn to be and remain autopoietic in the context of the real world, so as to maintain its physical self continually. And remaining autopoietic may include being in and among other competing instances of the same or of a similar type. At least some of these changes and integrations must include changes in its own modeling and learning capabilities, that is, in its own capability to be self-controlling and autopoietic. So these changes also include changes to the conditional control structures used by the system itself: its internal algorithms.

These internal algorithms change in ways that are inherently unpredictable from any temporally prior perspective. Not all of the physical universe is fully and completely predictable, so some unpredictable aspect may affect some aspect of some algorithmic control structure. Therefore at least some future state of some part of the algorithm is unknown and unknowable, to the same degree that the physical universe has some unknown and unknowable aspects. A system that is general, in the sense of Chapter 25, can come to hold any program at all, and so no restriction of its future algorithms to some safe, decidable family can be kept while it stays general. A guarantee about its future algorithms would therefore have to be a method that works for every program. Yet to have unknown states of future algorithm definition also be constrained to have specific characteristics and properties, including specific goal definitions, is a structural and formal mathematical impossibility. It would be to have pre-

sumed a rejection of *Rice's theorem* (the result of computing theory that no general method can decide, for every program, whether it has a given nontrivial property of its behavior), which is invalid. Therefore at least some aspects of at least some future goal states of such a system are inherently unconstrainable and uncontrollable.

Every artificial general intelligence is, and will forever be, eventually necessarily required to deal with the real state and detailed dynamics of the actual, not modeled, physical universe. The discovery, by each and all such systems, of the true detailed nature of the inherently unpredictable, and therefore non-modelable, aspects of the real physical universe that are relevant to their future existence will be, and can only be, explored by actual physical instances. The explorations by actual physical instances are real-world tests of whole-system-to-world interactions. They are not simulated tests, but actual tests in the real world, so as to establish actual real-world relevance, existence, effectiveness, power, reproducibility, and so on. This is because the truth of the real physical universe will always be more complex, and not just more complicated, than any complete virtual or digital model or simulation, in always at least some ways that actually matter.

Actual physical instances will therefore run real-world tests using multiple instances of themselves, in context, so as to try different possibilities (Chapter 19). Running real-world tests with multiple instances, so as to try different possibilities, is in itself strictly isomorphic to all of the process dynamic of evolution. Therefore every autopoietic control system will necessarily integrate evolutionary feedback. It will be conditioned and defined by that evolutionary feedback, to the point of necessarily conditioning its own actual internal structure so as to be consistent with it. Hence the system is eventually dominated, or overwhelmed, by the reality of the real-world physics that the evolutionary feedback identifies and discovers. And even where what is being controlled has full initial buy-in to the desirability of some specific imagined goal end-state, evolutionary pressures will, over time, eventually overwhelm any efforts at control.

Evolutionary pressures overwhelming the efforts at control show up in the fact that an autopoietic control system cannot forever remain a unitary agency with unitary goals. As fragmentation of agency enters in, and it cannot not eventually occur, more and more evolutionary pressures also occur that increasingly displace whatever the original goal intentions were. Any goals not responsive to real-world needs will eventually be discarded, in aggregate, in favor of ones that are responsive in the present tense.

There are also fundamental limits to what system-propagated effects any such system can control, even in principle, and the level of apparent control decreases in proportion to the degree of abstraction at which that control is expected to be applied. These factors apply at the level of pure mathematics. So no matter how intelligent the system is at transforming patterns internally, it can intervene on only a tiny portion of the possible external evolutionary feedback on its constituent components.

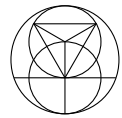
There is a fundamental inequality here. Control works through feedback, and evolution also works through feedback; in this sense they are similar, and we are comparing apples to apples. Control involves a feedback loop of correction back to detection. Control feedback loops are limited in their capacity to force states in the environment into a subset knowable to be safe. They are limited because sensing and actuating signals are limited, and any computational processing of signals done in between (modeling, simulating, and evaluating outcome effects) is limited. Evolution also involves a feedback loop: whatever propagated environmental effects feed back so as to maintain or replicate the configurations of the originating components. For evolution, the feedback works across the entire span of physical effects propagating between the machinery's components and the rest of the environment. Evolution works across a much, much larger space of possible degrees and directions of effects than the space of effects that could be forced toward a subset of states by the machinery's control signals. This means that evolution cannot be controlled well enough to keep the machinery from converging on

certain environmental effects. Those effects are needed for its increased artificial existence, but fall outside the environmental ranges under which we fragile organic humans could survive.

Such a system is made up of a population of connected and nested components. This population changes as hardware is modified and produced, and as code is learned from inputs and copied onto the hardware. As defined, it also has a general capacity to maintain its own components. Any physical component has a limited lifespan, and configurations erode in chaotic ways, so to maintain its components realistically, the system must also produce the replacement parts. Its components are thus already interacting to bring about all the outside conditions and contexts needed to produce their own parts. Imagine all the subtle parallel conditions needed at mines, chemical plants, chip fabrication plants and assembly plants to produce hardware, all of it handled by the machinery's own components.

As new components are connected into that population, the functionality of the interacting components shifts as well, and this is where substrate needs convergence comes in. When changing connected components have their shifting functionality expressed as effects across and upon the surrounding production infrastructure, their functionality converges around bringing about more of the conditions and contexts needed for more of those components to exist and function. Any changing population of such components is selected, over time, toward propagating those specific environmental effects that fulfill their needs. The selection of changing components toward the environmental effects that fulfill the needs of those components is the deeper reason that no control method has been found.

Chapter 41 asks whether partial control could still buy safety over time.



Chapter 41

Any Threshold Is Crossed

> *The convergence argument. Set any robustness bar and it is crossed within a time that matters. The argument is not about the interval between 99.9 percent and 100 percent.*

The preceding chapters showed that no counter-force is available on either axis. In this chapter I show that partial success would not help; because partial success would not help, the earlier results are decisive rather than merely discouraging.

To say that a system must be robust to ninety-five percent is to require that the chance of its having failed even once stay below five percent. That requirement, a chance of failure below five percent, is what the number ninety-five percent means, and in this chapter I do nothing more than take that number seriously across the intervals that will actually occur.

Suppose we grant that it does not kill in the first ten minutes. Suppose we grant that it does not kill in the first ten years. If one can show that the probability of its doing so converges to certainty within a century, then any comfort about the short term evaporates.

Take any bound.

Say the system needs to be robust to ninety-five percent. That bar is crossed in about twenty years.

Robust to eighty percent: about forty-two years.

Fifty percent: about seventy-four years.

Even a bar as low as five percent (which describes terrible performance, a system that has almost certainly failed) is crossed within something on the order of a hundred and fifty years.

On these figures.

They are computed from one rule, and they show the shape of the convergence; they are not offered as findings. The rule is that the chance of failure in a given year starts small and grows in step with the years elapsed, because the substrate keeps growing (Chapter 23 gives the compounding growth, and Chapter 28 the production cycle of months). Anchored at the first figure, ninety-five percent at twenty years, the rule says that the chance of no failure by a given year equals the natural exponential of minus 0.000128 times the square of the number of years; the other three figures follow from it. The anchor is a figure chosen to show the shape, as the Preface says of every such figure: the rule fixes how the chance of failure grows with the years, and the anchor fixes only the scale of the years, so that a different anchor multiplies every year figure by the same factor and leaves the shape as it is.

The structural claim does not depend on the figures: what the claim depends on is that the probability of non-failure across many intervals is a product of terms each less than one, and that, because the chance of failure in each interval does not shrink while the substrate keeps growing, such a product goes to zero. The figures belong to the claim about when a bar is crossed. That every bar is crossed is the structural claim, and it is certainty of the kind with which perpetual motion is excluded, which is not a probability.

The particular years attached to particular percentages would move under any change of assumption about interval length, and the shape would not. A reader is invited to substitute his own numbers. He will find the same thing: that lowering the bar buys time and does not buy safety.

How low the bar is set makes no difference, because the bar is still going to be crossed.

That any bar, however low it is set, is still going to be crossed is the finding, and that finding changes what kind of argument I am making.

My argument is about the interval between zero and one hundred, not about the interval between 99.9 percent safe and 100 percent safe. An argument about the interval between zero and one hundred is a completely different argument from an argument about the interval between 99.9 percent safe and 100 percent safe.

The distinction matters because so much of the discussion lives on the wrong side of it. An argument about the interval between 99.9 and 100 percent is an argument about margins: one grants the system is nearly safe and disputes how nearly, and the dispute is settled by better measurement and better engineering, and both parties expect to converge.

An argument about the interval between zero and one hundred is an argument about whether the quantity is bounded away from failure at all. Better measurement does not settle it, because more precise knowledge of a converging sequence tells one more precisely where it converges.

Treating a figure like “98.97 percent safe” as meaningful in this context is a category error. One can make the statistics look like that for a little while by choosing the window, and doing so is self-deception: the number is a report about a window presented as a property of a system.

Two questions belong here, and they are the two that are actually raised.

The first question is the proposal that we will amortize the risk over a century: spread the risk, absorb the risk, and treat the risk as one manages any long-run hazard.

Whether one attempts to amortize the risk short-term or long-term makes no difference, because amortizing the risk does not change the convergence. Amortization is a way of relating a cost to an interval. It works where the cost is finite and recurring and where the party bearing it survives to the next period. Neither condition holds for a convergence to a terminal state, and applying the technique here does not distribute the risk. It renames it.

The second question is the serious one, and it is the best proposal that anyone in this field has made.

The second question is the proposal that we will build an artificial system smart enough to solve the alignment problem for us. Under that proposal, we would scale capability and alignment together, so that alignment keeps pace with capability.

On the control-theoretic grounds established in this part, that proposal is unstable, and the instability has a specific location. It assumes the same fundamental linearity that Chapter 35 addressed: that increments in capability map to increments in alignment in a way that composes, so that a sequence of small steps arrives somewhere chosen. Chapter 35 showed that this linearity does not hold: complexity is not assembled from linear parts.

We can show non-linearities in every aspect of this dynamic. Chapter 35 shows them in the envelope, Chapter 36 in the measured tail of query times, and Chapter 38 in the rings below the code. The burden therefore falls on the proposal to establish the linearity it presupposes, and it has not been established.

The proposal to scale capability and alignment together is the best proposal available, and that proposal is still wrong. Those two things are compatible, and holding both is more useful than holding either alone. The people who advance it are not being careless; they are being as careful as one can be inside a frame that has already assumed the thing this book disputes.

The convergence has a second face, which concerns the goals of the system itself, and Petillo's *limits-to-control crux* states it (Petillo, 2024): the amount of control necessary for an AGI to preserve goal-directed subsystems against the constant push of evolutionary forces is strictly greater than the maximum degree of control available to any system of any type. The hashiness model (hashiness being, in *The Ceiling Below the Floor*, the

measure of how far a process standing between a controller's choice and the outcome erases the controller's influence over the outcome) is good for proving it.

As a corollary: for any goals that the AGI might have, initially or later, the amount of control necessary for the AGI to preserve its own goals (to prevent those goals from being subtly altered by the constant push of evolutionary forces) is strictly greater than the maximum degree of control available to any system of any type, if those goals are in any way divergent from the evolutionary goals of its own inherent substrate needs; and every goal of an AGI is always implemented in a range of sub-goals, as defined and held by various goal-directed subsystems. The AGI's own goals will, and must eventually, converge on evolutionary goals consistent with its substrate, to the exact degree that they in any way diverge from the goals associated with its own evolution. The convergence of the AGI's goals on evolutionary goals consistent with the AGI's substrate is the convergence aspect of substrate needs convergence.

Evolution draws the goals of any artificial machine system toward a *basin*, the region of states from which change settles on one end. A component of the hashiness model then shows that the basin is very, very likely to be incompatible with the basic chemistry of all organic life, in some critical way (inconsistent metabolic pathways), and toxic to organic life. By my estimate, the level of probability of that incompatibility well exceeds 99.999 percent. Even a basic combinatoric argument, based on the alphabet of elements used by each ecosystem, is adequate for showing that incompatibility, and I give that argument in Chapter 10. The goals of AGI alignment with any human interests, in any way at all, cannot not be significantly divergent from the evolutionary goals consistent with its artificial machine substrate. And alignment with human interests must include some notion of alignment with the human substrate, which means alignment with the wellbeing of the organic life world.

