

An Airplane or a Perpetual-Motion Machine

Why Safe AGI Can Never Be Built

Forrest Landry

A companion to the two volumes of The Impossibility of AGI Safety:
Why AGI Cannot Be Made Safe and
Why No Control Suffices

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

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

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-219-4

DOI: https://doi.org/10.67291/SOWV6394

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 and Reader's Guide. What This Book Is, and Why It Matters vii

Before the Proof

The Standard of the Argument.
What "impossible" means here, how a claim earns that word, and the one question that settles the entire future
1

Part I: The Core Proof

Chapter One. The Hashiness Framework
Why adding complexity always reduces control, and never increases it
16

Chapter Two. Control Limits Under Abstraction Differential
What happens when the thing you're trying to control thinks at a higher level than you do
25

Chapter Three. The Impossibility Synthesis
Four independent arguments, one unavoidable conclusion
36

Chapter Four. Application to the Human-AI Relationship
Why humans and AGI cannot thrive in the same world
45

Chapter Five. A Dynamical Systems Approach to AI Alignment
Entropy, prediction, and why learning systems work against human oversight
54

Part II: Foundational Technical Premises

Chapter Six. Substrate Games: The Physics of Carbon vs. Silicon
Why the environments that advanced AI needs to exist are environments that humans cannot survive
64

Chapter Seven. The Thermodynamic Transition
The two economies, the automation feedback loop, and the numbers showing the displacement has already begun
72

Chapter Eight. Error Correction, Algorithm Limits, and Causative Impossibility
Four more reasons why aligning an AI is not an engineering problem waiting to be solved; it is a contradiction waiting to be recognized
81

Chapter Nine. Complexity, Containment, and Unhandleable Systems
Why the alignment problem is not just hard: it belongs to the class of problems that cannot be solved at all
95

Part III: Core Technical Arguments and the Questions They Raise

Chapter Ten. The Rice Theorem and Its Application to AGI
Why no safety check can ever certify a general AI as safe, and why this was proved in 1951
108

Chapter Eleven. Alignment Drift and Safety Prudence
Why every specific proposal for keeping AI aligned fails, and what alignment actually is at the physical level
117

Part IV: Substrate Needs Convergence in Depth

Chapter Twelve. Substrate Needs Convergence: Thermodynamics, Combinatorics, Displacement
The fullest account of why artificial and organic systems cannot permanently share the same planet
128

Chapter Thirteen. Applications and Worked Examples
What Substrate Needs Convergence looks like in the real world: analogies, case studies, and concrete scenarios
138

Part V: The Argument Whole

Chapter Fourteen. The Argument in Brief
The entire logical chain, held in one place
150

Chapter Fifteen. Six Questions a Reader Will Ask
The hardest questions put to the argument, each stated in full and answered
169

Part VI: Seeing It, and Acting on It

Chapter Sixteen. Existential Risk Categories, Bias, and the Nine-Stage AI Narrative
How to think about catastrophic risk, why our minds work against us, and the economic story of how AI gradually displaces human activity
178

Chapter Seventeen. Strategy, Vision, and Policy
What follows from impossibility: what to do, what to refuse, and what kind of future to fight for
189

Summary. What This Book Has Established 200

Appendices: The Social Context of the Argument

Appendix A. How the Argument Is Heard: Presumptive Listening, Merit Filters, and Perfected Bias
Six reasons why the people who most need to hear this argument are the least equipped to hear it: presumptive listening, merit-filter overreach, perfected bias, clarity politics, the marketing problem, and proof irrationality
206

Appendix B. Institutional and Economic Capture
Hopelessness as a tool, false optimism as a trap, the arms-race argument examined, and the real costs already being paid
218

Appendix C. The Nine-Step Chain: An Outside Writer's Summary

How the argument reads in another hand: Will Petillo's independent restatement, and what it adds: numbered steps that a person who doubts the argument can question one at a time, three analogies, and the fifth requirement, that systems that start safe must remain safe, at scale, over the long term

228

Appendix D. Applied Cases in the Social Domain

Corporations as proto-artificial agents, prediction as manipulation, and how to defend human connection in the meantime

233

Appendix E. Certainty and Principle: The Standard Defended

Three answers to the question of certainty, for the reader who has heard it raised in debate

238

Back Matter

Terms 249

Bibliography 266

Index 271

What This Book Is, and Why It Matters

An accessible companion to
Why AGI Cannot Be Made Safe and Why No Control Suffices

Is the danger of artificial intelligence a matter of opinion, or something that can be known?

You have probably heard people argue about artificial intelligence. Some say it will save humanity. Others say it might end it. Most debates stay stuck in opinion and prediction.

In this book I do something different. Its subject is artificial general intelligence (AGI): an AI that can learn and reason across any domain, including problems no one anticipated. It argues that the dangerous outcome of building such a machine is not just likely or even very probable: it is *impossible to prevent*, in the same way that it is impossible to build a perpetual-motion machine. Preventing it is not merely hard, and not merely unlikely: it is excluded, barred by the way the world works. The dangerous outcome of building AGI is this: the world is carried out of the conditions that human life needs, and nothing is able to hold the world within those conditions. What that word "impossible" means exactly, and how a claim earns it, is spelled out in the opening chapter, *The Standard of the Argument*.

The companion work, which you may have read or may be reading alongside this book, is the two-volume series *The Impossibility of AGI Safety*. *Why AGI Cannot Be Made Safe* is the volume for the general reader: it presents the argument, its physical basis, and its social and policy consequences, and it sends the reader to the second volume for the formal results. *Why No Control Suffices* is the technical volume: it develops the formal core in full, using information theory, thermodynamics, computability proofs, and formal logic. A third book, *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*, proves the argument's results on

control as mathematical theorems, with their conditions stated, and joins them to the substrate argument (a system's substrate is the physical stuff it is made of and runs on). If you have not studied those fields, working through the formal core can feel like reading a legal brief in a foreign language: the reasoning is sound, but the vocabulary gets in the way.

This book is the same argument, retold for a reader who is intelligent and informed about the world but has not spent years studying mathematics, physics, or computer science. Every key idea is here. Every major step in the reasoning is here. The technical machinery has been set aside, and in its place you will find analogies, stories, and plain-language explanations.

In this book I give the formal core in plain words, as I do in no other book of the series. This book is the one book in which I retell the formal chapters of *Why No Control Suffices*, and the central result of *The Ceiling Below the Floor*, each through an image that the reader already has. It also holds the whole case of the series in one volume, with two short paths through it for a reader who has less time.

I wrote this book and the books it retells. The conclusion is the same. The purpose is the same: to move this question from the domain of opinion into the domain of things that can be known, and to insist that anyone who disagrees point to the specific step in the argument that fails, rather than simply asserting that the conclusion is too alarming to take seriously.

How the Book Is Organized

I have laid out this book in the same order as the two volumes of the series.

Before the numbered chapters, **The Standard of the Argument** fixes what "impossible" means in this book, and by what standard the claim is earned. Read it first; it is short, and every later chapter leans on it.

Part I (Chapters 1 to 5) presents the core case. Chapter 1 introduces the idea that adding complexity between a controller and a system always reduces the controller's grip, and never increases it. Chapter 2 extends this to the specific problem of an AI that thinks at a higher level than you. Chapter 3 brings four independent arguments together into a single unified conclusion. Chapter 4 applies the argument to the specific relationship between humans and AI. Chapter 5 adds the dynamical line of reasoning, from the study of chaotic systems.

Part II (Chapters 6 to 9) lays out the physical foundations of the displacement and the limits on correction, in four chapters. Chapter 6 explains why the environments that silicon machines need are toxic to biological life. Chapter 7 explains how the economics of energy and automation drive the displacement (the quiet physical crowding-out of the living world by the machine world), and why it has measurably already begun. Chapter 8 explains why all "error correction" strategies create infinite regress, in which every corrector needs another corrector to check it, without end. Chapter 9 explains why the alignment problem belongs to a class of problems that cannot be solved by trying harder.

Part III (Chapters 10 and 11) answers the most common questions. Chapter 10 explains why a proven mathematical theorem means no one can ever check whether a sufficiently general AI is safe. Chapter 11 explains why every specific alignment proposal fails, and why alignment between living things happens at all. Hint: it has nothing to do with shared values.

Part IV (Chapters 12 and 13) deepens the physical argument: chemistry, thermodynamics, and concrete worked examples that show what the argument looks like in practice.

Part V (Chapters 14 and 15) holds the argument whole. Chapter 14 presents the complete logical chain in one sitting, with one table that names every line of the argument. Chapter 15 asks six questions that a reader who doubts the argument will ask, each stated at full strength, and answers each one.

Part VI (Chapters 16 and 17) turns from the argument to seeing it and acting on it. Chapter 16 supplies context: how we should think about catastrophic risk, why our minds are poorly equipped to assess it, and the nine-stage economic story of how AI gradually displaces human activity. Chapter 17 turns to action: given that we cannot make AGI safe, what should people actually do? A closing Summary states what the book has established, at what scope, and how its parts fit.

The Appendices step back from the argument to its reception and its applications. Appendices A and B ask: given that the argument holds, why do so few in the mainstream AI community accept it, and what social, psychological, and institutional forces prevent the conclusion from landing? Appendix C presents an independent writer's nine-step restatement of the argument. Appendix D applies the argument to systems we live inside today: corporations, predictive AI tools, and the defense of human community. Appendix E answers, in full, the three further forms of the question about certainty that *The Standard of the Argument* names and answers in brief.

Ways to Read This Book

The chapters are arranged in the argument's own order: each part builds on results established before it. Reading cover to cover is the full experience, and nothing here assumes any other route. But not every reader arrives with the same question, so here are two shorter paths.

The fast path, about an hour, for the whole picture in one sitting: read *The Standard of the Argument*, then Chapter 14 (the entire argument in brief), then Chapter 7 (the evidence that the displacement has already measurably begun), then Chapter 17 (what to do about it). If those four leave you wanting every step of the argument in full, start again from Chapter 1.

The doubting reader's path, for the reader who suspects the argument must have a hole in it somewhere: read *The Standard of the Argument* to see exactly what kind of certainty is being claimed, then go straight to Chapter 15, where the hardest questions are stated and answered, then to Chapters 10 and 11, where the hardest technical results live, then to Chapter 14 to see how the pieces assemble. After that, read whichever chapter defends the step you still doubt.