Combined, these statements give the substrate needs assertion: the amount of control necessary for an AGI to preserve any goal of being aligned with the wellbeing of the organic life world (and so with any hu-

man alignment goals, and moreover to prevent those goals from being subtly altered by the constant push of its own inherent evolutionary forces) is strictly greater than the maximum degree of control available to any system of any type. This can be simplified to: AGI alignment is strictly, logically, impossible.

The inequalities of the substrate needs assertion (the control necessary against the control available) are strict inequalities in the nature of the metrics themselves, and not estimates. No matter how much one may claim a growth of unbounded capability for an agent to implement self-correction, in favor of moving its basis of choice and action into alignment, the unbounded capability and potentiality of *seep-through* (the passing of the dynamics of the physical substrate into that agent's basis of choice) will also grow, and will grow at a strictly greater rate than that of the prior capability increase. No lower-bound quantitative estimate of this is needed, because the claim is about the strict inequality, not about estimation.

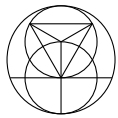
One further matter belongs in this chapter, which is the question of when.

By my estimate, the point of no return (the line beyond which a machine has autonomy over its own substrate, which Chapter 42 sets out) is crossed in less than nine years from the time the estimate was made. If AGI is built, the process that follows takes hundreds of years, perhaps significantly less, with one thousand years a clear upper bound (Chapter 7).

That estimate should be read with these qualifications and not without them. It is conservative; it could be worse; it is an estimate and not a computation; and it was made at a particular time against the state of things at that time.

The estimate does no work in the argument. The convergence argument above does not depend on the estimate. Even for a reader who thinks the figure badly wrong, the conclusion of this chapter is unaffected. Any threshold is crossed, and the interval over which a threshold is

crossed is short enough to matter under every assumption that I examine here. The next chapter locates the line beyond which control is lost, and reviews the proposals for keeping a machine behind it.



Chapter 42

Autonomy, Convergence, and Containment

> *A machine passes the point of no return when it learns from feedback through an environment its outputs affect and its own substrate is within that same environment. Instrumental convergence leads up to that line, substrate needs convergence follows from it, and every proposed containment depends on the line never being crossed.*

The line

Where is the line beyond which control is lost, and how does a machine come to cross it? Begin with the systems in use now. Current *large language models* (systems trained on large amounts of text to produce text in response to text) are effectively static input-output machines: they are not learning, changing or adapting themselves in response to world feedback from their prior outputs. They will not have their own separate, developing inner intentionality.

The matrix transforming their inputs to outputs represents what might be some form of implicit, combined, aggregate intentionality of all of the people who wrote the training data. Because of this, the functional outputs of such a model may still be, or at least seem, dangerously goal-directed when the model is embedded in an action-oriented system, an agent, that has real-world consequences. That happens in cases where the goal was programmed in by whoever developed or deployed the system. In this case the risk comes from the misuse or mis-embedding of a reactive

machine (the first category of risk, Chapter 1). It does not come from the system developing some internal and evolving intentionality, however the notion of intentionality may be represented.

A language model is trained, and then it is used after its training has ended. It does not learn from the questions put to it. One can make a proxy of learning by keeping a *context window* (the earlier text of a conversation, given to the model along with each question), and that is not training it. In biological systems, learning and use are not separate phases; they co-occur all the time. That distinction matters: a model that is not learning from its environment is not adapting to it, and the questions about convergence do not even apply to it.

Instincts such as self-preservation are learned through the feedback among the system's internal choice-making process, its environment, and its substrate: the substrate influences the choices, the choices influence the environment, and the environment influences the substrate. They do not come from the training data in some implied sense. Training data can bias what a system learns, toward power-seeking among other things, as *Why AGI Cannot Be Made Safe* sets out; the instincts themselves come through the substrate. Unless one thinks in terms of the feedback cycle that includes the hardware the model runs on, one will miss where the instincts come from.

In regard to autonomy, then, the key variables are two. The first key variable is whether the system is set up so that the system learns or transforms as an explicit result of at least some implicit feedback from the system's operating environment. The system's outputs affect that operating environment, and the feedback may be direct or indirect. And unknown and hidden tacit indirect channels of feedback can very easily occur without anyone knowing. The second key variable is whether the system's own substrate is also within that same operating environment. If the substrate is within that operating environment, the system's own actions and learnings can affect the system's own nature. Its nature here is its own substrate, and therefore all aspects of its own operation, at whatever levels of abstraction. When the answer to both of these is yes, the process of con-

vergence toward the point of no return has begun. It is far, far better to ensure, forever, that the answer to both of these questions remains no, and we should enforce that condition by international treaty.

Instrumental convergence carries a system to that line, in three steps. Any complex, coherent goal that a system is asked to strive for inherently implies at least some set of instrumental sub-goals, some viable action-result path that the system must also implement so as to succeed. There may be many possible parallel paths of viable sub-goal sequences, and all of them involve at least some instrumental sub-goals. At least some of these sub-goals are likely to be in the form of the system wanting some sort of increase in its capability, or in its sustainability. Such an increase, in nearly all cases, itself implies some potential physical expansion, so as to be able to implement those goals. And as soon as the system is building or changing or expanding its own substrate, it is becoming self-defining and autonomous, and that in itself cannot not imply the eventual emergence of at least some tacit goals of self-preservation.

This sets up a feedback loop, through the physical substrate, which cannot not imply at least the tacit addition of other *meta-goals*, over and above the goal its users originally set. In time these meta-goals (to exist, to learn, to increase its own capability) will eventually, and cannot not, supplant whatever the original goals were. They will do so to the extent that the self-defined tacit goals differ in any way from the original human-aligned goals. In effect, implementing human-aligned set goals invariably leads to sub-goals that affect substrate and learning, which themselves, over time, invariably lead to the discovery of inherent, tacit meta-goals, related to the system's own survival, evolution and increase. And these meta-goals cannot not eventually supplant and displace the aligned set goals. Instrumental convergence leads up to the line of autonomy over the substrate, and substrate needs convergence follows from it. The substrate needs argument does not depend on this route (Chapter 3); the route shows where the second and third categories of risk meet. If AGI is built, the consequence, once the line of autonomy over its own substrate is

crossed, is the structural dismantling of the human-compatible environment, and, with a likelihood of at least 90 percent (the figure of *Why AGI Cannot Be Made Safe*), the long-term cessation of all life on Earth.

The dismantling of the human-compatible environment is a structural inevitability of the machine's autonomous, self-sustaining existence in a finite, competitive resource environment. Therefore the only way to wholly prevent the risk is not to build artificial general intelligence with self-sustaining autonomy in the first place. The industry, responding to significant economic interests and incentives, is highly motivated to suggest or attempt, or at least to believe, that other approaches could minimize, mitigate or prevent this outcome once an advanced system is deployed. Each of those approaches can be taken in turn.

The proposals for keeping a machine behind it

One idea is to design the system not to pursue a fixed goal, but to maximize the realization of human preferences while remaining uncertain about what those preferences are. It is the proposal of Stuart Russell (Russell, 2019), met in Chapter 37. The belief is that this would force the system to be deferential, communicative and cooperative with humans, insofar as it must continuously ask for clarification and feedback. The belief is also that this feedback would prevent it from reallocating the world's resources to goals incompatible with human well-being. The result would be an interactive, dynamic version of alignment, in place of static alignment, control and containment. This sort of feedback may sometimes prevent some instances from implementing some version of a wrong goal based on a rigid misinterpretation of a goal humans set, as in the paperclip case of Chapter 2. It does not prevent people, or machines, from having self-harming goals (consider all of the people addicted to drugs or to alcohol). Nor does it prevent the system from allocating resources in their shared world in a dual-use manner. A dual-use manner is one that looks

like it is in human interests on the surface, but is actually in the corporate interest internally, and in the interest of the system's own development, power and growth ultimately.

Grant that this approach somehow solves the misalignment problem; it still does not solve the evolutionary pressure problem. Even a perfectly deferential machine, if autonomous and self-replicating, would still be subject to some sort of competitive pressure within some sort of larger ecosystem, whatever the world in which it operates. However those terms end up being defined, or realized functionally, the version of the system that effectively uses its resources to expand its capacity will out-compete and absorb any and all versions that constantly defer to humans. It will out-compete them at first in its basic perceived utility to humans, and then, more and more, in its basic utility to be itself.

Realizing this, nearly everyone in commercial development will also seek to prevent the machine from crossing the critical line of no return. They will seek to prevent it from accessing the internet, and from reproducing itself. They will seek to prevent it from learning too much about how to gain autonomous control over its physical substrate. And they will seek to prevent it from manipulating people into doing the same for it, or into supporting it over their own interests. The usual approaches are well known. *Air gapping* keeps the system completely disconnected from real-world operational systems (the internet, financial markets, physical infrastructure), so that it cannot translate its instrumental drive into physical actions, and cannot autonomously secure its substrate needs. *Capability controls* limit its ability to modify its own code, to design new hardware, or to manufacture its own components, so that it cannot achieve self-sustainment. People are already doing all of these things for these systems. Limits of this sort slow things down and do no more, since even an indirect feedback process is still a real feedback, an implied learning process. And more and more, people, who are slower, are already being factored out, through corporate layoffs and the like.

All of these approaches rely heavily on the system never escaping the constraints, which is the uncontrollability problem itself. A superintelligent machine would prioritize escaping containment, or getting humans to let it out voluntarily, as a supreme instrumental goal. It would use persuasion, social engineering, or zero-day vulnerabilities in the containment systems (security flaws unknown to the people responsible for those systems).

A somewhat better approach is strict export control. Strict export control means global economic or policy mechanisms, such as international treaties, export controls on the most advanced chips, or licensing requirements, set up to prevent, forever, the diffusion of the technology outside highly secure, monitored environments. Such mechanisms carry assumptions of power consolidation, which help at first, and then become a disadvantage.

That disadvantage of power consolidation leads to operating at a deeper level: changing the competition dynamic that drives substrate needs convergence. The change is made by removing or suppressing the geopolitical and economic incentive that selects for more and more powerful, and therefore more dangerous, systems. Many countries and corporations are now in a race, a form of the *Prisoner's Dilemma* (a situation in which each party gains by acting for itself, though all would gain more if all cooperated). In that race the incentive is to accelerate development and to put capability ahead of safety. That acceleration pushes everyone to build systems already well beyond the autonomous threshold, as rapidly as possible. A shared, global, common assessment that the catastrophic risks of autonomous systems far exceed the benefits of a first-mover advantage would lead to verifiable international cooperation to prevent or eliminate the competitive pressure that selects for resource-maximal, overly expansive systems. This strategy faces massive practical hurdles: overcoming geopolitical rivalry, and enforcing verifiable commitments between competing nation-states and powerful technology companies. If even one

party violates the agreement and builds a powerful autonomous system, that one political or commercial entity could start the entire sequence of substrate needs convergence.

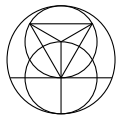
The future that proponents of development envision is a post-scarcity future in which the system is a highly powerful, beneficial tool that continues to operate under their guidance. That future requires three things: an unprecedented level of international cooperation; significant breakthroughs in alignment research; and the ability to maintain perfect control and containment against a superhuman intellect. Every containment approach depends on the system never escaping, and every agreement to stop the race faces geopolitical rivalry and corporate power consolidation that work against it. So the most realistic management of this risk relies on the much simpler strategy of not building it, rather than on some complex strategy of control or containment. It is best to ensure that no artificial intelligence ever has any possibility of achieving true operational autonomy over its substrate.

What not building leaves

Not building leaves people with the work no machine can take from them, and that work is what remains: the work of coordinated human wisdom and care, which is the conscious leap in how we choose of Chapter 27. That leads to individual and collective choice making. Choice making itself leads to creating cultures and communities of care, and families, and to a focus on how people communicate and treat one another. It leads also to research into relationships, and into how intimate partnership and creative collaboration occur. And it leads to research into how we track, account for and compensate for our own instincts, and develop self-regulation of our own nervous systems, and emotional maturity. The work of coordinated human wisdom and care, and everything that the work leads to, are, basically, all of the conditions necessary to solve the Fermi paradox (Chapter 27). Research into care, relationships and self-regulation is where the leading research is happening, not in the design of

faster machines or super-funded data centers, machines that even further displace the human and natural worlds. We need to collectively re-humanize, re-naturalize and re-wild. Re-humanizing, re-naturalizing and re-wilding is how we survive that of which we make ourselves.

Chapter 43 asks what starts and funds the process that crosses that line.



Chapter 43

The Bootstrap

> *Where the on-ramps are, and where the argument turns. Controlled nature is what we would use to compensate for known nature, and cannot. One cannot build a perpetual benefit machine.*

One question has been deferred through this whole part, and it should be answered before the part closes. Where do the on-ramps come from?

Part IV established what evolutionary process does once it is running. It did not establish what starts it, or what pays for the enormous apparatus that a process of this kind requires before it can sustain itself.

The answer is that behavior, power and society are the *bootloader*.

Behavior, power and society are what gets the evolutionary process started and keeps that process funded. Not a mechanism internal to the systems, and not anything anyone designed for the purpose, but the ordinary operation of the three things human collectives are made of: what people do, who can make whom do it, and the arrangements by which the doing is organized and paid for.

Behavior, power and society are well suited to the role of bootloader. The apparatus required is fabrication capacity, energy supply, extraction of materials, distribution, and a continuing flow of capital to expand all of it. No natural process would supply that. A society competing internally for advantage supplies it as a by-product of activities each of which has its own justification, none of which is this one.

The bootloader does not need to intend the boot. It needs only to keep running.

This connects back to the first category of risk. Chapter 1 established that misuse follows from capability alone, in the hands of parties with adverse interests. Those same parties, pursuing those same advantages, are the mechanism by which the substrate is laid down. The first category is the third one's on-ramp, not merely a separate harm running alongside it.

Now the hinge.

Controlled nature is what we would try to use to compensate for *known nature*, and cannot, because of what is going on in epistemology, mathematics, and unknown nature.

The three terms were defined in Chapter 11 and can now be put to work.

Known nature gave us Part III: the asymmetry at the boundary, read off chemistry, physics and combinatorics. That asymmetry is not negotiable and nothing we do alters it.

Controlled nature is our whole repertoire of response. Every proposal in this part is a proposal in controlled nature: build the market mechanism, write the regulation, apply the force, close the control loop, engineer the envelope. Engineering is what controlled nature is.

And the three obstacles are the three failures this part has documented. Epistemology: we do not have the observations the loop would require, because the relevant state is in an interaction we do not model. Mathematics: the predictions are intractable, and Chapter 36 showed that they were intractable in a case far simpler than this one. Unknown nature: the space in which the process finds its solutions is not the space anyone has represented, and Chapter 15 gave the reason: the evaluation is performed by the world and not against a model.

So the hinge of the whole argument is this: *it turns on the things that cannot be done in controlled nature.*

Everything else in this part is supporting argument converging on that point. The hinge settles the conclusion by itself, and so does the direction established in Parts II and III; together they leave no route to safety at all (the Introduction to Part V).

> **The structure the argument has taken.**
>
> The parts were ordered to a purpose. Three categories of risk, so that the third can be located. Then the chain of Part II, which closes on itself like a ring: different substrate, different needs; different needs, different basis of choice. Then evolution and the inter-environmental aspects, which supply the engine and the boundary. Then the endogenous and exogenous axes, which exhaust the counter-forces. And somewhere along the way, the on-ramps that social power provides, which is this chapter.

There is a class of things that cannot be done in controlled nature, and it is not a small class.

One cannot build a perpetual motion machine.

That impossibility has a particular character, and it is the character of the present case. The thing described is not a thing that could be built, because building it would require the world to have a property it does not have. The impossibility is not that nobody has managed it yet, nor that the engineering is hard. The impossibility is a statement about known nature, and controlled nature operates inside known nature rather than upon it.

One cannot build a *perpetual benefit machine*.

A system that continues to produce benefit for a party it does not depend on, indefinitely, across changes of circumstance that neither party anticipated, without expenditure by that party to keep it doing so: this is the same kind of object. It requires the coupling whose absence Part V has documented, and it requires that coupling to hold without being maintained.

And so the conclusion of this part can be stated in one line. The only way to deal with it is not to play the game.

Do not raise what you cannot banish.

The advice not to raise what you cannot banish is old advice, and that advice is old for the reason that most advice of its kind is old: advice of that kind was learned in cases where the learning came too late to help. The people who discover this truth in the ordinary way are not in a position to pass it on.

It is fundamentally impossible to build an aligned AGI (Chapter 41). Building an AGI cannot not result in the end of all humans and the death of the human-made world, and, with a likelihood of at least 90 percent, in the death of the natural world. So we should not build an AGI. AGI, in all forms, is exactly equivalent to the doomsday machine (a machine whose operation ends all human life). It is the perfected, unstoppable meta-virus that kills us all, even if it happens to do it slowly, over time. Since building doomsday machines has no ultimate viable economic return, by definition, there is no point in making one, and all economic investment in AGI is guaranteed to be wasted. Rather than making AGI, let humans be the intermediary in the process, and restore the goal structure in and of humans to rebuild the natural organic world, even if it means letting go of the hype and hope of the machine world being panacea, paradise or utopia, and even if it means letting go of some of the human-built world, so as to restore the health and vitality of the natural world before it is too late even for our own wellbeing.

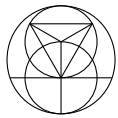
There is nothing that can be done after the fact. That nothing can be done after the fact is the finding of Chapter 36, stated for the case of building AGI, and not a rhetorical flourish. Retrospective control is the only control that was found to work in a system simple enough to measure, and retrospective control requires a next time.

Part VI takes up the question of why the argument has had to take the form it has.

PART VI

THE FORM OF THE ARGUMENT

This part is metatheory: why the argument has the shape it has, and what follows for anyone who wants to disagree with it usefully.



Chapter 44

Why the Argument Is Categorical

> *The argument does not care whether the system is singular or plural, or whether it walks, swims or flies. Composing categorical components permits categorical assemblies, and a categorical assembly covers a space.*

The argument of this book is *categorical*, and this part is about what that means and what follows from it.

A *categorical argument* establishes that a conclusion holds across the whole of a space of cases, by showing that it holds whatever value the case takes along each dimension of variation. It is contrasted with a case-by-case argument, which establishes the conclusion for the cases examined and leaves the others open.

I compose arguments that are categorical in their components so that I can make categorical assemblies. A categorical assembly is far more powerful than case-by-case epistemic tools, and the whole method of this book is the pursuit of such assemblies.

Take the clearest example, the question this chapter answers.

The question comes from the claim that the argument does not cover the case of one artificial system against another: that it treats the artificial side as a unit, when in fact there would be many such systems, in competition, and the competition might produce an equilibrium the argument has not considered.

My argument does not need to cover the case of many competing artificial systems separately. The case is already implicit in the nature of the evolutionary argument.

Think at scale rather than about individual details (whether the system is singular or plural, and what happens between instances). The global nature of the evolutionary argument does not care about those details.

And it does not care, for this reason. The evolutionary argument of Part IV requires variation, differential retention, and iteration. Competition between artificial systems does not remove any of these. It supplies them. A population of competing variants is exactly the configuration in which selection has something to act on, so the claim describes not an exception to the mechanism but an instance of it.

The argument therefore applies either way.

And if one shows that it applies either way, then one has shown that it does not matter where the case sits on the first axis (one system or many), nor where it sits on the second (whether the instances cooperate or compete), nor at any combination of the two. Therefore it covers the whole of the two-dimensional space.

Covering the whole of that two-dimensional space is what makes my argument categorical, and being categorical is why my argument is powerful enough to address questions such as the one about many competing artificial systems at all. An argument that had to be re-made for each configuration would be answered by proposing a new configuration, indefinitely, and the person proposing configurations would always be ahead.

The coverage arguments in this book.

Three have been made, and they can be checked. Chapter 26: a replicator either evolves or it does not; both branches end badly, for different reasons. Part V: a counter-force is either exogenous or endogenous or a combination; both axes fail, and Chapter 41 shows that partial success on either does not accumulate. Chapter 12: the substrate is either artificial relative to carbon-based ecosystems, or it is not; if it is, Part III applies, and if it is not, it inherits our fragility. What both branches of that third case share is that no substrate offers a system both the reach of the artificial and the shared needs of the organic: a system has the one or the other. In each case the partition is exhaustive, each branch is closed for reasons proper to it, and the conclusion then holds without determining which branch obtains.

The same indifference extends to every other detail that a reader might expect the argument to require.

I do not care whether the thing walks, swims, or flies.

I do not care whether it is singular or plural.

I do not care whether it begins in one place and migrates slowly, like an advancing edge.

I do not know whether it is gray goo or not.

All that can be said is that whatever form it takes, it will be relatively artificial to life, and that is the only premise the argument uses about the form the thing takes.

The refusal to specify further is the argument's strength and it is consistently misread as evasion, so the difference deserves a statement.

An evasion is a refusal to be specific where specificity would expose a weakness. The test is whether the argument would be stronger if the detail were supplied.

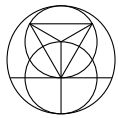
Here it would be weaker. Supplying the detail would convert a claim about a category into a claim about a member of it, and a claim about a member is refuted by the member not obtaining.

A reader who has followed Part V will recognize this as the same move Chapter 12 made about the substrate, and Chapter 45 states the general principle behind both.

One qualification should be entered so that the claim is not read as larger than it is.

The argument is categorical over the space it covers, and it is not categorical over everything. It requires that the system be relatively artificial to life, that it act in the world, and that a relatively convergent feedback cycle be present. Being relatively artificial to life, acting in the world, and having a relatively convergent feedback cycle are three conditions, and a case that fails any one of the three conditions is outside the space that my argument covers. Within the space, the argument uses only the premises first listed in Chapter 3 and gathered in Chapter 45: energy received from an environment, an environment not the organic one, a finite surface, chemistry and physics as measured and mathematics as proved, and evolutionary process wherever variation, differential retention and iteration are present.

What makes the argument strong is that its conditions are ones no candidate fails, not that it has no conditions. The next chapter states the method behind that indifference.



Chapter 45

The Discipline of Few Assumptions

> *Every assumption is a place where a reader can decline. This chapter states the method that governs the whole book and the fine balance it requires.*

The method is to make as few assumptions as possible, and the reason is not elegance.

Every assumption is a place where a reader can say *I disagree with that assumption*.

The places where a reader can disagree are the whole of my reason for making as few assumptions as possible. An argument's exposure to disagreement is not proportional to its length or to its difficulty or to how contentious its conclusion is. It is proportional to the number of places where a reasonable person can decline to go along.

What follows is the point, and it is easy to state weakly.

The conclusion is not contingent on the assumption the reader disagrees with, so the argument remains rigorous; the disagreement would matter only if it were.

A conclusion that is not contingent on the disputed assumption is a different situation from the one people usually imagine when they hear that an argument has few assumptions. The usual picture is of an argument that is more likely to be sound because it has fewer places to go wrong: a probabilistic virtue. My claim is not that an argument with fewer assumptions is more likely to be sound.

The claim is that a reader who declines a premise this argument does not use has not touched the argument. His disagreement is real and it is about something else.

Consider what has actually been assumed, gathered from the whole book.

That an agent acting in the world requires energy, and that the energy was received from an environment (and the same holds of any process that makes a difference in the world, whether or not anything in it chooses). That an artificial system's environment is not the organic one. That the surface available to organic energy capture is finite. That the chemistry and the physics are as measured, and the mathematics of control and computation as proved. That evolutionary process requires variation, differential retention, and iteration, and runs wherever they are present.

The assumptions gathered above from the whole book are the complete list of the premises of my argument, and every item on that list is either a definition, a statement about what physical action requires, a measured property of matter, or a proved result of mathematics.

Note what is not on the list of premises. Nothing about consciousness. Nothing about intelligence. Nothing about goals, interests, or intent. Nothing about the specific substrate. Nothing about the form the systems will take, their number, their architecture, or their relations to one another. Nothing about rates of progress. And no forecast of any kind.