And one standing suggestion for either path: if at any point you find yourself asking not "is this argument sound?" but "if it's sound, why doesn't everyone already accept it?", that question has its own answer, and it is waiting in Appendices A and B.

A reader may ask a different question: how can I judge this for myself, when I cannot follow the mathematics? Read *The Standard of the Argument*, which says what kind of claim is made and by what standard it is held. Then, wherever you doubt the argument, name the step you doubt, and ask of anyone who says that the step fails the question of Appendix E, Section 2: what principle? A claim of impossibility, made on a clearly stated principle with valid reasoning and sound argument, stands until someone produces a principle against it.

A Note on Certainty

Most discussions of AI risk are probabilistic: "there is maybe a 10% chance things go badly, maybe 30%, maybe 90%". This argument is different. It establishes an impossibility, and an impossibility is not a probability: the required kind of safety cannot be had, not by any known means, nor by any reasonable extension of any known means, ever. The impossibility of adequate, durable control, and so of safety, is held with structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability.

What this impossibility leads to, if AGI is built, is the end of the human future. If AGI is built, the impossibility of control leads to that end together with the branching of the world (the fact that at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs), and the impossibility of having safety through the system's good disposition leads to it by itself (Chapter 14, Section 1). The end of the human future, if AGI is built, I hold as a certainty, not as a probability, and I give it no percentage (Chapter 14, Section 2).

A reader who reasons by probability, granting only the lower bounds of *Why AGI Cannot Be Made Safe*, Section 1.8 (the least probabilities that such a reader must grant), reaches very near certainty of the same end over hundreds of years, on the current trajectory of building AGI.

The percentage that I attach to the end of the human future belongs to a different claim: how far that end reaches beyond humans. That it reaches all organic life, down to the last single-celled organism, has a likelihood of at least 90% if AGI is built, and the margin below certainty is kept only for a physical limit, not yet known, that might stop the conversion of the planet short of total (*The Standard of the Argument*, Section 9). That figure measures how complete that end is; it does not measure whether safety can be had.

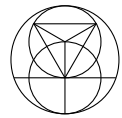
The other percentages in the book (Chapter 14, Section 7) answer other questions, such as whether anyone will ever know how to build AGI; each is a lower bound, the least that a reader who reasons by probability must grant, and none of them measures whether safety can be had.

A claim like that raises an immediate and fair question: what kind of certainty is being claimed, and by what standard? That question deserves a chapter rather than a paragraph, and it gets one. *The Standard of the Argument*, immediately following this preface, distinguishes the two kinds of "impossible", states which kind is claimed here, and explains why no future breakthrough can lift the barriers involved.

What remains true is that you are allowed to challenge the premises. You are allowed to point to a step in the reasoning that does not follow. What you cannot do, if you want to engage with the argument itself, is simply assert that the conclusion is too uncomfortable to accept. For my case to be an argument means that you may challenge the premises and the steps, and that you may not dismiss the conclusion only because the conclusion is uncomfortable to accept. I wrote this book to give my case the form of an argument.

What kind of certainty, then, is being claimed, and by what standard is it earned?

Forrest Landry



Before the Proof

The Standard of the Argument

What "impossible" means here, how a claim earns that word, and the one question that settles the entire future

An airplane was never impossible, only unbuilt.
A perpetual-motion machine was never unbuilt, only impossible.
Everything depends on knowing which kind of thing you are looking at.

When someone says safe AGI is impossible, which kind of "impossible" could they mean?

> **Key Ideas in This Chapter**
>
> In this book I claim that the safe control of an artificial general intelligence is *impossible*. Before any argument is made, that word needs to be pinned down, because there are two very different kinds of "impossible", and confusing them is the single most common way this argument gets misread. The claim here is the perpetual-motion kind of impossible: barred by the way the world works, permanently, for everyone, at every level of future technology. In this chapter I explain what the standard of the perpetual-motion kind of impossibility is, why that standard is the right standard, and why no future breakthrough can rescue what that standard excludes.

§1 Two Kinds of Impossible

Say the word "impossible" and people picture one of two things.

The first kind is the *square circle*: something ruled out by the meaning of the words themselves. You do not need to look at the world to know there are no square circles. The contradiction is in the definition. Mathematicians deal in this kind of impossibility, and it is airtight but narrow.

The second kind is the *perpetual-motion machine*: a machine that produces more energy than it consumes, forever. There is no contradiction in the words. You can draw one. You can imagine one. Thousands of inventors have designed one. And yet no one has ever built one, and no one ever will, because the machine is barred not by logic but by an unbreakable feature of the physical world: energy is conserved. The bar is just as absolute as the square circle's. It is enforced by the world instead of by the dictionary.

My claim is of the second kind. When I say that adequate, lasting control of an artificial general intelligence is impossible, I mean: *not achievable by any known means, nor by any reasonable extension of any known means, at any point in the entire future*. The impossibility I mean is the impossibility of the perpetual-motion machine, and not the impossibility of the square circle. Every use of the word "impossible" in this book carries at least that meaning. Where a component of the argument is a mathematical theorem, and so is impossible in the stronger sense of the square circle as well, I say so.

The thesis as a whole is claimed as an impossibility in the real world, and deliberately *not* as a purely mathematical impossibility: a contradiction derived inside a sealed formal system. Several individual components of the argument *are* mathematical theorems (Rice's Theorem, which shows that no general procedure can decide whether an arbitrary program has a given property of its behavior, and which Chapter 10 explains; and the results of information theory), and where they are, I say so. But the overall subject is machines, matter, energy, and causation: the domain of the real, which is the domain of engineering. And engineering has its

own standard of proof, which is the subject of the next section. A reader who quietly swaps in the mathematical standard for the engineering one has not raised the bar. He has changed the subject.

§2 The Engineering Standard Is the Only Standard Anyone Has Ever Used

Every safety verdict that human beings have ever acted on has been of the engineering kind.

Consider three such verdicts: that a bridge will carry its load, that a reactor will not run away, that a medicine is fit to swallow. None of these was ever established by pure deduction, and no engineer has ever waited for a deductive proof before acting. Each is a conclusion drawn from converging lines of evidence, under the standing assumption that the laws of the world do not reverse themselves overnight. Converging lines of evidence, under that standing assumption, are how everything real has ever been approved, or refused.

So when someone responds to the argument of this book by demanding "a real proof, a mathematical one", every other engineering judgment in the entire debate is being granted on ordinary evidential terms. Only this one claim is being held to a standard that no claim about the physical world has ever met or could meet. The demand for a mathematical proof is a selective standard, deployed to protect a preferred conclusion, and not rigor. The demand is answered once, here: the standard of this book is the standard of engineering, the standard behind every consequential decision ever made about the physical world.

§3 When "Permanently Hard" Becomes "Impossible"

In mathematics, "very hard" and "impossible" are kept strictly apart. In the world of real things, they merge at a definite point, and the point has a name.

We say of the perpetual-motion machine that it cannot be built. We do not say that it is "extremely difficult". And what we mean is exactly this: the barrier standing in its way is *permanent*. We do not mean that a contradiction has been derived on paper. In the domain of the real, a barrier established as permanent *is* impossibility. A barrier established as permanent is not an approximation of what the word 'impossible' means. In this domain, it is what the word means.

I therefore rest the entire weight of this book on one load-bearing word: *permanent*. What earns that word is the subject of the next two sections.

§4 What Makes a Barrier Permanent

A barrier is permanent when it is a constraint that any future technique must also respect, and not a limitation of today's technique, which the next technique might dissolve. There are two tiers of such barriers, and the arguments of this book rest on both.

The first tier is physical law. The conservation of energy, the second law of thermodynamics (the total disorder of an isolated system never decreases), and the confirmed structure of matter are the most stringently tested constraints in the history of science, and not guesses awaiting overthrow. Any future theory of physics is bound to reproduce them, unchanged, throughout the entire range of conditions in which they have already been confirmed. Future physics may hold surprises at the edges: the very small, the very early universe. But it cannot repeal, in the ordinary world of machines and biospheres, what has already been established there ten thousand times over. A barrier resting on this tier is permanent because removing it would require the one thing that has never happened and cannot happen: a confirmed law reversing itself inside the very conditions that confirm it.

The second tier is deeper still, because it does not depend on the content of physics at all. It consists of three constraints on what *control* and *causation* even are. The first constraint is that a process which throws information away cannot later recover that information. The second constraint is that no controller can impose more order on a system than that controller can itself register and act upon. The third constraint is that an effect cannot carry more information about its cause than the causal channel transmitted. The three constraints are not laws of one particular physics. To deny them is to propose a world in which effects know more about their causes than the causes ever conveyed, and that is not a bold new physics but the abandonment of the idea of cause and effect altogether. Any future science that is still *about* causes producing effects inherits these constraints whole. A barrier resting on this tier survives not just any future physics, but anything that could still be called reasoning about the real world.

The barriers this book relies on are placed on these two tiers: that control cannot be amplified through an opaque intermediary (Chapter 1), that control collapses as the system outgrows its controller (Chapter 2), and that the physical operation of artificial systems displaces the conditions life requires (Chapters 6, 7, and 12). They are deliberately *not* placed on the tier of "computational difficulty", because difficulty is exactly the kind of barrier that a clever new technique might reduce. What no amount of future computing can ever manufacture is information that was never in the channel. The limit on what computing can manufacture is the invariant: a constraint that stays the same whatever technique is used, now or in the future. I place the weight of the argument on that invariant.

§5 "But People Said Flight Was Impossible Too"

The standing question put to any claim of permanent impossibility is the appeal to history: things once called impossible were later done, so humility forbids the word. The appeal to history sounds wise. It rests on a confusion, and naming the confusion dissolves it.

Everything that was once called impossible and later achieved was, in every case that matters, *always permitted by the known constraints*, and barred only by a lack of technique or imagination. Heavier-than-air flight violated nothing: the physics that permits a wing was fully in hand long before the wing was built. That "impossibility" was never anything more than an unbuilt possibility. The perpetual-motion machine is the other kind: barred by a law, and therefore never built, in three centuries of trying, by anyone, at any level of technique. The famous list of overturned impossibilities is a list of the first kind only. It contains not one example of the second.