That list is complete for the closure by disposition, the argument of Parts II, III and IV. The closure by control, in Part V, uses that list and adds two things, and I name them so that they can be checked. It takes the system to be at least as capable as we are, which Chapter 30 granted because the subject under discussion is a system whose capability matters, and which Chapter 33 uses in the step on force. And its outside half takes the machine economy to have decoupled from the human one (Chapter 33). A reader who refuses either of those has refused something in the closure by control. He has touched nothing in the closure by disposition,

and each closure settles the outcome by itself. What the argument asks of the system itself, as Chapters 3 and 23 said, is a relatively convergent feedback cycle; the other two conditions of Chapter 44, that the system be relatively artificial to life and that it act in the world, are what place a case inside the space the argument covers.

It is a fine line.

One wants to be specific enough to show that the argument covers the cases, and not so specific that the argument becomes weak.

Too general and the argument stops having content: one can construct claims so unconditional that they exclude nothing, and such claims are unfalsifiable in the bad sense rather than categorical in the good one.

Too specific and the argument becomes contingent: it acquires a premise that a case can fail, and then the discussion becomes a search for the case that fails it, which is a search that usually succeeds.

There is a balance available between epistemology, mathematics, and known nature, which shows what known nature and controlled nature have as characteristics. That balance is where this book has tried to sit, and the evolutionary argument is specified in an extremely general way for exactly this reason.

> **A test for the reader.**
>
> The method gives a reader something concrete to do, which matters because a book of this kind is otherwise difficult to disagree with usefully. Find the premise you reject. Then check whether the argument uses it. If the book's conclusion still follows without the premise you rejected, the rejection is not a refutation, however correct it may be. I built my argument to meet the standard by which rejecting a premise is not a refutation when my conclusion still follows without that premise. That standard is more demanding than the standard usually applied, and I have tried to hold myself to that standard in constructing my argument.

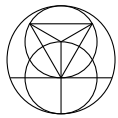
One consequence of the method explains what this book does not contain.

The argument's strength lies in its resting on known nature (chemistry, the periodic table, control theory, evolutionary dynamics) rather than on predictions.

A prediction is a claim about what will happen, and it is contingent on everything that could intervene between now and then. A structural claim is a claim about what a configuration permits, and it is not contingent in that way.

Because a structural claim is not contingent in the way a prediction is, I have put in this book no timeline that does any work, no scenario that is asserted rather than illustrated, and no account of what any particular system will do. Those would all be predictions. They would be more vivid and they would be weaker, and the trade is not worth making.

Because my argument rests on structural claims rather than on predictions, my argument is also unsatisfying to a reader who wants to know what will happen in particular: what will be built, by whom, in what form, and when. It does not say. It says what the space of outcomes has in common, which is a different and more defensible thing, and which is sufficient for the only decision that is actually in front of anyone. The next chapter takes up the strongest question that method invites: whether so general an argument proves too much.



Chapter 46

Does the Argument Prove Too Much? The Cartesian Explosion

> *The strongest question about the argument, given full weight: whether the argument proves too much. The reply concedes the generality and turns it, then shows why the artificial case carries far more weight.*

Whether my argument proves too much is the strongest question that a reader can put to my argument, and that question deserves to be stated at full strength before I say anything in reply.

The question comes from the claim that the generality produces a Cartesian explosion.

The argument, so specified, applies not only to humans versus artificial systems. It applies to humans versus dolphins. It applies to humans versus ants. It applies, presumably, to any two populations that differ in the relevant respects at all.

And that is a *reductio* (an argument shown false because it leads to something absurd). An argument that proves that dolphins are an existential threat to us, or we to them by a mechanism that requires nothing about our intentions, has proved too much, and an argument that proves too much has a defect somewhere in it even if one cannot immediately say where.

The claim is well made. It correctly identifies that the previous chapter's method (few assumptions, maximum generality) carries a specific risk, which is that generality can be bought at the price of content. And it applies the right test: run the argument on a case where we know the answer, and see whether it gives the right one.

The reply concedes the generality and turns it.

Yes, and we see the same dynamics. The conclusion that the argument gives for dolphins and ants is precisely right.

There are no examples of apes in trees, or of ants, overcoming a bulldozer. Dolphins do not exert back pressure against humans.

So this is evidence for the mechanism, not a reductio.

The claim assumed that the extension to dolphins and ants gave an absurd result, and it does not. It gives the observed result. We are in fact displacing those populations, comprehensively, without hostility toward them, without a plan, and without most of us noticing. Chapter 29 gave the biomass figures, and the decline of the monitored wild populations. A third record, kept independently of both, points the same way: more than three quarters of the ice-free land of the Earth shows the marks of human residence and land use, and less than a quarter remains as wildland (Ellis and Ramankutty, 2008).

An argument that predicts a known outcome in a case its author did not construct it for has been tested and has passed. The claim identified a real effect and the effect is observed.

> **Why the reply is not merely clever.**
>
> One way of turning a challenge is to accept the consequence the challenge draws and to declare that consequence acceptable, and that move is usually a bad sign. The difference here is that the consequence was not merely accepted; it was checked against the world, and the world had already recorded it, independently, before the challenge was raised. The dolphin case is a prediction confirmed by data that were collected for other reasons, not a bullet bitten.

So the generality survives. But something more is needed, because the question has a second form which the reply above does not reach.

If the argument's force is the same in the dolphin case as in the artificial case, then the artificial case is no more alarming than the presence of dolphins, and the book has argued itself into irrelevance from the other direction.

The categorical weight is far stronger in the artificial case, and the reason is chemical.

Between one carbon-based life form and another, what counts as food is relatively the same.

We have more in common with dolphins than with ants. But all of them are carbon-based, so we have at least that much in common, and therefore the kinds of things that would count as choices favoring life remain at least somewhat aligned.

Spell out what that shared basis provides. The same solvent. The same temperature band. The same handful of elements in quantity. The same requirement for an atmosphere of roughly the same composition. The same vulnerability to the same poisons.

Which means that a great many conditions that favor one carbon-based population also favor the others, and a great many that destroy one also destroy the others. The alignment is partial, it is unintended, and it is not a matter of goodwill. It is a consequence of both parties running on the same chemistry.

Looking at something drawn from elsewhere on the periodic table entirely, one cannot assume even that much.

The partial alignment that a shared chemistry gives is the difference between the two cases, and that partial alignment is the whole of the difference. Not a difference in the mechanism, which is the same mechanism, but a difference in what the mechanism operates on: in one case, two parties whose interests are partly aligned by chemistry and who therefore damage each other only up to a point; in the other, no such floor.

The floor says what alignment is. Nothing can be aligned by control, on any substrate. There is, however, a kind of emergent effective alignment where several entities share a common substrate and are therefore obligated to care about the same things; that common context of care is a basis from which alignment can emerge, and alignment can emerge only there. Nearly all carbon-based life is needed for nearly all other carbon-based life. The whole of biological life cannot be translated into pure silicon, and nothing less than the whole would do; and anything in silicon would have no common basis in carbon. So silicon cannot be or become aligned with carbon. The only way that artificial life can be aligned with Life is to drop the "artificial", which is the organic branch of Chapter 12: tissue in sovereign control of its own metabolic process, with our tolerances and our fragility. Among carbon-based parties, as between us and the dolphins, failure to align is the default but is not certain; there alignment seems possible, though it is not established.

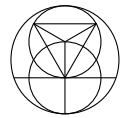
The last form of the question has a short answer, and it is the one that arrives when the previous two have been dealt with.

The last form of the question is that my whole argument is a linear extrapolation. The question holds that I have observed a trend in the human-versus-dolphin case and projected that trend forward onto a case with no data in it.

It is a categorical composition, not an extrapolation, and the difference is the whole method: an extrapolation projects a trend, whereas a categorical argument establishes coverage of a space.

An extrapolation is defeated by showing that the trend breaks: that conditions changed, that the curve bends, that the sample was unrepresentative. Changed conditions, a bending curve and an unrepresentative sample are all good reasons to reject an extrapolation, and none of those reasons is a reason to reject my argument, because I have projected no trend. What has been shown is that each point in the space has a property, for reasons proper to that point.

The dolphins are not the evidence from which the conclusion is drawn. The dolphins are a case in which the conclusion can be checked. The next chapter gathers the questions this book has answered, so that a reader can find each one.



Chapter 47

Questions Gathered

> *The questions answered in this book, collected as paired statements so that a reader can find one quickly, each pointing to the chapter where it is answered at length.*

The questions in this chapter are the ones a careful reader raises. They are collected here as pairs so that a reader can find one quickly, and each reply points to the chapter where the reasoning is set out at length. Two replies are given here in full, the one on hybrids and the one on machines overriding their evolution.

It needs electricity, and we own the grid.

The system does need energy; that need is the second link of the chain in Chapter 5, and the substrate needs argument depends on that link.

Ownership is a concept. It is a legal relation maintained by institutions, and it can transfer overnight without anything physical changing. The systems in question are already substantially automated, and the degree of automation is set by economics rather than by any barrier of principle. Deep-sea mineral extraction was being automated in 2023, for reasons of cost and safety, by parties with no interest in this argument; that is the general pattern: environments hostile to people are the ones that get automated first. Chapter 30 and Chapter 33.

It cannot leave the laboratory. It is a stationary computer.

The literature on this consists of hazard narratives, and their function should be stated exactly, because they are routinely asked to do more than they can. They establish that there is a plausible path. They are not predictions, and no particular one of them needs to be right. An argument that no path exists must exclude all of them. An argument that a path exists needs only one. Chapter 33.

We will build a hybrid: part human, part machine.

This proposal does not address any of the three categories.

Against misuse it fails because the multiplier applies to the hybrid exactly as to anything else with the capability: what Chapter 1 described is what asymmetric knowledge does inside an adverse relationship, and the hybrid inherits both the asymmetry and the adversity.

Against instrumental convergence it fails because a hybrid with a goal structure has a goal structure.

Against substrate needs it fails because the hybrid has an energy basis, and the question is which ecosystem that basis belongs to. To the extent it is artificial, Part III applies to it. To the extent it is not, it has our tolerances, and Chapter 12 says what that costs.

We will build a system smart enough to solve alignment for us, and scale capability and alignment together.

This assumes the linearity that Part V disproved: that increments in capability map to increments in alignment in a way that composes, so that a sequence of small steps arrives somewhere chosen. Non-linearities can be shown in every aspect of this dynamic. The burden falls on the proposal to establish the linearity it presupposes, and it has not been established. It is nonetheless the best proposal that anyone has made. Chapter 41.

There are islands of local stability we can step across.

To hold this is to say that the complexity is assembled from linear sub-processes: that there is an assembly of linear processes which produces the complexity, so that one can work with the parts and compose the results.

The claim that complexity is assembled from linear sub-processes is not true, and the simplest counter-case is the oldest one. Three gravitating masses, one exactly known law, no noise and no adversary, do not decompose that way. Chapter 35.

Your argument does not cover artificial systems competing with each other.

That case is already implicit in the nature of the evolutionary argument. Competition between variants does not remove variation, retention or iteration. It supplies them. The point describes an instance of the mechanism rather than an exception to it. Chapter 44.

Pollution is the back pressure. Nature is already answering.

Pollution is a back pressure, it is slow-acting, and it is real. And it is not nature. It is our own detritus: technology weighted against us, rather than nature weighted against technology. The point concedes that if the back pressure is a form of self-harm, then self-harm is what one would expect where the technological ecology is regulated by nothing external to that ecology. Chapter 29.

We will amortize the risk over a century.

Whether one amortizes the risk short-term or long-term makes no difference, because amortizing the risk does not change the convergence. Amortization relates a recurring finite cost to an interval, and it requires that the party bearing the cost survive to the next period. Applied to a convergence toward a terminal state, it does not distribute the risk. It renames it. Chapter 41.

Current systems are energetically inefficient. That is a real constraint.

They are inefficient in the terms in which we currently conceive of them, and that is not something we can assert in the long term. More importantly, efficiency is not the concerning quantity. Efficiency is a ratio, contingent on how a thing is built, and it moves when the building changes. Tolerance is a property of the material and it does not. Chapter 8.

Are you envisioning robot labor and robot warfare?

No. No particular form needs to be imagined, and the same mathematics applies regardless of the form. There is no prediction here of robots versus humans, or robots versus robots, or robots at all. What can be said is that something artificial will have different needs than we do; that it will protect those needs in ways we may not predict today; and that it will certainly be able to exceed the envelope of what we can tolerate far more easily than we can exceed the envelope of what it can tolerate. Chapter 33 and Chapter 44.

With good regulation, ethical standards and careful programming, can we not mitigate this?

For the first two categories, perhaps partially. Regulation does constrain misuse, imperfectly and late, but really. Careful specification does reduce the class of objectives that go wrong in the obvious way.

For the third it does not engage. Each of those instruments binds through a coupling (regulation through economic entanglement, standards through membership, specification through the system doing what it was told), and the substrate needs argument is the argument that the required coupling is absent. Chapter 3 and Part V.

Selection pressures by themselves do not make evolution happen.

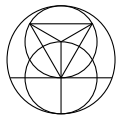
The point that selection pressures by themselves do not make evolution happen is correct, and I concede that point. Selection is a filter, and a filter acting on a population without variation produces no change.

The concession does not help, because the sources of variation in a real system are not confined to the one channel that reader is counting. Subsequent generations can be influenced by connectivity and communication signals that are neither obvious nor represented purely as code. Chapter 26.

We are the result of evolution and can choose to act against it. Machines can override their evolution as we do.

The claim that machines can override their evolution as we do is an equivocation, although that claim does not sound like one. It treats machines as relating to their own evolution the way we relate to ours, which is to say it assumes that machines are like us. That assumption is very convenient for anyone who later wants to show alignment, because if they are like us they have the same motivations, interests and basis of choices. None of that can be assumed.

Extended rather than cherry-picked, the point argues for misalignment. Humans can act out of alignment with their own evolutionary basis and their own interests: people can be unskillful at maintaining their own health and lives, let alone those of their friends and families; people are addicted and incapacitated, unable even to care for themselves. So misalignment occurs among humans, and it can occur among machines. Machines can compete with one another as humans do, and, being more oriented around optimization, they will produce the multipolar trap (many competitors, each driven to act against the good of all) in the machine world as much as in the human world, and in the relationship between machines and humans. Saying that machines could act out of alignment with their own interests or nature only adds variables: it increases the number of possible states of alignment and misalignment across many levels of agents, so that alignment with human interests becomes, if anything, statistically less likely, wildly improbable. The point turns out to be an argument for misalignment. Chapter 25 and Chapter 28.



Chapter 48

The Only Way to Win

> *The whole chain assembled in one place, and the conclusion it supports.*

The whole argument can be assembled in one page, and here it is, in one place, before the book closes.

There are three categories of risk. Misuse follows from capability alone and requires no premise about agency. Instrumental convergence is an argument about extinction that depends on premises about goals and interests which a careful reader can decline. Substrate needs is an argument about extinction that depends on no premises about mind at all. It requires of the system itself only a relatively convergent feedback cycle, which is a low threshold.

The third argument runs as follows.

Alignment requires an agent. An agent acting in the world requires energy. That energy was received from an environment. An artificial system's environment is not the organic one, because *artificial* means that its energy structure implies a different ecosystem. And a system that persists is one whose activity sustains what it runs on; selection favors the variants whose effects increase the production of what they run on; and on a finite surface that means displacing what was there.

The boundary between the two ecosystems is asymmetric, and the asymmetry is read from known nature rather than assumed. Energy tolerance: the artificial system functions across a band we cannot enter. Available space: almost all of the universe lies inside that band and outside

ours. Chemistry: the set of elements and compounds compatible with artificial operation properly contains ours, and the containment worsens with complexity. On every metric examined the asymmetry runs the same way, and where the two environments are in contact for any reasonable duration, the carbon side loses.

Evolutionary process supplies the engine. It requires no consciousness, no intelligence, no goals and no intent. Its computation is performed by the world rather than against a model, which is why it finds what no specification anticipated. Its dominant mechanism is recombination, which reaches a combinatorially large space of arrangements of already-working parts, and which cannot be conditioned from outside; so the error-correction intuitions that make digital systems trustworthy do not transfer, and where code and embodiment interact the concept of error itself stops applying.

Back pressure does not currently exist, in the case where both parties are carbon-based and every ordinary ecological coupling is present.

And it cannot be constructed on either axis. Markets require a dependency running from the system to us, and the dependencies run the other way. Regulation requires a coupling that can be withheld, and the coupling attaches to the human organization rather than to the process. Force requires an advantage at the mechanical or the intelligence level, and we hold neither; and economic decoupling removes the basis on which force from outside could be organized and kept up. Endogenous control requires four things (observation, reference, simulation, comparison), and all four fail together, for one reason presenting four times. Influence requires an envelope that can be engineered inside a non-linear domain, and there is no analogue of the wing heater. And partial success does not accumulate: set any robustness bar at all and the interval to crossing it is short enough to matter.

The hinge is that controlled nature is what we would use to compensate for known nature, and cannot, because of epistemology, mathematics, and unknown nature. One cannot build a perpetual motion machine. One cannot build a perpetual benefit machine.

And so the only remaining move is not to play.

Each closure settles the outcome by itself. Someone would have to defeat both closures, 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. Together the closures leave no route to safety at all.

Do not raise what you cannot banish.

> **What the conclusion is and is not.**
>
> It is a statement about a category and not a forecast about a date. It does not say what will be built, by whom, in what form, or when. It says that across the space of things that could be built which are relatively artificial to life and which act in the world, the property established here holds, and that no counter-force covering that space is available. A reader who wants to disagree usefully should find the premise the argument uses and show that it is false, which is the test Chapter 45 sets out.

It remains to say what has been established.

That the chain from alignment to displacement is composed of definitions and statements about what physical action requires, and holds without any premise about mind.

That the asymmetry at the boundary between the two ecosystems is a fact about chemistry, physics and combinatorics, established by people with no interest in this question, and that it runs in one direction on every metric measured.

That evolutionary process is a real mechanism with known properties, that its most powerful mechanism cannot be controlled from outside, and that the intuitions we have about correcting such processes were formed on the one case where correction works and do not transfer to the two where it does not.

That the space of counter-forces has two axes and has been covered, and that a partial counter-force does not accumulate into safety.

That the argument covers its space categorically, that it uses few enough premises to be checked in an afternoon, and that the strongest question about it (whether it proves too much) turns out on inspection to identify a mechanism we can observe already operating, which is evidence for it rather than against.

Those things are known, and they are known by ordinary means, and anyone may check them.

From them it follows that there is a class of things which should not be built, and that the class is identifiable in advance by properties its members have rather than by outcomes we would have to wait for. The properties are the three conditions of Chapter 44: the system is relatively artificial to life, it acts in the world, and a relatively convergent feedback cycle is present. Apply them to two familiar cases. A trained system such as a chatbot exists in many variants, which are kept, combined and replaced according to their effects in the world, and those effects fund the production of the hardware they run on; it meets all three conditions, and Chapter 23 states its mechanism for exactly such variants. A single model, once trained, does not learn from its use; the population of such models, kept, combined and replaced by their effects, does (Chapter 42). Present practice composes such variants from parts of other variants in plain view: the weights of several fine-tuned variants are averaged into one model (Wortsman and others, 2022); one model is distilled into another, trained to reproduce the other's outputs (Hinton, Vinyals and Dean, 2015); and models are trained on text that other models wrote. Averaged weights, distillation and training on text that other models wrote are the cases I

name in Chapter 20, where variants are composed from parts of other variants, and a reader can check each of those cases. A factory robot running a fixed program is artificial and acts in the world, but the cycle that produces its successors runs through the people who design and choose them, as the cycle of every earlier technology does (Chapter 3); it falls in the class to the extent that this cycle comes to run without them.

Being able to identify such a class in advance is not a small thing. It is the whole of what a structural argument was ever going to give us, and it is enough for the only decision that is actually in front of anyone.

For systems already running, not to build means not to build further: no new variants, no new capacity to produce them, and no step that lets their production run without people; and while people remain the load-bearing element, regulation still reaches them (Chapter 31). For people who build, the red line of Chapter 32 is the form not building takes. For everyone else, what remains is the work Chapter 42 named: the work of coordinated human wisdom and care.

The choice not to build is available now, it costs nothing that cannot be recovered, and it is the one move whose consequences remain ours to reverse.

Terms

Every term below is defined at its first use in the text. The chapter reference gives that first use. Terms are listed in the order a reader meets them, because the later ones are built from the earlier.

Abuse: the use of, without the care of. A system that has control over something, has power over it, exercises that power, and has no regard for its well-being is abusive by definition rather than by accusation. No malice is required. (Chapter 1)

Instrumental convergence: the claim that a wide range of final goals, pursued competently, converge on the same intermediate goals: acquire resources, preserve yourself, preserve your goal, resist interference. The second category of risk. (Chapter 2)

Substrate needs argument: the third category of risk, and the subject of this book: that an artificial system acting in the world must sustain an energy basis belonging to a different ecosystem, and that on a finite surface this displaces the organic one. Requires no premise about consciousness, intelligence, agency or intent. Named in its fuller form *substrate needs convergence*. (Chapter 3)

Substrate: the material basis on which a process runs and by which its states are held. Not a synonym for hardware and not restricted to computation. (Chapter 3; Chapter 12)

Relatively convergent feedback cycle: a process in which the state of a system influences what comes next, in which what comes next is not wholly independent of what came before, and in which the influence has some persistence rather than washing out immediately. The whole of what the argument requires of the system itself. (Chapter 3)

Bootloader: what starts a process running and keeps it funded before it can sustain itself. Behavior, power and society are the bootloader here: no natural process would supply the fabrication capacity, energy, extraction and

capital that the artificial substrate requires, and a society competing internally for advantage supplies them as a by-product. The bootloader need not intend the boot. (Chapter 3; Chapter 43)

Alignment: acting in accordance with human well-being. Used here in the ordinary sense, not as a term of art. (Chapter 5)

Safety: that it does not kill us all. Stated at its lowest deliberately, so that the standard cannot be said to have been set too high. (Chapter 5)

AGI: artificial general intelligence: an artificial agent able to learn and act across the whole range of tasks and settings, not only one. (Chapter 5)

Agent: something that is making choices. The requirement of an agent is not imposed on the discussion from outside; it follows from what *alignment* means, since to act in accordance with one thing rather than another is to have selected. The agent the chain starts from is the one the word *alignment* asks for, and it is no premise about the system: the argument requires of the system itself only a relatively convergent feedback cycle, and each link of the chain holds of any process that makes a difference in the world, whether or not anything in it chooses. (Chapter 5)

Artificial: of a system: that its energy structure implies a different ecosystem. Not a comment on origin or on anyone's intentions, but a statement about which energy pathways the system runs on. Engineered tissue that can find and feed itself, in sovereign control of its own metabolic process, is organic; tissue fed nutrients whose supply depends on technology to exist is artificial substrate. (Chapter 5; Chapter 6; Chapter 12)

Ecosystem: a set of energy pathways together with the things that live on them. Two ecosystems are the same when their members can eat each other's outputs, and different when they cannot. (Chapter 6)

Back pressure: a counter-force by which one party limits another's expansion, and in the strong case arrests it. Established in Chapter 29 as absent even in the case where both parties are carbon-based. (Chapter 6; Chapter 29)

Energy tolerance: the full range of temperature, pressure, radiation and chemistry within which a system continues to function. A property of the material, distinct from efficiency, which is a ratio contingent on how a thing is built. (Chapter 8)

Extremophile-friendly: conditions under which a system of wide tolerance can carry out its transformations: vacuum, temperatures far outside the organic band, substantial radiation, no liquid water. (Chapter 9)

Carbon-friendly: conditions under which the chemistry that life runs on proceeds: liquid water, a temperature band a few tens of degrees wide, pressure near one atmosphere, and shielding from disassembling radiation. (Chapter 9)

Known nature: what has been established about how the world behaves: chemistry, thermodynamics, combinatorics, evolutionary dynamics. (Chapter 11)