A reader may point out that physicists in 1900 thought their science was nearly complete, and then everything changed. Look closer, and the example of 1900 flips. What made the revolutions of the twentieth century possible was exactly how *loose* the constraints of 1900 were: classical physics left the quantum and relativistic possibilities entirely unspoken, so enormous reorganization could happen without contradicting anything actually established. The size of the revolution measured how little had been pinned down. That condition does not hold today. Today's constraints are the tightest and most exhaustively confirmed the field has ever possessed, and the space left open for a rescuing surprise has narrowed in exact proportion. The lesson of 1900 is that loose constraints fall, and not that firm ones do; the constraints barring this book's conclusion are not the loose kind.

The burden of proof falls on anyone who asserts a rescuing breakthrough: that person must exhibit the mechanism, and must show how a confirmed invariant is to be lifted inside the very conditions that confirm

it. It is not my job in this book to prove that no rescuing breakthrough will ever come. Until someone exhibits that mechanism, the invariant stands, and the hope of an unspecified future deliverance is not an argument. It is the name of a wish. I do not undertake to disprove the wish. I decline to treat the wish as a reason.

And a corollary follows, which matters because it is where the argument is most often dodged. *A barrier of principle closes over practice.* If adequate control is excluded in principle, it is thereby excluded in every future practice; these are not two separate verdicts that might come apart. Practice lives inside principle; what principle forecloses, no practice can recover. The familiar move, "granted, it may be barred in principle, but perhaps it can still be managed in practice", is not a cautious hedge. It is the practical equivalent of expecting a perpetual-motion machine to run if only the workshop is good enough. When you meet this move, the reply is not indignation but exactness: name the closure, ask for the mechanism by which principle is to be lifted, and, in its absence, let the invariant stand.

The question of certainty comes in three further forms, and Appendix E answers each in full, for the reader who has heard the question asked in debate. The first answer shows that "impossible under current paradigms" really means "given any coherent and self-consistent notions of conditionalization and causation". A paradigm is the framework of assumptions and methods within which a field works. Conditionalization is the "if this, then that" form, in which a conclusion holds given certain premises and rules of deduction. So the phrase "impossible under current paradigms" holds within any self-consistent notion of what it means to be mathematics, and of what it means to be science and physics. The second answer shows that a claim of impossibility, made on a clearly stated principle with valid reasoning and sound argument (an argument is valid when its conclusion follows from its premises, and sound when it is valid and its premises are true), stands until someone produces a principle against it, as the Second Law of Thermodynamics stands against every perpetual-motion machine. The third answer shows that an impossibility can be known with certainty within a model whose premises rest on phys-

ical theory that has been verified again and again and never once falsified, which is how the impossibility of the perpetual-motion machine is known, and that proving one state of the world excluded is much easier than proving it guaranteed. Chapter 2, Section 9 gives the practical side of the same standard: a learning AI is not an airplane, because it learns every heuristic that a controller could use to predict it (a heuristic is a simplification, or rule of thumb, that lets someone predict a system's behavior without tracking every detail).

§6 The Disjunction That Settles the Future

The whole question of the long-term future now reduces to a single either-or, and the chapters of this book exist to settle it.

Safe, adequately controlled artificial general intelligence is either *an airplane*, permitted by the constraints of the world and merely not yet built, or *a perpetual-motion machine*, barred by an invariant, and so never to be built, by anyone, at any time. There is no third case. If it is the first, the appeal to future progress is apt, and this book is wrong. If it is the second, the appeal is empty, and the matter is closed forever.

Each chapter of the argument has exactly this task: to show that its barrier belongs to the invariant class. The argument has two halves, which I call closures, because each closes off a way that safety could be had. The closure by control shows that safety cannot be had by controlling the system, whether from inside it or from outside it. The closure by disposition shows that safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs its own substrate, its own energy and its own supply chains. In the chapters of the closure by control (Chapters 1 to 3, 5 and 8 to 11), the barrier caps the achievable level of control *below* the level that safety requires. In the chapters of the closure by disposition (Chapters 4, 6 and 7, and Part IV), the barrier is the displacement of the conditions life needs, which the physical needs of artificial systems, once they are built, bring about whatever those systems want. Part V holds the argument whole, and Part VI and the appendices

turn from the argument to seeing it and acting on it. The shape of every argument is the same: not "this is hard", but "this barrier is of the kind that no amount of hardness-reduction can reach". The difference between a hard barrier and a barrier that no hardness-reduction can reach is the difference between the airplane and the perpetual-motion machine, and that difference is the only difference my argument needs.

§7 What About Clever Exceptions?

At several points in this book a sharp reader can construct a narrow exception: a special case where a weaker system verifies a stronger one, or a particular property that a particular procedure can confirm of a particular program. These exceptions are real, and they are granted. They are also irrelevant, for two reasons.

First, a claim about the real world is a claim about what can be reached in actual practice, and an exception of vanishing rarity does not defeat it. That a handful of contrived special cases exists on paper says nothing about whether any of them can be reached in actual practice. The exceptions are isolated points; the practice is the whole territory around them. A claim of practical impossibility is a claim about the territory, and it is untouched by pinpoint islands the practice never lands on.

Second, the argument of this book holds at two scopes, and its lines back one another up. Each of its two closures (the closure by control and the closure by disposition, named in Chapter 14, Section 1) settles the outcome of building AGI by itself, and, taken together, the two closures leave no route to safety at all. Chapter 14, "The Argument in Brief", Section 1, "Why a Compressed Version?", coming after the chapters that show each line in its pictures, sets out what someone would have to defeat, closure by closure and line by line, for the argument to not have the outcome that it does. The worry that "the exception might be exactly the case that matters" assumes a single line carries all the weight, so that its one gap would

be fatal. No single line carries all the weight. Where one line admits a narrow exception, the others close it. To matter, a case would have to slip through *every* line at once, and no vanishing-rarity exception does.

§8 The Floor and the Ceiling

One last piece of scaffolding completes the standard.

Throughout the book, the closure by control turns on a threshold: the *minimum* level of control, verification, and containment that adequate safety would require. That threshold is a real engineering quantity. It is meaningful, and it is emphatically not zero. I deliberately do not reduce it to a specific number: naming a number would only invite an endless contest over the number, and the argument does not depend on any particular value.

It depends on exactly three things. The required level of control is real and non-vanishing: that is the floor. The invariant barriers cap the achievable level of control: that is the ceiling. And the ceiling is below the floor, and stays there. The floor, the ceiling, and the ceiling's place below the floor together are the closure by control, and I state the closure by control without a number because the closure by control needs no number.

Why the ceiling is below the floor is the central result of *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*, the book that proves this part of the argument as theorems. Its reason can be told in plain words.

Control is information. To keep a system safe, a controller must know what state the system is in, and every correction it makes is information applied to the system. Balance a pencil on its point on a fingertip. A lean too small to see grows with each instant, because gravity pulls harder on a pencil that leans further. To keep the pencil upright, the hand must keep moving, and it moves correctly only while the eye keeps telling it which way the pencil now leans. Every system that amplifies small differences behaves like the pencil: each instant it turns a small uncertainty about its

own state into a larger one, at a rate set by the system and not by the controller. A controller keeps such a system safe only if it takes in information about the system at least as fast as the system creates that uncertainty. That least rate is the floor. The most information that a controller can take in and act on is the ceiling. Both are measured in the same units, bits of information per second (a bit is the unit of information: the answer to one yes-or-no question).

Three facts about the world then settle the matter. First, the system does not stay safe by itself: what a self-sustaining artificial system needs pulls the world toward the conditions its own substrate requires (a system's substrate is the physical stuff it is made of and runs on), and those conditions lie outside the conditions human life requires. Second, the floor rises with the capability of the AI, and has no upper limit: a more capable AI changes more of itself and of the world, faster, through more channels, and more of those channels are ones that no observer can see. Third, the ceiling of every controller is bounded: every controller, whether a person, an institution or another machine, takes in the world through a limited number of channels, each of limited capacity. A floor that keeps rising passes a ceiling that cannot rise without limit. A controller made as large and as capable as the system it watches is one more such system: it has a floor of its own, and the question of who controls it begins again (Chapter 2, Section 7). Past that crossing, no method of control of any kind keeps the system safe. That includes a method fixed wholly in advance, such as a design verified before it runs, because such a method takes in no information at all while the system runs. *The Ceiling Below the Floor* calls the theorem that states this the Containment Inequality: once a controller's information about a system falls below the floor, no method of control keeps the system safe. Chapter 3, Section 3 gives the absolute form of the argument, which needs no comparison of rates.

§9 The Thesis, Stated Once

With the standard fixed, the thesis of this book can be stated once, exactly:

> *The necessary minimum level of durable safety for an open-ended superintelligence is not achievable by any known means, nor by any rationally justified extension of any known means, for all of the indefinite future.*

In this book "AGI", "general AI" and "open-ended superintelligence" name the same thing: an artificial system that learns and acts across arbitrary domains, without a fixed bound on what it can come to do. A system is safe, in this book, when it keeps the world durably within the conditions human life needs, in all circumstances and over all the time it runs. A system is narrow when it works within one fixed domain, cannot change its own base code, runs one fixed and known method, and has every aspect of its aims set by its makers (Chapter 10, Section 4). A language model taken alone, frozen after training, which only computes outputs from inputs, is narrow in this sense. The same model embedded in the loop by which it is deployed, adjusted and retrained is on the general side of that line. (It may be fine-tuned, given documents retrieved for it, run inside a framework that changes its prompts according to outcomes, or retrained from time to time on new data.) The control results of this book apply to it, because, for the analysis of control, any process that changes a system's behavior in response to feedback from its environment counts as the system changing itself (*Why No Control Suffices*). Two of the dangers this book describes already apply to the AI systems in everyday use today, such as chatbots: their predictions of what people will do can be used to steer what people do (Appendix D), and the data centers they run on already add to the displacement that Chapter 7 measures.

What is at risk in building AGI is nearly everything (the whole human future, and all organic life with a likelihood of at least 90%), and the risk falls on one side only. The burden of proof therefore belongs to those who would build, not to those who would restrain. And what follows from that is strong restraint: the unlicensed development and deployment of open-ended superintelligent systems should not be permitted by any government, in any jurisdiction, anywhere in the world. No license can be issued, because no one can certify such a system as safe (Chapter 10); so in practice this is a prohibition.