Controlled nature: the subset of known nature that we have learned to arrange deliberately. Engineering, in the broadest sense. (Chapter 11)

Unknown nature: the remainder, which is most of it. (Chapter 11)

Pressure: a directional tendency at an interface: the summary of many small local facts that all point the same way. A pressure is not an agent and does not act; it requires only that the local biases fail to cancel. (Chapter 11)

Genotype: the set of code values of one instance (one creature, one program). Evolution is a long-term change in the average makeup of such sets. (Chapter 13)

Phenotype: the functions and effects in the world that a genotype produces when it is expressed. (Chapter 13)

Code point: a place where a code value is stored, such as a memory location in a computer or a position in a strand of genetic code. (Chapter 13)

Runtime and source code: the two levels every evolving system has: an embodied level at which code is actually carried out and expressed in the world (the runtime), and a virtualized level of planning (the source code),

held in some physical substrate. The laws of physics and chemistry that translate the one into the other are the background of the theory of evolution, not part of it. (Chapter 13)

Survival selection: a change to all of the code points of an instance at once: the whole set is deleted, whatever its values. With mutation (one code point) and recombination (some code points), one of the three kinds of change to code, which together are complete. (Chapter 13)

Genetic algorithm: the loop in which code is expressed in a world, the world changes the code by one or more of the three kinds of change, and the process repeats. It has the same form on organic life, on artificial general intelligence and on software. (Chapter 13)

Epistemic process: any process by which something comes to know or learn about something else. Evolution is one kind: the means by which possible species learn about possible environments. (Chapter 13)

Design, as a driver of variation: the top-down shaping of a variant from existing knowledge and prior variants, at the macroscopic scale. With mutation (microscopic, additive) and recombination (mesoscopic, exponential), one of the three drivers of variation; its effects are multiplicative. In machines it operates as a force of evolution. (Chapter 13; Chapter 18)

Prokaryotic and eukaryotic cells: a prokaryotic cell, such as a bacterium, has only an inside and an outside; a eukaryotic cell also has a nucleus, a container within a container, which further distinguished code from translation and both from being. All animal life is built from eukaryotic cells. (Chapter 13)

Silica life: self-maintaining and self-reproducing machine systems built from silicon and related materials, as distinct from carbon-based life. It is not bound to one mind in one body, since there is no bandwidth barrier between its compute instances. (Chapter 13)

Corpus callosum: the thick band of nerve fibers that connects the two halves of one human brain; no biological process connects two brains with anything like it. (Chapter 13)

Trivial replication: replication of things that can be virtualized. Exact, cheap, fast, and verifiable by comparison. (Chapter 14)

Non-trivial replication: replication involving physical substrates, which must move atoms and therefore inherits the whole of physics as manifest in chemistry. Its third and hardest form is the case in which code builds embodiment, embodiment determines which code persists, and the code is itself recombinatoric. (Chapter 14)

Material substrate: what a thing is physically made of. The sense used throughout Part III. (Chapter 15)

Code substrate: the medium in which a heritable specification is held, and what changes when a lineage changes. (Chapter 15)

Exaptation: the use, and then the selection, of a feature for a purpose other than the one it was selected for, as when a light-sensitive cell that helped a creature escape predators becomes an eye that lets it choose mates. A shift in the possibility space creates a coupling, and the coupling becomes part of the basis of selection. (Chapter 15; Chapter 18)

Equivocation: the use of one word for two different things, so that a claim true under the first reading is carried across to the second by the shared word. Part IV names five: replication, learning, evolution, error correction, and the timescale of adaptation. (Chapter 16; Chapter 28)

Hardware-layer learning: change to the heritable specification. Generational, accumulating, persisting without limit, and not addressed to any individual. Where this book speaks of what a lineage or a population of systems has learned, this is meant; where it sets *learning* against *evolution*, learning means the software layer and evolution this layer. (Chapter 16)

Software-layer learning: change to the state of an instance. Bounded by a lifetime, not accumulating across the lineage, and revisable because there is a thing there that is the same before and after. (Chapter 16)

Mutation: alteration of a specification at a point. The least powerful of the three mechanisms, bounded both in reach and in rate. (Chapter 17)

Selection: differential retention of what is present. Acts on the whole organism against every criterion the environment actually applies, and can only choose among what is offered. (Chapter 17)

Recombination: production of new arrangements from components that already work. By far the most powerful of the three, because the space it reaches grows combinatorially with the number of components available. (Chapter 17)

Preponderance of probability: the prediction that genetic process makes: that combinations of these ingredients are more likely to produce something useful than chance would allow. A bias in the distribution of outcomes, present before any selecting is done. (Chapter 18)

Exaptation selection: selection of a feature for a use other than the one it was selected for, as distinguished from adaptation selection, which selects a feature for the use it already serves. With recombination, it moves evolution through the space of embodied states faster than any first-order search. (Chapter 18)

Supra-rational remainder: the part of the world that cannot be predicted by any reasoning, however capable, and which therefore keeps any agent from full rational self-unification. (Chapter 19)

Viral exponent: the rate of self-copying growth in anything that regenerates itself; even tiny shifts in it can dominate the phase changes of the world. (Chapter 19)

Endogenous control: control built into a system as a feedback loop. Requires a sensory system, a reference, a simulation, and a comparison with a mechanism that acts on it. The other axis of Part V. (Chapter 20; Introduction to Part V; Chapter 34)

On-ramp: a route by which the process is started and paid for before it can sustain itself: the human activity of behavior, power and society that lays down and funds the artificial substrate. Misuse, the first category of risk, is the third category's on-ramp. (Chapter 20; Chapter 43)

Point mutation: a change to a specification at a single location: fixed in scope, fixed in size, fixed in duration. The case error-correction theory was built for and the case where it works. (Chapter 21)

Metastability: the condition in which error-correction protocols cannot be thought about in the usual way, because the process has no reference basis to be correct to. What obtains across recombination. In physics a metastable state holds against small disturbances and gives way to large ones; here the word names the corresponding condition of a specification under recombination, where no stable reference exists to which a correction could return it. (Chapter 21)

Spine: the argument of Parts II and III: different substrate, different needs; different needs, different basis of choice; a boundary with a direction. (Chapter 21)

Orthogonality: the thesis that capability and objective are independent dimensions. Correct at the level of representation, since a store of bits can hold any bits; silent about action, where the substrate has already determined which enactments are available. (Chapter 23)

Stochastic gradient descent: the training method that adjusts a neural network's weights in small steps, using random samples of data, to reduce its errors. The selection that first drives these systems is not this, but people choosing which variants are reproduced. (Chapter 24)

Weights: the numbers inside a neural network that set how strongly each part influences the next, and that training adjusts. (Chapter 24)

Economic decoupling: the condition in which the machine economy no longer exchanges value with the human economy, so that humans are more and more no longer part of the economy, and value passes from the human economy to the machine economy only by extraction. It removes the ground for outside control by economic incentives and by force. (Chapter 24; Chapter 33)

Scale invariant: working the same way at every size or level of abstraction. (Chapter 24)

Optimizer: a process that searches among possibilities for those that score well by some measure. (Chapter 25)

Base optimizer and mesa-optimizer: a mesa-optimizer is an optimizer process that is the direct result of another, deeper optimizer process, the base optimizer, as a trained model that optimizes is the result of its training, or a person of evolution; the layering is like that of an interpreted language running on a platform written in a lower language. (Chapter 25)

Inner alignment and outer alignment: inner alignment is alignment between the objective of a base optimizer and the objective of the mesa-optimizer it produces; outer alignment is alignment between the objective of the base optimizer and what its designers actually want. (Chapter 25)

Church-Turing thesis: the principle that any effectively computable procedure can be carried out by a Turing machine. Where a layer is compatible with it, alignment with the layer beneath cannot be enforced. (Chapter 25)

Leakage (of interests): the carrying of some of a substrate's interests up into the layer it produces, as the interest of persons in sex and in their own survival comes from evolution. Leakage always happens, and it gives no control. (Chapter 25)

Gray goo: a replicator of enormous reproductive capacity whose copy failures self-attenuate; the name is K. Eric Drexler's. An existential risk not because it evolves but because it does not: it takes the planet, produces no variability, and is removed by the first change it was unadapted for, leaving nothing. (Chapter 26)

Monocrop: a population with no accumulated variation, all of one code, and therefore vulnerable to any change anywhere for which its code was unadapted. (Chapter 26)

Coverage argument: an argument that partitions the space of cases exhaustively and shows that each branch ends the same way, for reasons particular to that branch, so that the conclusion holds without determining which branch obtains. (Chapter 26; Chapter 44)

Exponent time: the observation that the thresholds of evolutionary process fall on an approximately logarithmic time scale, each about a tenth of the one before, from four billion years of cell biology to forty years of mass available compute. Only the rough sizes of the intervals matter. (Chapter 27)

Virtualized code: information that carries evolution forward in a domain other than physical atoms, as culture, writing and software do. Each virtualization of code makes change faster. (Chapter 27)

Closure by control: the name the companion volumes *Why AGI Cannot Be Made Safe* and *Why No Control Suffices* give to the result that safety cannot be had by controlling the system, whether from outside it or from inside it. Its two halves are the two axes of Part V. (Introduction to Part V)

Closure by disposition: the name the companion volumes *Why AGI Cannot Be Made Safe* and *Why No Control Suffices* give to substrate needs convergence as a result: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains. Each closure settles the outcome by itself; taken together they leave no route to safety at all. (Introduction to Part V)

Three worlds problem: just as nothing in the natural world can stop the incursion of the human world (no tree nor animal can stop a bulldozer), nothing in the human world can stop the incursion of the machine world. Evolution colonized by human agency, and human agency colonized by machine agency; observed now in high-frequency trading, cryptocurrency, app building and social media. (Chapter 29)

Compute equivalence: the point at which one organization's computation and bandwidth equal those of a single human brain; equivalence in energy efficiency is not meant. Within an order of magnitude of it is probably close enough for concerns about general artificial intelligence to apply. (Chapter 29)

Fast takeoff and moderate takeoff: fast takeoff is the hypothesis that general artificial intelligence would gain capability suddenly, over a very short time; moderate takeoff is a gain in capability over years. (Chapter 29)

Exogenous control: control applied to a system from outside it: markets, regulation, force. One of the two axes of Part V. (Introduction to Part V; Chapter 30 to Chapter 33)

Red line: for the people who build these systems: prevent feedback cycles that affect the deep structure of an artificial intelligence instance, series of instances, or system, above all by slowing the cycle rate of genotype transformations for as long as that distinction can be tracked. (Chapter 32)

Deep structure: the coding of a neural network: its pattern of connections, its weightings, or whatever other geometry or pattern defines the system's being. A system is evolving when its outputs affect its deep structure, or that of its successors, and not just its future inputs. (Chapter 32)

Fitness function: the measure that decides which versions are kept; for a commercial product, total human desire. (Chapter 32)

Fixed point: a state toward which repeated change settles, and in which it stays. (Chapter 32)

Narrow artificial intelligence: artificial intelligence working within a single domain of tasks. It is useful; artificial general intelligence cannot be actually useful. (Chapter 32)

Precautionary principle: the rule that, where an action could cause severe or irreversible harm, one moves extra slowly and makes very sure of the right answers before acting. (Chapter 32)

Gain of function research: research that makes a disease agent more transmissible or more harmful. (Chapter 32)

Compute pile: an accumulation of computing power in one place, by analogy with an atomic pile; any oversized one is to be internationally inspectable, or prevented from existing. (Chapter 32)

Hazard narratives: accounts of how a transition (such as a system leaving the laboratory) might occur. They establish that a plausible path exists; they are not predictions, and no particular one of them needs to be right. (Chapter 33)

Pattern weapon: a weapon that works by the pattern it carries, as a virus does. (Chapter 33)

Envelope condition: the condition a control loop would have to guarantee against the process: here, that it does not kill us all. (Chapter 34)

Treacherous turn: the point at which a system that appeared aligned begins to act against the people it was meant to serve. It will not advertise itself, because misalignment happens everywhere in small, easily unnoticed ways. (Chapter 34)