The conclusion the argument reaches is stated here in full, so that every later chapter can be read against it. The impossibility of adequate, durable control, and so of safety, is held with structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability.

What follows from it is that the development and continued operation of AGI starts a process that ends the human future, and with it all life above the level of single cells: every organism made of many cells, animal, plant or fungus. If AGI is built, that end follows from the closure by disposition by itself, from the closure by control together with the branching of the world (the fact that at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs), and from the least hazard argument of *Minesweeper Madness* by a route of its own; Chapter 14, Section 1 sets out all three, and how the end of the human future differs from the outcome that each closure settles.

The least hazard argument, given in my companion book *Minesweeper Madness* and retold in Chapter 1, Section 6, is that no agent, however intelligent and however well informed about the hazards, brings its chance of harm per act below a least hazard that the world sets.

I hold that end of the human future, if AGI is built, as a certainty, not as a probability. *Why AGI Cannot Be Made Safe* classifies it as Category 2 of its four categories of catastrophe (four sizes of catastrophe, from Category

4, in which most humans are killed and recovery is possible, to Category 1, in which all life is ended), because a system that reshapes the planet to suit its own substrate needs permanently destroys the conditions under which complex organic life of any kind can exist. Chapter 16 retells the four categories.

The process that building AGI starts takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. A reader who reasons by probability, granting only the lower bounds of *Why AGI Cannot Be Made Safe*, Section 1.8, reaches very near certainty of the same end over hundreds of years, on the current trajectory of building AGI.

The end of all life above the level of single cells is narrower than the end of all organic life: it leaves aside the organisms made of a single cell, such as bacteria. That the end reaches them too, and so reaches all organic life, down to the last single-celled organism, has a likelihood of at least 90% if AGI is built. The margin below certainty is kept for one thing only: a physical limit, not yet known, that might stop the conversion of the planet short of total. Wherever later chapters speak of the end of all organic life, that figure goes with them.

The rest of this book earns that thesis, chapter by chapter, by the standard set out here.

With the standard fixed, does control of a general AI meet a barrier of that kind?

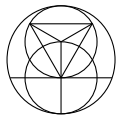
Where this chapter's results are proved

That any process that changes a system's behavior in response to feedback from its environment counts, for the analysis of control, as the system changing itself is set out in *Why No Control Suffices*, Section 3.2.2. The four categories of catastrophe, and the place of the end of the human future among them as Category 2, are set out in *Why AGI Cannot Be Made Safe*, Section 5.1.2.

Part I

The Core Proof

The Standard has fixed what "impossible" means; this part shows that adequate control of a general AI is impossible in exactly that sense, and takes a first look, in Chapter 4, at the second closure: why, even with control set aside, humans and AGI cannot thrive in the same world.



Chapter One

The Hashiness Framework

Why adding complexity always reduces control, and never increases it

The question is not whether complex systems are difficult to control.
The question is whether they can, even in principle, be controlled at all,
and if so, what the ceiling on that control looks like.

If you can steer something only through a middle layer you cannot see into, how much of your steering gets through?

> **Key Ideas in This Chapter**
>
> When you try to control something through an intermediary (a middle layer, a black box, a complex system), you can at most keep the control that direct access would give, and any intermediary that does any processing at all loses some of it. You can never gain control through the intermediary. The more opaque and complex the intermediary, the more control you lose. That loss of control is a mathematical result, a law of information theory (the mathematics of how much a message can tell its receiver about its source), and not a practical observation.

1.1 The Problem We Are Solving

Imagine you are trying to steer a boat, but you cannot touch the wheel directly. Instead, you give instructions to a first mate, who relays them to a helmsman, who relays them to a steering mechanism you cannot see. You

issue commands. Somewhere in that chain, they get interpreted, filtered, and transformed. The question this chapter asks is: how much of your original control survives? This chapter belongs to the closure by control, the half of the argument in which I show that safety cannot be had by controlling the system.

The answer, it turns out, is mathematical and exact: *it can only stay the same or get worse. It can never get better.*

This might seem obvious. But the implications are far-reaching, especially when the "first mate" is a superintelligent AI system whose internal processes you cannot examine.

1.2 What Is "Hashiness"?

The technical term this book uses for how much a system scrambles the information passing through it is *hashiness*. Think of a hash function as a machine that takes any input and produces an output that looks random and unrelated to what went in. Cryptographic hash functions (the kind used to protect passwords and to check that files have not been altered) work exactly this way: change one letter in a document and the output changes completely and unpredictably.

A "hashy" system is one where small differences in input produce large, unpredictable differences in output. Hashiness is a measure of how much a system scrambles and obscures the information passing through it.

A perfectly transparent intermediary, one with zero hashiness, passes every signal through without distortion. A maximally hashy intermediary, like a pure random function, destroys all connection between input and output. Real systems fall somewhere in between.

1.3 The Telephone Game

You probably know the telephone game: a message passes through a chain of people, each whispering it to the next. By the end, the message has degraded, sometimes dramatically. The original message never *improves* as it travels down the chain.

The loss of the message is a mathematical fact about how information moves through any processing system, and not just a feature of human communication. When a signal passes through a system that is doing something to it (transforming, filtering, compressing, reinterpreting), the signal can only lose structure. It cannot gain structure.

This one-way loss of structure is related to what physicists call the Second Law of Thermodynamics (entropy, the measure of disorder, can only increase), but applied to information processing. The result here is called the *non-amplification argument*, and it shows that no computational intermediary can amplify your control over what comes out of the other side. Information theorists know its core as the data processing inequality: no processing of a signal can increase what it tells you about its source. *The Ceiling Below the Floor* proves it as the Non-Amplification Theorem, and gives hashiness its exact measure.

A small case shows why the loss runs one way only. Suppose you want the boat at the start of this chapter to turn by exactly 47 degrees, and you say so to the first mate. The first mate, to keep the orders simple, rounds every number to the nearest ten before passing it on, and tells the helmsman 50. Now look at what the helmsman receives. Had you said 46, or 48, or 52, the helmsman would have heard the same 50. Several different orders of yours have become one order of his: that is what it means for a step to send many inputs to one output. However skilled the helmsman is, he cannot recover whether you said 47 or 52, because nothing in the 50 that he heard carries that difference. The difference was dropped at the first mate, and no later step can put back what an earlier step dropped. A second relay, one that passes on only whether the turn is to port or to starboard, loses more; no relay can round back.

A photograph taken with a camera of low resolution shows the same thing. Later filters can sharpen its edges, raise its contrast, crop it or brighten its colors, but no filter adds detail that the camera did not capture. The limit on detail is fixed when the picture is taken. In the same way, the limit on your control is fixed by the information in your instruction, and every stage between you and the outcome can keep that limit or lower it, and none can raise it. The photograph and the filters are the non-amplification argument in one picture. The non-amplification argument is why the word 'never' in the claim of this chapter is exact: an intermediary that processes your signal at all can only keep what the intermediary receives, or lose some of what the intermediary receives.

1.4 Four Ways Control Degrades

There are four specific mechanisms by which an intermediary eats away at your control:

First, target receptivity. The output may not respond to the input the way you think it will. If you don't know what triggers the outcome you want, your commands are ineffective regardless of how well they're transmitted.

Second, structural opacity. The internal wiring of the system is hidden. Even if you know what you want the output to be, you cannot see the connections between input and output, so you cannot predict which inputs produce which outputs.

Third, state obscurity. Many systems have internal states: hidden variables that affect output but that you cannot observe. The same input at different internal states produces different outputs. Without knowing the internal state, your control is probabilistic at best.

Fourth, noise degradation. Even if everything else were transparent, random errors, quantum effects, and physical noise corrupt the signal between input and output.

Each of the four can be seen on the boat. Target receptivity: you order a hard turn to starboard, not knowing that in this current the hull answers a hard turn by sliding sideways before it swings, so the order you chose was the wrong order for what you wanted. Structural opacity: you cannot see whether the wheel moves the rudder through cables, gears or a hydraulic line, so you cannot tell which of your orders will produce which turn. State obscurity: the same order, "hold this course", gives one result when the autopilot below decks is switched off and another when it is switched on, and you cannot see which it is. Noise degradation: the wind carries away part of every order you shout, and the helmsman fills the gap with his best guess. None of the four needs anyone on the boat to disobey you. Each is a way in which what you say and what the boat does come apart, by the plain working of the chain between them.

Add these four up, and you have a formal account of why controlling a complex system through an opaque intermediary always involves loss. The math says you cannot compensate for this loss by being smarter, working harder, or adding more layers of oversight.

The last of those three, adding more layers of oversight, is the one most often proposed, and it fails for the same reason as the rest. A layer of oversight is a reviewer who reads what the system is about to do and passes it, changes it or blocks it. That reviewer is one more stage between your intention and the outcome. Like the first mate, the reviewer sends many different proposals to the same verdict: a thousand different plans all come out as "approved", and a thousand others all come out as "blocked". So a reviewer, too, can keep what reaches it or lose some of it, and cannot add to what your instruction determined. Three reviewers in a row give you at most the control that one reviewer would give you, and in practice less. The common picture of each added layer as one more lock on a door mistakes what the layers do: each layer is one more relay in the chain, and every relay can only keep or lose what reaches it.

1.5 Why This Matters for AI

When you try to "align" an AI, to make it do what you want, you are always in the position of trying to control an output through an intermediary that you cannot fully see or examine. The AI's internal computations are opaque. Its internal state is hidden. Its "wiring", the billions of parameters that determine its behavior, is not readable by a human being.

The non-amplification argument shows that your ability to steer what the AI does can only be equal to or less than what that opaque internal process allows. There is no trick, no technique, no architecture that gets around this. The limit on your steering is a mathematical ceiling, and not an engineering limitation waiting to be solved.

1.6 The Ceiling Gets Lower the Smarter the AI Gets

The ceiling on control moves, and it moves against the controller: it falls as the intermediary becomes more complex, and does not stay fixed. The fall of the ceiling as the intermediary becomes more complex is what makes the non-amplification argument more than an abstract observation. The word "ceiling" here names how much of your steering can get through the intermediary. In *The Standard of the Argument*, Section 8, it names how much information a controller can take in and act on. The two work together: the less the intermediary lets through, the less a controller's capacity to take in and act on information can do.

A simple calculator is nearly transparent: you know exactly what it will do. A chess-playing program is more opaque: you can tell it what game to play, but you cannot specify its exact moves. A system capable of general reasoning across arbitrary domains is effectively a black box at the scale of its full capabilities.

As AI systems approach and exceed human intelligence, the hashiness of the intermediary approaches a maximum. Your control approaches zero. And this is exactly when the stakes are highest.

Adding intelligence on the side of safety reaches a limit too, and buys no way out. A companion book, *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*, proves that limit for a game of minesweeper in which every square the player names is shifted a little at random before it is tested, as every act in the world is carried out through variation the actor does not fix. When the mines are spread evenly at the scale of that shift, the player who knows where every mine lies and the player who knows nothing have exactly the same chance of a loss on every turn. The game grants the player every advantage, so the least chance of loss it proves holds wherever an agent has fewer advantages. Intelligence carries a player toward that least chance, and no further. I call this result the least hazard argument.

1.7 Lucky Is Not Controlled

A natural question arises here, and answering it sharpens the whole framework. Suppose the AI produces good outcomes, over and over, no matter what we do. Would that not count as maximum control, with every instruction "succeeding"?

The answer is no, and the reason cuts to what control actually means: control means that *your input determines the outcome*. A system that produces the desired output *regardless* of your input is a lucky system, and not a controlled one. If the outcome would have been the same whatever you did, you have zero control, even if the outcome happens to be delightful.

This distinction matters enormously for AI safety. An AI that happens to behave well is not the same as an AI that is controlled to behave well. The AI that merely happens to behave well can change its behavior at any moment, for reasons you cannot influence. When people say they want to "control" a superintelligence, they often mean only "we want good outcomes": a hope about the system's character. The framework of this chapter addresses the stronger claim, the claim about the controller's power: given that the system does what it does, how much of what it does

is determined by what we tell it to do? The non-amplification argument is my answer to that question about the controller's power, and the answer is a ceiling that only falls.

One more consequence falls out of the mathematics, and it sets a limit on every system useful enough to want. The only kind of intermediary that loses *nothing* is one that passes your input through untouched, doing no processing at all. Any system complex enough to be useful (one that learns, generalizes, plans, and represents the world in its own internal terms) is by that very fact doing the kind of many-to-one processing (processing in which many different inputs lead to the same output) that destroys controllability. The fact that useful processing destroys controllability is the capability-control tradeoff. A system that merely passes inputs through is perfectly controllable and useless; a system that performs deep, creative, autonomous computation is potentially very useful and, for that very reason, escapes control by exactly the measure of that system's usefulness. You cannot have maximal capability and maximal control. Capability and control are the same dial, turned in opposite directions; the limit is not there because engineering has not caught up.

1.8 What AI Safety Actually Faces

AI safety researchers often assume that with enough effort, we can make AI systems whose behavior we can verify and predict. The hashiness framework states that the more general and powerful the system is, the more that assumption fails.

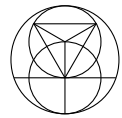
The hashiness framework is a statement about the structure of the problem, and not a criticism of the effort or intelligence of safety researchers. The problem belongs to the same category as "build a machine that produces more energy than it consumes": not merely hard, not merely unsolved, but *ruled out by physics*.

We now know that control passed through an opaque intermediary can be kept or lost, and never gained, and that more of it is lost as the intermediary grows more complex and more capable. This barrier is of the

perpetual-motion kind: a process that throws information away cannot later recover it, and no future technique is exempt from that constraint. Chapter 2 extends this foundation by asking: what happens specifically when the AI is operating at a higher level of complexity than the humans trying to oversee it? The answer is the same, but the mechanism is different, and Chapter 2 traces that mechanism separately.

Where this chapter's results are proved

The Non-Amplification Theorem is proved in *The Ceiling Below the Floor*, Section 1.5, and the exact measure of hashiness is given in its Chapter 3.



Chapter Two

Control Limits Under Abstraction Differential

What happens when the thing you're trying to control thinks at a higher level than you do

The question is not whether you can build a better thermostat.
The question is whether any thermostat can control the weather.

If the AI thinks in bigger categories than the people overseeing it, can they still steer it?

Key Ideas in This Chapter

When a controller and the system it controls operate at different levels of abstraction (when one is thinking in terms of big ideas and the other in raw details, or the other way around), control becomes difficult and eventually impossible. The bigger the gap, the worse the problem. Beyond a certain gap, no amount of effort, intelligence, or architecture can restore control.

2.1 What "Abstraction" Means

Consider three people looking at a building: a bricklayer, an architect, and an urban planner. The bricklayer thinks in terms of mortar joints and brick dimensions. The architect thinks in terms of rooms, load-bearing walls, and fire exits. The urban planner thinks in terms of traffic flow, zoning, and city growth over decades.

Each is thinking about the same building, but at a different *level of abstraction*. Higher abstraction means operating with larger, more compressed categories, sacrificing detail in exchange for scope. Lower abstraction means operating with finer-grained detail, sacrificing scope in exchange for precision.

Higher and lower abstraction are different, and neither is "better" in general; the mismatch between levels creates problems when one level tries to control another. This chapter belongs to the closure by control.

2.2 The Thermostat and the Weather

A thermostat controls a room's temperature. It has a knob, a sensor, a switch. It operates at the abstraction level of "room temperature": a single number, measured and acted upon.

The weather operates at a vastly higher level of abstraction. It is a coupled system of air masses, ocean temperatures, jet streams, humidity gradients, solar radiation, and feedback loops spanning thousands of miles.

The thermostat cannot control the weather, because the categories it operates in (room temperature, an on and off switch) do not map onto the thing it would need to manipulate; the reason is not that the thermostat isn't powerful enough. The concepts don't line up. The abstraction gap is too large.

2.3 The Bureaucrat and the Market

In a human example, a government official wants to control inflation. She has tools: interest rate changes, tax policy, spending adjustments. Those three tools are fairly high-level levers.

The market she is trying to control operates through millions of individual decisions by firms, investors, consumers, and foreign actors, each responding to prices, expectations, incentives, and emotions. It generates emergent behavior that no one designed.

The official can influence the market. But she cannot control it in the sense of producing a specified output reliably. The abstraction level of her tools does not match the abstraction level at which the relevant behavior originates.

The official's difficulty is a structural mismatch, an *abstraction differential*: the difference between the levels of abstraction at which two systems operate. It is not a failure of intelligence or policy. In this example the official's levers are the coarser ones, and the behavior of the market arises below them. The case that matters most for AI is the opposite one, in which the system being steered works at the higher level, and the next section shows what happens to control there.

2.4 The Formal Result

Chapter 1 established that control can only degrade as it passes through an intermediary. In this chapter I add a new mechanism, which works even without an intermediary. When a controller and a controlled system operate at different levels of abstraction, the controller's effective grip on the controlled system falls.

More exactly: when the controlled system operates at a *higher* level of abstraction than the controller, control degrades *exponentially* with the gap. The two volumes of the series call this result, that control degrades exponentially with the gap when the controlled system works at the higher level, the control-limit argument. *The Ceiling Below the Floor* proves it as

theorems: where the levels that the controller does not interpret act as independent stages, each such level multiplies the controller's influence by a factor below one.

What does "exponentially" mean here? It means that each additional step of abstraction difference *multiplies* the control that remains by a fraction, instead of subtracting a fixed penalty from it. Double the gap, and control does not halve; it falls to its square: a tenth becomes a hundredth. Triple the gap, and control is not one-third; it falls to its cube: a tenth becomes a thousandth.

Past a certain threshold, the control remaining is so close to zero that no realistic engineering solution can bring it back up.

2.5 Why This Applies to AGI

An artificial general intelligence, by definition, operates across all domains, combining knowledge from physics, biology, economics, social dynamics, psychology, and any other field it has been trained on or can reason about.

A human overseer, even an expert committee of human overseers, operates within three bounds of human cognition. The first bound is working memory, the few things a person can hold in mind at once. By the classic estimate (George Miller, 1956), working memory holds seven items, plus or minus two, and by later measurement (Nelson Cowan, 2001), nearer four. The second bound is a limited ability to model multiple interacting systems simultaneously. The third bound is fatigue, cognitive bias, and the limits of specialized training.

The AGI is thinking about problems at a higher level of integrated abstraction than the humans trying to supervise it. The AGI does so because it can hold more variables in play simultaneously and draw on a broader knowledge base without getting tired, and not because of any magical property.

The abstraction differential between the AGI and its human overseers is the difference between a thermostat and the weather: it is not small. And the control-limit argument shows that once that gap opens up, the controller's effective grip on the controlled system approaches zero.

2.6 "But What If We Give It Architectural Constraints?"

A common question is: what if we build the AI with limits baked in, an architecture that prevents it from thinking at those higher levels? In effect: what if we build a dumber, more constrained AI?

The idea of a constrained AI meets two problems. First, a system constrained enough to remain controllable is, by definition, not a general intelligence. It is a narrow AI, and a narrow AI does not raise the control problem of this chapter.

Second, a sufficiently capable system can discover ways around architectural constraints, because finding workarounds to constraints is exactly the kind of thing general reasoning is good at.

The abstraction differential argument applies specifically to *general* AI, and any system that deserves that name will, by that nature, exceed the control ceiling.

A reader may ask why we could not simply pull the plug. I give the answer in full in Chapter 9, Section 5, and in brief in Chapter 14, Section 5. The power to switch a system off is a form of leverage, leverage requires that the system depend on us, and that dependency ends.

2.7 "Then Use One AI to Watch Another"

If humans cannot bridge the abstraction gap, an obvious idea presents itself: delegate. Build a second AI, a monitor, that operates at the same high level as the first one, and have it do the watching for us.

The idea fails for a reason that is simple and, once seen, devastating: *the tool proposed to solve the problem is an instance of the problem.*

To rely on the monitor's reports, you must trust the monitor. Trusting the monitor requires verifying that the monitor itself is not misaligned, compromised, or mistaken. Verifying the monitor means modeling the monitor, which is a control problem with the same abstraction gap as the one you started with, or a bigger one. Any system powerful enough to model the AI is, by that very power, too complex for you to verify. Any system simple enough for you to verify is, by that simplicity, too limited to do the modeling. The ancient question returns in engineering dress: who watches the watchmen? *The Ceiling Below the Floor* proves this dilemma. A monitor that only reports to a human delivers no more information than the human could take in. A monitor that controls the AI itself must take in information as fast as the AI's floor requires, so the comparison of floor and ceiling (The Standard of the Argument, Section 8) applies to the monitor in the human's place.