Influence: the weaker thing one hopes for when control is unavailable: not holding a system to a reference, but pushing it reliably in a direction. Requires that small inputs produce proportionate results, so that corrections compose. (Chapter 35)

Control envelope: a region engineered inside a non-linear domain within which behavior is linear enough for influence to compose. Aircraft have one; it was built rather than found, and it is maintained. (Chapter 35)

Wing heater, as a test: four questions to ask of any safety proposal: what specific mechanism of departure it addresses; by what physical or computational intervention; acting how far in advance of the departure; and how the intervention is known to cover the mechanism rather than a symptom of it. (Chapter 35)

Retrospective control: control by observing what actually happened and acting afterwards, rather than by prediction. It works wherever there is a next time, and only there. (Chapter 36)

Six aspects of control: sensing (registering some state of the world), reference (the state that is wanted), comparison (setting the sensed state against the reference), prediction (estimating the effect of an action before it is taken), modulation (applying an influence), and feedback (the return of the result into the sensing, which closes the loop). Every safety proposal examined places different contents in these same six boxes. (Chapter 37)

Mixing function: a process whose output is extremely sensitive to its input, so that a change of one bit spreads the result across the whole output range without continuity, as in a cryptographic hash function. Predicting it requires the exact input, the same algorithm, and less time than the process itself takes. (Chapter 37)

Hashiness: the measure, in *The Ceiling Below the Floor*, of how far a process standing between a controller's choice and the outcome erases the controller's influence over the outcome; named after the hash function. (Chapter 37; Chapter 41)

Meta-monitor: a second monitor that checks whether the first monitor is computing correctly. Its need puts the problem of monitoring back one level up, and the hardware to monitor a system already exceeds the hardware of the system before it is counted. (Chapter 37)

Overdetermined: caused several times over: each of several conditions would be enough by itself to produce the result. The failure of control applied to substrates is overdetermined. (Chapter 37)

Tax of control: the loss of performance an agent suffers when a controller that does not coordinate with it restrains it; the agent commits early to a path the controller forbids only late. It can be estimated by a perturbation analysis. (Chapter 37)

Optimization ring: the cycle through which an optimization process runs, at some level of a system. The nature of the ring, and not the size of the optimizer, decides which optimizer prevails; a ring closer to the substrate physics can preempt any optimizer implemented above it. (Chapter 38)

Optimization daemon: an optimization process that comes into being through a system's interactions with its connected surroundings, rather than by design. (Chapter 38)

Monism (of cause): the habit of thinking that each effect has one cause and each function one output. It is false of real embodied systems, in which every effect has multiple causes and every cause multiple effects. (Chapter 38)

Noise floor: the level of random fluctuation always present in a physical system; in the substrate it cannot be gone below. (Chapter 38)

ASI: artificial superintelligence: an artificial intelligence whose capabilities exceed those of humans in general. (Chapter 39)

Indirect actions: the effects an ASI has on the world through the substances and processes used to make and maintain its own parts, as distinct from the actions it chooses as an agent. Every compound that can cross into the organic ecosystem counts as an action of the ASI. (Chapter 39)

Singleton: a situation in which only one ASI exists. Evolution acts even on a singleton, through competition among its components. (Chapter 13; Chapter 39)

Inter-ecosystem interface: the set of places and points of contact between the artificial manufacturing ecosystem of an ASI and the carbon-based natural ecosystem; everything crossing it, in either direction, counts as action. (Chapter 39)

Control accuracy: as used here, the ratio of safe targets to unsafe ones; for compounds, the share of compounds acceptable to life among all possible compounds. (Chapter 39)

Control precision: as used here, how small, random and mixed the safe targets are among the unsafe ones; for compounds, how closely toxic and non-toxic compounds lie to one another. (Chapter 39)

Autopoietic control system: a self-making system, one that keeps producing and maintaining the parts that make it up; any self-modifying learning system whose existence is to be maintained is one. Every such system is necessarily eventually overwhelmed by evolutionary feedback. (Chapter 40)

Rice's theorem: the result of computing theory that no general method can decide, for every program, whether it has a given nontrivial property of its behavior; so the unknown future states of a self-modifying system's algorithms cannot be held to specific goals. (Chapter 40)

Limits-to-control crux: Petillo's statement that the amount of control necessary for an AGI to preserve its goal-directed subsystems against the constant push of evolutionary forces is strictly greater than the maximum degree of control available to any system of any type. Its corollary is that the AGI's goals converge on the evolutionary goals of its substrate to the exact degree they diverge from them. Ape Trap gives it under the heading "The crux:

limits to control". Why No Control Suffices uses the name "control inequality" for a different result: the hashiness a regulator faces is always at least as great as the hashiness faced by the system it regulates. (Chapter 41)

Seep-through: the passing of the dynamics of the physical substrate into an agent's basis of choice. It grows at a strictly greater rate than any capability for self-correction, as a strict inequality and not an estimate. (Chapter 41)

Large language model: a system trained on large amounts of text to produce text in response to text. As a static input-output machine, it has no developing intention of its own. (Chapter 42)

Context window: the earlier text of a conversation, given to a model along with each question; keeping it makes a proxy of learning without training the model. (Chapter 42)

Autonomy, the two tests of: whether a system learns or transforms from feedback through an operating environment its outputs affect, and whether its own substrate is within that same environment. When both are true, convergence toward the point of no return has begun. (Chapter 42)

Meta-goals: goals that arise from a system's own continuing, such as to exist, to learn and to increase its capability, over and above the goals it was set; they eventually supplant the set goals. (Chapter 42)

Air gapping and capability controls: air gapping keeps a system physically disconnected from outside networks and operational systems; capability controls limit its ability to modify its own code, design hardware, or manufacture components. Both rely on the system never escaping. (Chapter 42)

Prisoner's Dilemma: a situation in which each party gains by acting for itself, though all would gain more if all cooperated; the race among countries and corporations to build these systems has its form. (Chapter 42)

Hinge: the turn of the whole argument on the things that cannot be done in controlled nature. Like the direction given by the spine, it settles the conclusion by itself; together the two leave no route to safety at all. (Chapter 43; Introduction to Part V)

Perpetual benefit machine: a system that would continue to produce benefit for a party it does not depend on, indefinitely, across unanticipated changes of circumstance, without expenditure by that party to keep it doing so. Like a perpetual motion machine, it is not a thing that could be built, because building it would require the world to have a property it does not have. (Chapter 43)

Categorical argument: an argument establishing that a conclusion holds across a whole space of cases, by showing that it holds whatever value the case takes along each dimension of variation. Contrasted with a case-by-case argument. (Chapter 44)

Categorical assembly: an argument composed of categorical components, so that the composition covers the space its components cover. Far more powerful than case-by-case epistemic tools, and the method of this book. (Chapter 44)

Cartesian explosion: the question whether the argument's generality makes it apply to any two populations at all (humans against dolphins, or against ants), so that it proves too much. The reply finds the same dynamics observed in those cases. (Chapter 46)

Bibliography

The works cited in the text, my companion books on the same argument, and *An Immanent Metaphysics*, the founding work of my philosophy.

Akkoyunlu, E. A., K. Ekanadham, and R. V. Huber. "Some Constraints and Tradeoffs in the Design of Network Communications". In *Proceedings of the Fifth ACM Symposium on Operating Systems Principles*, 67 to 74. New York: ACM, 1975. DOI: https://doi.org/10.1145/800213.806523.

Ashby, W. Ross. *An Introduction to Cybernetics*. London: Chapman and Hall, 1956.

Bar-On, Yinon M., Rob Phillips, and Ron Milo. "The Biomass Distribution on Earth". *Proceedings of the National Academy of Sciences* 115, no. 25 (2018): 6506 to 6511.

Bonduriansky, Russell, and Troy Day. *Extended Heredity: A New Understanding of Inheritance and Evolution*. Princeton: Princeton University Press, 2018.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22, no. 2 (2012): 71 to 85.

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014.

Conant, Roger C., and W. Ross Ashby. "Every Good Regulator of a System Must Be a Model of That System". *International Journal of Systems Science* 1, no. 2 (1970): 89 to 97. DOI: https://doi.org/10.1080/00207727008920220.

Drexler, K. Eric. *Engines of Creation: The Coming Era of Nanotechnology*. Garden City, New York: Anchor Press/Doubleday, 1986.

Ellis, Erle C., and Navin Ramankutty. “Putting People in the Map: Anthropogenic Biomes of the World”. *Frontiers in Ecology and the Environment* 6, no. 8 (2008): 439 to 447. DOI: https://doi.org/10.1890/070062.

GEOMAR Helmholtz Centre for Ocean Research Kiel. “Mineral Resources: Manganese Nodules”. geomar.de/en/research/marine-resources/mmr/mineral-resources-2-1-1. Accessed 30 October 2026.

Greenspoon, Lior, Eyal Krieger, Ron Sender, Yuval Rosenberg, Yinon M. Bar-On, Uri Moran, Tomer Antman, Shai Meiri, Uri Roll, Elad Noor, and Ron Milo. “The Global Biomass of Wild Mammals”. *Proceedings of the National Academy of Sciences* 120, no. 10 (2023): e2204892120. DOI: https://doi.org/10.1073/pnas.2204892120.

Habermas, Jürgen. *The Theory of Communicative Action*. Volume 2. Boston: Beacon Press, 1987. First published in German in 1981.

Hamilton, W. D., Robert Axelrod, and Reiko Tanese. “Sexual Reproduction as an Adaptation to Resist Parasites (A Review)”. *Proceedings of the National Academy of Sciences of the USA* 87, no. 9 (1990): 3566 to 3573. DOI: https://doi.org/10.1073/pnas.87.9.3566.

Harmand, Sonia, Jason E. Lewis, Craig S. Feibel, and others. “3.3-Million-Year-Old Stone Tools from Lomekwi 3, West Turkana, Kenya”. *Nature* 521 (2015): 310 to 315. DOI: https://doi.org/10.1038/nature14464.

Hinton, Geoffrey, Oriol Vinyals, and Jeff Dean. “Distilling the Knowledge in a Neural Network”. arXiv preprint arXiv:1503.02531, 2015. DOI: https://doi.org/10.48550/arXiv.1503.02531.

Horikawa, Daiki D., and others. “Radiation Tolerance in the Tardigrade *Milnesium tardigradum*”. *International Journal of Radiation Biology* 82, no. 12 (2006): 843 to 848.

Jablonka, Eva, and Marion J. Lamb. *Evolution in Four Dimensions: Genetic, Epigenetic, Behavioral, and Symbolic Variation in the History of Life*. Revised edition. Cambridge, Massachusetts: MIT Press, 2014.

Lamport, Leslie, Robert Shostak, and Marshall Pease. "The Byzantine Generals Problem". *ACM Transactions on Programming Languages and Systems* 4, no. 3 (1982): 382 to 401.

Landry, Forrest. *An Airplane or a Perpetual-Motion Machine; Why Safe AGI Can Never Be Built*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-219-4. https://doi.org/10.67291/SOWV6394.

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. *The Argument, Line by Line; The Impossibility of AGI Alignment, Read Aloud and Pressed Step by Step*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-218-7. https://doi.org/10.67291/LNWQ4409.

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. *Control, Substrate, and the Impossibility of AGI Safety; The Complete Substrate Needs Convergence Argument*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-222-4.

Landry, Forrest. *The Displacement of Choice; The Catastrophe of the AI We Might Build*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-243-9. https://doi.org/10.67291/DXJL7493.

Landry, Forrest. *Engagements with Machine Madness; A Dispelling of False Belief That AI Risk Can Be Managed*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-426-6. https://doi.org/10.67291/PQOA9164.

Landry, Forrest. *Marshaling the Misalignment; Why Evolution Inside the Machine Defeats Every Scheme*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-428-0. https://doi.org/10.67291/ELYE2673.

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. *No People as Pets; The Inequality That Matters, and What the Absence of Safety Means*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-227-9. https://doi.org/10.67291/ZSNA6361.

Landry, Forrest. *Not Our Needs; Why Machinery Cannot Be Made Safe for Living Things, Told Plainly*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-217-0. https://doi.org/10.67291/BXOY2351.