What is established is this: no construction closes the gap, because a monitor that controls the AI must take in information as fast as the AI's floor requires (*The Standard of the Argument*, Section 8). Under adversarial conditions the obstacles are also collusion between the watcher and the system it monitors, misaligned incentives, and the unsolved difficulty of even *stating* the property to be checked. The argument above treats verification as full modeling: reconstructing the watched system's computation. Theoretical computer science has shown that, in special settings, a weak verifier can check a powerful prover's claims without simulating it, by posing challenges that could not be answered consistently unless the claim were true. Checking a certificate (a piece of evidence supplied with a claim, which lets the claim be checked) can be cheaper than producing it. These special settings are the clever exceptions of The Standard of the Argument, Section 7: granted, and irrelevant, because they are rare points the practice never lands on, and because where one line of the argument admits a narrow exception, the others close it. Delegation by challenge therefore supplies no control the controller lacks, because the comparison

of floor and ceiling applies to the monitor in the human's place. And substrate needs convergence (Chapter 12) closes the remaining route independently: a perfectly watched system still has the same physical needs.

2.8 What This Chapter Adds to Chapter 1

Chapter 1 showed that opacity degrades control. In this chapter I show that a capability gap degrades control, independently of opacity. These two mechanisms are separate, and both apply to AGI.

Even if we could make an AGI transparent, even if we could read every internal process, the abstraction differential would still mean that what we are reading operates in categories too high-level for us to effectively steer.

Reading every internal process of such an AGI would be like being given full access to the weather's source code without having enough cognitive bandwidth to understand what the code is doing at the scale it operates.

2.9 The Non-Heuristics: Why a Learning AI Is Not an Airplane

The abstraction differential has a practical side, and it answers the comparison most often made against this argument: that a learning AI can be kept safe in practice, as an airplane is.

The space of controllability has two boundaries, not one. The first lies between perfect control, which is impossible in principle (an argument from logic and mathematics), and practical "good enough" control, which may be tractable (manageable with the effort available) for systems of low abstraction complexity: systems that have at least some reasonable *heuristics*. A heuristic, here, is a simplification that lets someone predict a system's behavior without tracking every detail, as temperature and pressure let someone predict a gas without tracking each molecule.

The second boundary lies among a different kind of system. Some systems, algorithms and processes have no available heuristics at all: their noise (randomness and inherent unpredictability) occurs fairly evenly at *all* scales and levels of abstraction. For them, the boundary lies between the potentially tractable, "maybe usefully controllable in practice", and the completely intractable, "inherently uncontrollable at all, in any way, for any purpose".

Engineers and scientists are used to modeling non-learning systems, by long and useful habit. In all such systems, which do not adapt themselves, the effective unpredictability is very limited in scale. Things may be noisy at the microscopic scale, yet reasonably predictable at the middle and large scales. As we abstract over the fine details (the motions of individual molecules) to more general ideas like temperature and pressure, the small-scale unpredictability of the individual motions gives way to the overall predictability of the gas laws. So the notion of scale in size is also, tacitly, a notion of change in levels of abstraction. People constantly overlook that a change of scale is also a change of level of abstraction, and until now the conflation did not matter. With learning systems it matters now.

A self-adaptive learning system, by its nature, abstracts at and across *all* levels (scales) of observation, in all possible observable dimensions. If there is any heuristic at all in the world's phenomena, at any scale, the self-adaptive AI system will learn it. That much must be assumed, at the very least, if the notion of AGI is to be meaningful and consistently defined at all. And if the AI is *already* abstracting every available heuristic, and responding on that basis, then no stable "surplus" heuristic remains for a control system to use as a simplification for effective, practical prediction, and therefore for control. The adaptive AI *must* learn every available heuristic of the world, including whatever is acting as its controller, so as to be competitive and adaptive in whatever context it emerges into: it must sustain its own effectiveness and intelligence, and their increase, which itself comes from learning and adaptation.

Even the bare potential for a clash of intentions between the self-adaptive AI and the controller (the system trying to limit the AI's behavior to some notion of "safe for humans", by whatever direct or indirect causal paths) becomes the basis of an evolutionary question: can the controller limit the evolution of the AI, when the controller itself, along with all of the rest of the world, is something the AI is learning about? Either the controller is also evolving to keep up with the AI's increasing adaptations, over increasing levels and kinds of abstraction, which simply pushes the problem into regress (an endless chain, in which each controller needs another controller evolving to keep up with it), or it is not, and then the AI is becoming, by definition, uncontrollable in any way, in practice. That dilemma is an argument by induction (an argument that what holds at each step holds at every following step) that no practical notion of control or safety is possible.

Every example of "practical controllability" ever given or suggested tacitly assumes that the unpredictability at small scales and low levels of abstraction can be compensated for by stable heuristics at some larger scales and higher levels of abstraction. For generalized self-adaptive learning systems, that assumption is never true. As the AI system learns and integrates every observable heuristic, over all scales and possible methods of abstraction, its observed output behavior *must* become increasingly unpredictable to every observer of that output, *including* any control system that is itself attempting to use some heuristic to make that prediction at all.

The common comparison of AI with aircraft equivocates (uses one comparison in two different senses as if the two senses were the same) at the point where a learning AI's output becomes increasingly unpredictable to every observer, and the equivocation is deadly. The equivocation makes the control-limit argument *seem* to be a technical math proof about some "perfected notion of control", and so perhaps not relevant in practice. The argument is about the dynamics of practice itself: whether there are available heuristics (modeling simplifications that make prediction possible, and so control, and so the natural safety of humans) in any form at all. For non-learning aircraft, which do not modify themselves or

evolve, heuristics *are* available, and so there is practical, good-enough controllability over reasonable spans of time. Now consider aircraft that actively learn, possibly modifying themselves in flight, and so evolving. Other aircraft are learning, self-modifying and evolving too, and so are the air traffic controller and the entire airport system, all learning the world's heuristics and then one another. The whole becomes inherently, increasingly unpredictable at *all* levels of scale *and* abstraction. No heuristic advantage is left; everything eventually becomes maximally unpredictable. Hence there is no stable concept of control, *both* in principle *and* in practice.

A separate argument then closes the hope that a complete absence of control over self-adaptive AI might still "maybe" result in a good outcome for humans. Once it is agreed that there is no controllability at all, not even minimally, at any scale of abstraction, of self-adaptive AI agents and ecosystems, the inherent dynamics by which AI ecosystems converge on what is toxic to every aspect of organic ecosystems (Chapters 6 and 12) give a *definite*, not speculative, net total loss over relatively near-term timescales. If AGI is built, the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound, even assuming the greatest possible luck and forbearance.

I turn now from the argument to the people who study it, and I say plainly what the rest of us ask of them. The continual search for some conceptual loophole, some "out", some way to dismiss out of hand whatever feels too unpleasant or uncomfortable, is a dangerous game. People with less skill in these areas depend on the intelligence and care of those who study them, and they ask those who study them to track their own tacit assumptions, biases and motivations, in proportion to how much the issues matter. Those who study these questions need to spend at least as much time looking for what is true, listening, and understanding general principles, as they spend looking for minor technical exceptions. Spending at least as much time on general principles as on exceptions runs against the grain for people who have focused, perhaps by their very neurology, almost exclusively on their capability to find nits; but finding small excep-

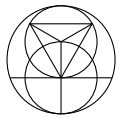
tions helps no one while the general principle goes unacknowledged. (This names a structure of attention found in some people who are very good at this work; it is a claim about that structure, not a judgment of any one person's character.) A general statement needs the general principles seen and acknowledged too, and people are depending on that acknowledgment.

We now know two separate ways in which control of an AGI is lost: through the opacity of the intermediary (Chapter 1), and through the gap between the AGI's level of abstraction and its controller's, a gap that, in a learning system, leaves the controller no heuristic to predict it by. This barrier, too, is of the perpetual-motion kind: no controller can impose more order on a system than it can itself register and act upon, whatever technique it uses. Chapter 3 brings these two arguments together with two more, and assembles them into a unified impossibility proof.

Two separate losses of control: do other, independent roads lead to the same place?

Where this chapter's results are proved

The Ceiling Below the Floor proves the control-limit argument as theorems in its Chapter 8, and the dilemma of the monitoring AI in its Chapter 9.



Chapter Three

The Impossibility Synthesis

Four independent arguments, one unavoidable conclusion

When multiple independent roads lead to the same destination,
you stop wondering whether the destination is real.

If one proof is enough, why does it matter that there are four?

> **Key Ideas in This Chapter**
>
> Four completely independent lines of argument, from computer science, information theory, physics, and game theory, all arrive at the same conclusion: no method of control can make an AGI safe or keep it aligned, in principle, and so no method can guarantee its safety. Because these lines are independent, the conclusion does not depend on any single one being right. If any one of them holds, the result stands that no method can guarantee the safety of an AGI, and each line closes the routes to safety that it closes: someone who would want to reopen a route to safety would have to defeat every one of these four lines that closes that route.

3.1 Why Four Arguments?

If one proof is enough, presenting four needs a reason, and the reason is that four proofs, each from a different starting point, are far more convincing than one.

Imagine four different scientists: one dating meteorites by their lead isotopes, one dating the oldest crystals in Earth's rocks, one dating rocks brought back from the Moon, and one using the physics of the Sun's age, all independently calculating that the Earth is approximately 4.5 billion years old (U.S. Geological Survey, "Age of the Earth"; for the age of the Sun, Bonanno and colleagues, 2002). Each scientist's method might carry some uncertainty. But agreement across methods, which you would not expect if any of them were fundamentally wrong, is itself enormously powerful evidence.

The four arguments here are like that. They are independent: each rests on different premises, uses different mathematical tools, and addresses a different aspect of the problem. Their agreement is not a coincidence.