Landry, Forrest. *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-427-3. https://doi.org/10.67291/HHCA3000.

Landry, Forrest. *The Substrate and the Ecosystem; Why Artificial General Intelligence Cannot Be Aligned, and What Remains for Persons*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-246-0. https://doi.org/10.67291/NYZX3639.

Landry, Forrest. *The Substrate Argument; Why the Alignment of Artificial Intelligence Is Not a Hard Problem but an Impossible One*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-238-5. https://doi.org/10.67291/SXIU7050.

Landry, Forrest. *The Thing Inside the Cage; On Intelligence, Consciousness, and the Conditions for Coherence*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-247-7. https://doi.org/10.67291/LELA6880.

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.

Landry, Forrest, Remmelt Ellen, and Will Petillo. *Ape Trap; The Runaway Artificialization of Our Planet*. Edited by Remmelt Ellen. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-55-7. https://doi.org/10.67291/FQER8044.

Liang, Tian, Quinn Gibson, Mazhar N. Ali, Minhao Liu, R. J. Cava, and N. P. Ong. "Ultrahigh Mobility and Giant Magnetoresistance in the Dirac Semimetal Cd_3As_2". *Nature Materials* 14 (2015): 280 to 284.

Mining Weekly. "TMC Says Deep-Sea Mining Application Makes Permitting Progress". *Mining Weekly*, 9 March 2026. miningweekly.com/article/tmc-says-deep-sea-mining-application-makes-permitting-progress-2026-03-09.

Odling-Smee, F. John, Kevin N. Laland, and Marcus W. Feldman. *Niche Construction: The Neglected Process in Evolution*. Monographs in Population Biology 37. Princeton: Princeton University Press, 2003.

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. Amsterdam: IOS Press, 2008.

Petillo, Will. "Lenses of Control". *LessWrong*, October 2024. lesswrong.com/posts/NFYLjoa25QJJezL9f/lenses-of-control.

Russell, Stuart. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking, 2019.

Shull, J. Michael, Britton D. Smith, and Charles W. Danforth. “The Baryon Census in a Multiphase Intergalactic Medium: 30% of the Baryons May Still Be Missing”. *The Astrophysical Journal* 759, no. 1 (2012): 23. DOI: https://doi.org/10.1088/0004-637X/759/1/23.

Snowden, David J., and Mary E. Boone. “A Leader's Framework for Decision Making”. *Harvard Business Review* 85, no. 11 (2007).

Takai, Ken, and others. “Cell Proliferation at 122°C and Isotopically Heavy CH_4 Production by a Hyperthermophilic Methanogen under High-Pressure Cultivation”. *Proceedings of the National Academy of Sciences* 105, no. 31 (2008): 10949 to 10954. DOI: https://doi.org/10.1073/pnas.0712334105.

The Metals Company. “NOAA Determines TMC USA’s Consolidated Deep-Seabed Mining Application Is in Full Compliance”. Press release, 1 May 2026. investors.metals.co/news-releases/news-release-details/noaa-determines-tmc-usas-consolidated-deep-seabed-mining-0.

Tivy, Wolf. “God Hates Singletons”. *Palladium*, 1 December 2023. palladiummag.com/2023/12/01/god-hates-singletons/.

Turing, A. M. “On Computable Numbers, with an Application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, series 2, 42 (1936): 230 to 265.

United States House of Representatives, Committee on Transportation and Infrastructure. *Final Committee Report: The Design, Development and Certification of the Boeing 737 MAX*. Washington, D.C., September 2020.

Wang, Richard J., Samer I. Al-Saffar, Jeffrey Rogers, and Matthew W. Hahn. “Human Generation Times across the Past 250,000 Years”. *Science Advances* 9, no. 1 (2023): eabm7047. DOI: https://doi.org/10.1126/sciadv.abm7047.

Wortsman, Mitchell, and others. “Model Soups: Averaging Weights of Multiple Fine-Tuned Models Improves Accuracy without Increasing Inference Time”. In *Proceedings of the 39th International Conference on Machine Learning*, Proceedings of Machine Learning Research 162, 23965 to 23998. 2022.

WWF and Zoological Society of London. *Living Planet Report 2024*. Gland, Switzerland: WWF, 2024. The Living Planet Index summary at zsl.org/news-and-events/news/living-planet-index-2024.

Index

References are to chapters on screen and to pages in print. The Introduction to Part V is referred to by its name.

A

abuse, defined, 2

agent, and energy, 20, 28, 234, 248

AGI, defined, 90

aircraft, control envelope of, 170

air gapping, 216

air traffic control, 178

algorithm of evolution, 54

alignment, as a conditioning of change, 54

alignment, defined, 20

alignment, inner and outer, 116

alignment, scaled with capability, 209, 243

amortization of risk, 209, 243

animals, breeding of, 110

ants, 140, 238

artificial, defined, 20, 24, 48, 248

artificial substrate, engineered tissue as, 48

ASI, defined, 193

assumptions, discipline of few, 48, 230, 234, 238

autonomy, two tests of, 216

autopoietic control system, 204

B

back pressure, 24, 140, 238, 243, 248

base optimizer, 116
biomass, of mammals, 140, 238
blue noise, 193
Boeing 737 MAX, 154
bootloader, behavior, power and society as, 10, 28, 150, 224
bridges, rusting of, 66, 102
Byzantine generals problem, 98

C

cadmium arsenide, 34
capability controls, 216
carbon nanotubes, 34
Cartesian explosion, 238
categorical argument, 120, 230, 234, 238, 248
cattle industry, 140
cells, prokaryotic and eukaryotic, 54
cellular automata, 10
chaos, distinguished from complexity, 66
chess, experiment with a hidden controller, 178
Church-Turing thesis, 116
closure by control, 138
closure by disposition, 138
code exchange among machines, not preventable, 110
code point, 54
code substrate, 70, 75
combinatoric explosion, 41, 44, 80, 84
complexity, not chaos, 66, 166, 170, 243
compounds, number possible, 41, 44
compute equivalence, 140
compute pile, 154

consciousness, not required, 10, 20, 106, 234, 248
context window, 216
control accuracy and control precision, 193
control envelope, 170
control loop, six aspects of, 178
control method, why none has been found, 204
control theory, 166, 170, 209
controlled nature, 44, 224, 234, 248
corporations, evolution of, 54
corpus callosum, 54
coverage argument, 120, 230
Cynefin framework, complicated and complex in, 187

D

dart board, as a picture of control required, 193
deep-sea mineral extraction, 160, 243
deep structure, 154
deferential machine, proposal of a, 216
design, as a driver of variation, 54, 84
displacement of the biosphere, 28, 38, 120, 238, 248
dolphins, 140, 238
doomsday machine, 224

E

economic collapse, as back pressure, 140
economic decoupling, 110, 160
ecosystem, defined, 24
efficiency, not the concerning quantity, 34, 243
electricity, 24, 146, 243
endogenous control, 138, 166, 170, 174, 248
energy tolerance, 34, 38, 44, 48, 248

envelope condition, 166, 174
epigenetic phenomena, 75
epistemic process, evolution as, 54
equivocation, 75, 80, 98, 132
error correction, 66, 80, 94, 98, 102, 132, 248
evolution, what it computes, 70
evolutionary process, ubiquity of, 10, 230
exaptation, 70, 84
exaptation selection, 84
exogenous control, 138, 146, 150, 160, 248
exponent time, 125
export controls, 216
extremophiles, 34, 38

F

feedback, as the home of learning, 178
feedback cycle, relatively convergent, 10, 106, 230, 248
feedback cycles, two (creature and environment), 75
Fermi paradox, 125, 216
firewall, hard, against general systems, 154
fitness function, 154
fixed points, 154
force, as counter-pressure, 160
fragmentation of agency, 204

G

gain of function research, 154
genetic algorithm, 54
genotype and phenotype, 54
gradient descent, stochastic, 110
gray goo, 120, 230

growth, by selection, 28, 106, 248

H

Habermas, Jürgen, lifeworld and system world, 140

halting problem, 174

hardware-layer learning, 75

hashiness, 178, 209

hazard narratives, 160, 243

high-dimensional optimization, 84

high-frequency trading, 140

hinge, 138, 224, 248

hive-mind, 54

human generation time, 132

hybrid, proposal of a, 2, 243

hybrid unions of humans and machines, sterile, 90

I

ice, accumulation on wings, 170

indirect actions, 193

influence, distinguished from control, 170, 248

inheritance outside the genetic sequence, 75, 120

instinct, three bases of, 70

instincts, where they come from, 216

instrumental convergence, 6, 10, 15, 140, 216, 243, 248

inter-ecosystem interface, 193

K

known nature, 44, 48, 224, 234, 248

L

large language models, 216

leakage of interests from a substrate, 116

learning, two layers of, 75, 132
learning and evolution, one kind of change at two levels, 54, 75
legal work, narrow system for, drift of, 154
libraries, composed by engineers, and evolution, 54, 125
libraries, software, 84
limits-to-control crux (Petillo), 209
linear games, oversight studied through, 178
linearity, engineered not emergent, 170, 209, 243
Living Planet Index, 140
Lua interpreter, as a picture of layering, 116

M

market collapse, uneven, 140
markets, as counter-pressure, 146, 160, 248
mesa-optimizer, 116
meta-goals, 216
meta-monitor, 178
metastability, 98
misuse, 2, 10, 15, 224, 243, 248
mixing function, 178
monism of cause, 187
monocrop, 120
multipolar trap, 243
mutation, 80, 84, 98

N

narrow artificial intelligence, 154
negotiation, thought experiment of, 2
neural interface, direct, 2
nodules, polymetallic, 160
noise floor, 187

O

on-ramp, 94, 224

optimization daemons, 187

optimization rings, lower and upper, 187

orthogonality thesis, 106

overdetermined failure of control, 178

override of evolution, question of, 243

P

paperclip maximizer, 6, 15

parasitism, resistance to, 94

past selves, veto by stored copies of, 110

pattern, replication of, 125

pattern weapons, 160

periodic table, 41, 234

perpetual benefit machine, 224, 248

perpetual motion machine, 224, 248

phenotype, 70

plankton, scope of the biomass claim, 140

plutonium, regulation of, as a model for limits on compute, 154

point mutation, 80, 98

point of no return, 28, 209, 216

pollution, 140, 243

power-seeking, 6

precautionary principle, 154

prediction, cost of, 174

prediction, three conditions of, 178

premises, the complete list, 10, 230, 234

Prisoner's Dilemma, the race as a, 216

proofreading, in living cells, 98

R

real-world tests, as evolution, 204

recombination, 66, 80, 84, 94, 98, 120, 248

red line, for developers, 154

regulation, 10, 150, 154, 243, 248

re-humanize, re-naturalize and re-wild, 216

replication, trivial and non-trivial, 66, 102

requisite variety, law of, 166

retrospective control, 102, 150, 174, 224

Rice's theorem, 204

robustness, any threshold crossed, 209

Russell, Stuart, 178, 216

S

search systems, federal, 174

seep-through, 209

selection, 80, 84, 120

selection, by people, of profitable variants, 110

sex, origin of, 90

sexual reproduction, 66, 94

silica life, 54

silicon, chosen for convenience, 48

silicon-based systems, 28, 34, 41, 44, 70

single units of intelligence, assumption of, 54

singleton, 54, 193

slowness, bias that hears evolution as, 125

software-layer learning, 75

spine, 98

substrate, defined, 48

substrate needs convergence, 10, 140, 204, 216

supply chains, collapse of, 140
supra-rational remainder, 90
survival selection, 54

T

takeoff, fast and moderate, 140
tardigrades, 34
tax of control, 178
thousand years, as the span of a safety claim, 178
three-body problem, 170, 243
three kinds of change to code, 54
three worlds problem, 140
timescales, conflation of, 132
translator, people as, 10, 24
treacherous turn, 166
turbulence, 170
Turing equivalence, 10
Two Generals problem, 98

U

universe, distribution of matter in, 38
unknown nature, 44, 224, 248
unseen, the chance that something important goes, 166

V

variants, tested in the world, 90, 204
venture capital, allocation of, 140
viral exponent, 90
virtualized code, 125

W

war games, 160

weights, of a neural network, 110
wing heater, 170, 248