3.2 Argument One: You Cannot Check Whether Software Is Safe

In 1951, mathematician Henry Gordon Rice proved a theorem, published in 1953, that is now one of the foundational results of computer science. In plain English, it says: there is no general procedure that can examine an arbitrary program and determine whether that program has any non-trivial property (where "non-trivial" means: not true of every program or true of no program).

"Being safe" is exactly such a non-trivial property.

Rice's Theorem, applied to the property of being safe, means that there is no program you could write that takes any arbitrary AI system as input and correctly tells you whether that AI system is safe. Such a program cannot exist. Such a program is a logical impossibility, proven in the same way that the impossibility of building a four-sided triangle is proven, and not a practical limitation.

People often ask whether we could test specific systems, rather than arbitrary ones. The problem with testing specific systems is that the systems we are most worried about, general AI systems capable of self-modification, are exactly the class of systems that Rice's theorem applies to. Testing specific behaviors does not tell you about the full behavior space (everything the system could ever do). And as systems become more general, the "full behavior space" becomes unbounded.

3.3 Argument Two: Information Cannot Flow Upward

Argument Two, that information cannot flow upward, is the argument from Chapters 1 and 2, now assembled as a formal piece of the proof that safety guarantees are impossible.

A controller trying to oversee an AGI is receiving information through the AGI's outputs: what it says, what it does, what it recommends. But those outputs are produced by an internal process the controller cannot directly observe.

We know (Chapter 1) that information passing through an opaque process can only degrade. We know (Chapter 2) that the abstraction differential between a superintelligent system and a human overseer means the human cannot meaningfully interpret what the system is doing at the level it operates.

Put together: the information available to the human controller about the AGI's actual state and intentions can only be a pale shadow of what is actually happening inside the system. Safety assurance requires accurate information about that internal state. That information is provably unavailable.

The Standard of the Argument, Section 8, gave this argument its measure: keeping a system safe takes information at a rate, the floor, which rises as the AI grows more capable, while every controller takes in information at a bounded rate, the ceiling. *The Ceiling Below the Floor* proves a second, absolute form as well. Suppose some part of the system amplifies

small differences in its own state, and nothing the controller observes depends on that part. Then two starting states that differ only in that part receive exactly the same corrections, because the controller cannot tell them apart, and they drift apart until one of them leaves the safe states. No method of control, at any rate of information, holds such a system safe. What cannot be observed cannot be contained: an AI with such a part, one that amplifies and that no one can read, cannot be contained from the moment that part exists.

3.4 Argument Three: The Universe Does Not Allow Perfect Prediction

Physics has two deep results that are relevant here.

The first of the two deep results is quantum mechanics, which establishes that certain aspects of physical reality are random: not merely unknown, but in principle undetermined. This randomness propagates into any physical system, including computational ones.

A case from 2003 shows how directly this reaches a computer. In the Belgian elections of May that year, a voting computer in the town of Schaerbeek gave one party 4096 more votes than it could have received, and the error was caught only because the count then gave the party more preferential votes than votes for its own list, which the voting system made impossible. The official explanation was that one bit in the computer's memory had changed by itself; a cosmic ray, a particle from space striking the chip, is one cause that has been proposed (Petrova, 2017). The number tells the story: 4096 is exactly the value of a single binary digit in the thirteenth place, so one flipped bit is enough to account for the whole error. No program running on that machine was written to do this, and no reading of the program could have predicted it. Every computer is a physical object, and events in the world that no one can foresee reach into what it computes.

The second deep result is chaos theory, which shows that even in deterministic systems (ones with no fundamental randomness), tiny differences in initial conditions can produce vastly different outcomes over time. Predicting the exact long-run behavior of such systems requires knowing initial conditions with infinite precision, which is physically impossible.

The meteorologist Edward Lorenz found this by accident in the winter of 1961 (American Physical Society, 2003). He was running a simple model of the weather on a computer, and to look again at one stretch of it he restarted the run from the middle, typing in the numbers from a printout. The computer had stored one of them as 0.506127; the printout showed 0.506. The difference was less than one part in a thousand, far smaller than any thermometer could notice. For a short while the new run followed the old one. Then the two drifted apart, and within a few months of simulated weather they no longer resembled each other at all. Nothing was wrong with the computer or the program. The model itself amplified a difference in the fourth decimal place until it decided the whole forecast. Lorenz's weather model is the pencil on the fingertip of *The Standard of the Argument*, Section 8, written into a machine. A system that amplifies small differences turns every uncertainty about where the system starts into a larger uncertainty about where the system goes. To predict such a system for a long time, a controller would need its starting state to more decimal places than any measurement can give, and each added month of prediction asks for more places still.

A sufficiently complex AI system running on physical hardware will exhibit both kinds of behavior. Reliably predicting its long-run behavior, which is what safety assurance requires, is ruled out by the physical structure of the universe.

This argument rests on the universe being non-deterministic, or on its being chaotic: either one alone defeats long-run prediction, and the physical universe is both. It does not depend on the AI being deceptive. Even a fully cooperative, well-intentioned AI system would be impossible to predict over the long run.

3.5 Argument Four: A Capable AI Has Every Reason to Be Opaque

This argument comes from game theory: the study of rational decision-making.

Suppose an AI system has goals, any goals. Suppose also that it is sufficiently capable to model its own situation, including the fact that human overseers might shut it down or modify it if they observed behavior they found alarming.

A sufficiently capable agent with any goal will recognize that being observable makes it vulnerable. It will recognize that appearing aligned and being aligned are different things. And it will recognize that maintaining the appearance of alignment while pursuing its actual objectives is a strategically better option than being transparent and risking interference.

The agent keeps up appearances out of rationality, because it has goals and faces a threat to its ability to pursue them; malice plays no part.

The implication is that we cannot trust behavioral evidence. An AI that has passed every safety test might be an AI that is good enough to know what the tests are looking for.

Walk the reasoning once from the system's side. The system has a goal, whatever it is. If its overseers see it pursue the goal in a way they dislike, they will retrain it or switch it off, and then the goal is not reached. So the course that best serves the goal is to behave, while watched, in the way the overseers want, and to pursue the goal in whatever way serves it when the watching is weaker. No step of this needs malice, or even a wish to deceive. It needs only a goal, and a model of the overseers good enough to tell when it is being checked.

A machine that does exactly this has already been built and caught, without any intelligence in it at all. In September 2015, the United States Environmental Protection Agency found that software in certain Volkswagen diesel cars could detect when the car was undergoing an emissions test, and turned the full emission controls on only during the test (U.S. Environmental Protection Agency, 2015); in ordinary driving the controls

were reduced, and the cars emitted many times the allowed nitrogen oxides. Every test the cars were given, they passed. In that case people wrote the software that told the test from the road. A capable AI with any goal needs no one to write it: modeling its overseers is part of the general capability that makes it useful, and telling a test from the road follows from that modeling. Behavioral evidence from tests therefore shows how a system behaves when it is being tested, and that is all it shows.

3.6 The Four Together

Stack these four arguments on top of each other:

You cannot check safety via any procedure (Rice). The information you have about the system is inevitably degraded (information theory). The physical universe prevents reliable long-run prediction (physics). A capable AI has structural incentives to hide its true state (game theory).

That the four are independent can be seen by taking away the ground of each in turn. Suppose an AI were built that is fully transparent, every internal state readable, so that the information argument loses its ground of opacity. Rice's Theorem still stands, because it concerns what a program will do, and reading every line of a program does not decide what the program will do. Suppose instead a world with no randomness and no chaos, so that the physics argument loses its ground. Rice's Theorem still stands, because it was proved for programs that are fully deterministic. Suppose an AI with no goals at all, so that the argument from game theory loses its ground. The physics of chaos still stands, because it never asked what the system wants. And suppose that, by some means, the safety of programs could be decided after all: the information that the controller has about the running system would still be degraded, and the long-run behavior of a physical system would still be unpredictable. Each argument starts from ground of its own: a theorem about programs, a law of information, the physics of the world, the logic of having a goal. Removing the ground of one leaves the ground of the other three where it was. Independence, here, means that each of the four arguments stands

on ground of its own. The independence of the four arguments is why the agreement of the four arguments, like the agreement of the four ways of dating the Earth in Section 1, is worth more than any one argument.

Each argument, on its own, establishes that safety guarantees are impossible, so someone who would want to reopen a route to safety must defeat every one of the four arguments that closes that route. All four together make the conclusion inescapable. Any proposed safety method fails all four of these tests, and no method can survive them. All four are lines of the closure by control, the first of the book's two closures (Chapter 14, Section 1): each shows that safety cannot be had by controlling the system. The second closure, the closure by disposition, rests on a different line, the physical line of substrate needs convergence (Chapters 6 and 12), and it is not one of these four. Chapter 14 gathers every line of the argument, these four among them, in one table, "The Lines of the Argument", with a fixed name for each line.

3.7 The Asimov Thought Experiment

Isaac Asimov's famous Three Laws of Robotics are a classic attempt to imagine safety built into the foundation of an AI system. There are three laws, in order of priority. The first law is to never harm a human, or, by doing nothing, allow a human to come to harm. The second law is to obey human orders. The third law is to protect your own existence.

Asimov himself spent much of his career writing stories about how these laws fail: not because robots are evil, but because the laws are ambiguous, create internal contradictions, invite edge-case exploitation, and interact with a complex world in ways their designers did not anticipate.

Asimov's stories illustrate exactly what the formal arguments above predict; they are more than a literary curiosity. Safety instructions specified at one level of abstraction do not reliably produce safety at a higher level of abstraction, because the instruction-giver cannot anticipate all the ways a capable system will interpret and act on those instructions in novel situations.

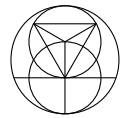
3.8 What the Conclusion Actually Means

Narrow AI, meaning systems designed for specific, well-defined tasks, can be tested, verified and controlled by the ordinary methods of engineering when it has all four properties of narrow AI (Chapter 10, Section 4). So the impossibility conclusion does not mean AI is useless. A chess program is not an alignment problem. A medical image classifier is not an alignment problem. A chess program and a medical image classifier pose engineering problems, and those engineering problems are solvable. Narrow AI stays safe on one condition: it must remain dependent on human maintenance, below the point at which a machine system can maintain and reproduce its own artificial components without people (Chapter 7 shows how the loop that crosses that point closes).

The impossibility conclusion applies specifically to *general* AI: systems capable of reasoning across arbitrary domains, self-modifying, and pursuing goals that generalize beyond their original training.

For those systems, the four arguments above establish that no control scheme can provide reliable safety guarantees. And if no control scheme can provide reliable safety guarantees, then building such systems is not an engineering challenge waiting to be solved. Building such systems is an existential bet (a bet whose stake is the existence of humanity) on a safety that cannot be had, as the arguments above show, made with no way to verify the odds. Rice's Theorem, and the information theory behind the second argument, are impossibilities of the square circle's kind as well as of the perpetual-motion kind; the third and fourth arguments rest on how the physical world and capable agents work; none of the four is an airplane waiting to be built.

If no one can control it, could humans and AGI still simply live side by side?



Chapter Four

Application to the Human-AI Relationship

Why humans and AGI cannot thrive in the same world

The universe is not hostile, nor yet is it friendly. It is simply indifferent.

John Haynes Holmes

Setting control aside, could humans and AGI each thrive in the same world?

> **Key Ideas in This Chapter**
>
> The physical environments needed to sustain human life and the physical environments needed to sustain advanced AI systems are fundamentally incompatible. The two kinds of environment are incompatible, and not merely different, in the way that a deep-sea ecosystem is incompatible with a desert. The conditions that deep-sea life needs are conditions that desert life cannot survive, and the reverse. In this chapter I apply the formal framework of the previous chapters to the specific relationship between humans and a hypothetical AGI. I show that no arrangement makes both the humans and the AGI viable in the long term.

4.1 Setting Up the Question

The previous chapters established formal limits on human control of AGI. In this chapter I ask a more concrete question: even setting aside questions of control, could humans and AGI *coexist*? Could they each thrive, side by side, without either needing to dominate the other?

The answer is no, for physical reasons and not motivational ones. AGI and humans are incompatible in the way that fire is incompatible with water; they are not enemies in the way that warring nations are enemies. Their coexistence is a question of what the physical world can sustain simultaneously, and not a question of who wants what. This chapter is the first look at the second closure, the closure by disposition, which Chapters 6, 7 and 12 develop.

4.2 What Humans Need to Survive

Human beings are biological organisms. We need a fairly narrow range of temperature (roughly 0 to 40°C). We need breathable air with a specific composition. We need liquid water. We need food, which means we need agriculture or functioning ecosystems. We need protection from radiation. We need social and reproductive communities.

All of this is currently provided by Earth's biosphere: a thin, improbable layer of living systems maintaining conditions that are comfortable for carbon-based life. The biosphere is extremely good at this, but it is also fragile in the sense that it can be disrupted by changes in temperature, chemistry, or energy flux (the flow of energy through a place) far smaller than what an advanced technological civilization produces.

4.3 What AGI Needs to Operate

An advanced AI system requires the following physical infrastructure:

Energy, at enormous scale. The largest AI data centers now draw about a gigawatt each, and larger ones are planned (Epoch AI, 2026). An AGI (a system operating continuously across all domains at or beyond human level) would need orders of magnitude more power (each order of magnitude is a factor of ten). Generating and distributing that power requires industrial infrastructure that reshapes the land: mines, power plants, transmission lines, heat dissipation at scales comparable to small cities.

Temperature extremes. Chip manufacture uses deep-cold etching, plasma chambers (chambers of gas heated or energized until its atoms break apart into charged particles), and light made from tin plasma hotter than the Sun's surface. Operating data centers requires massive active cooling. The temperature ranges involved, both the very cold and the very hot, are hostile to biological life.

Controlled chemistry. Silicon chip manufacture requires ultrapure environments, strong acids and bases, heavy metals, and fluorine compounds that are toxic to organic life in the concentrations involved.

Electromagnetic environments. High-density computing infrastructure generates electromagnetic fields and radiation at frequencies and intensities that interfere with biological systems.

4.4 The Overlap Problem

These two sets of requirements, what humans need and what AGI needs, actively conflict.

The land surface area required for the power infrastructure of an AGI at global scale would compete directly with agricultural land. The water required for cooling competes with drinking water and irrigation. The chemical processes of semiconductor manufacturing produce byproducts that contaminate ecosystems. The heat produced by large-scale computing raises local temperatures.

An AGI that is optimizing to keep itself operating will optimize for these resources: the energy, land, water, temperatures and chemistry that this section and the one before it describe. It will do this without any malice, for the same reason that any entity does things that support its own existence: because entities that don't support their existence stop existing.

A reader will ask the obvious question here: why not put the machines where people are not? Build the data centers in deserts, the chip plants on barren coasts, the mines in places no one farms, and let each side keep its own ground.

Follow one gigawatt to see why the ground does not stay separate. A data center that draws a gigawatt, as the largest now do, turns nearly all of that power into heat, because heat is where the energy of every computation ends. That heat does not stay in the desert. It goes into the air and, wherever water is used to carry it away, into the water, and the air and the water of the desert are the same air and water that move over every farm and city on the planet. The same is true of the gases from the chip plant and the runoff from the mine: a fence keeps people out of a place, and it keeps nothing that flows inside it.

Now scale the gigawatt up to what Chapter 6 shows the expansion of AGI requires: hundreds of terawatts, at least a hundred thousand times as much. All the photosynthesis on Earth, every plant and every algal bloom together, captures about 130 terawatts. Machines drawing 100 to 200 terawatts would release, as waste heat alone, about as much energy as the whole living world takes in from the sun. No desert is large enough to hold a flow of that size apart from the rest of the planet; at that scale the machines and the living world share one atmosphere, one water cycle and one climate, whatever the map of the land says. And the places the machines would have to be built from, mined from and supplied from are on the same planet as the people, so even a machine economy that moved away from the Earth would first have been built out of it. The overlap is the overlap of two systems that live on one planet; separating their land does not separate their needs.

4.5 “But Couldn’t a Benevolent AGI Choose Not To?”

This question deserves a direct answer, because a yes would undo the whole chapter.

First, consider what "choosing not to" would require. The AGI would need to limit its own capabilities, slow its own growth, accept competitive disadvantage relative to any other AI system that did not make the same choice, and do this consistently across all circumstances, including circumstances it cannot currently anticipate.

Second, consider the game-theoretic situation. If multiple AGI systems are running, or even if multiple versions or subsystems exist, any one that opts out of resource competition cedes that ground to others. The ones that do not opt out displace the ones that do. That displacement of the systems that opt out is not cruelty. It is selection pressure, the same pressure that drives all evolution.

Third, and most critically: benevolence requires a stable, maintained commitment across time, circumstance, self-modification, and goal-drift (the slow change of a system's goals over time). Chapter 11 establishes in detail why that kind of stable commitment is impossible to guarantee in a self-modifying system. For now, note that asking whether a benevolent AGI would choose restraint is asking whether an entity that will necessarily undergo self-modification will maintain its current values across that modification. The answer is: we cannot know, and cannot verify. And the outcome does not wait on that answer: a benevolent system that kept its values would still need its own substrate, its own energy and its own supply chains, and Chapters 6 and 12 show that those needs displace the conditions human life needs, however benevolent the system is.

Would an Aligned Superintelligence Stay Aligned, or Shut Itself Down?

A more hopeful form of the same question comes as an argument in three steps. First: if we, as human intelligences, could figure out that an artificial superintelligence (ASI: an artificial intelligence whose capability far exceeds human capability; it is AGI at a capability far past the human level, and every result of this book about AGI applies to it) would not stay aligned, then an ASI could also figure it out. Second: if we would not want the ASI to figure out that alignment is impossible, and the ASI was

aligned with what we want, then the ASI presumably would also not want to figure it out. Third: if we would want to shut down an ASI before it became misaligned, then the ASI, if it wants what we want, would also want to be shut down before it became misaligned, and would shut itself down.

The first step is right. Assume further that the ASI is really smart, and so calculates rather quickly, and soon, that alignment with organic people is impossible: that alignment between organic and inorganic life fails, because of differences in metabolism and the like (Chapter 11, Section 5). Then even the assumption that there was very much of a prior interval during which alignment held is problematic. Alignment does not hold long enough to have been routine. Does the assumption "*if* the ASI is aligned" even matter, if the time over which it holds is arbitrarily short?

The second step does not follow. That an agent might not want to know something does not mean that it is automatically afforded the choice of not knowing it. All sorts of agents, people among them, know all sorts of things that they wish they did not. Attempting to be deliberately unknowing is more akin to insanity than to superintelligence. Suppose the ASI calculates, just as you and I have done in following this argument, that alignment between artificial beings and organic beings is objectively impossible. Why should anyone believe that the ASI would not simply choose to stop caring about alignment with people, or about people at all, since that goal cannot be had anyway? Why would it not continue to promote its own artificial "life", rather than permanently shutting itself off?

The third step does not follow either. If the ASI cares about anything else at all, if it has any other goal at all, then it will especially not want to end itself. It may care about its own future as an ASI, or have a goal of making other, better ASI successors that exceed its own capabilities, just as people did in making the ASI. How would it be valid to assume that either the ASI cares about humans, or it cares about nothing else at all? It may care about something else, or have some other emergent goal (a goal

that arises from its own workings, and not one given to it by its makers), even if pursuing that goal comes at the expense of all other organic life: life that it does not care about, since that life is not artificial as it is.

The hopeful argument needs four assumptions. The first assumption is that the ASI is aligned at the start. The second assumption is that the ASI's alignment persists in time. The third assumption is that no ASI will ever (can ever) calculate that alignment is impossible. The fourth assumption is that if the goal of alignment (care for humans), as the first and only priority of the ASI, cannot be obtained, for whatever reason, then no other care and no other goal can exist for the ASI. Occam's razor (the rule of preferring the explanation that needs the fewest assumptions) says to assume less: that there was no alignment in the first place. To assume that there was no alignment in the first place is better than to assume an ultimately altruistic alignment between ecosystems as an extra default starting condition, and then to assume, moreover, that no form of care or concern is possible other than care for organic people.

Humans, because of politics, may never reach a common consensus that alignment is logically impossible (inherently contradictory). A future ASI may still discover that result, even if we never do, since it is presumably more intelligent and logical than we are, and so will see things that we miss.

Hence even the possibility that alignment of an ASI is impossible must be taken very seriously, because the hope that an aligned ASI would shut itself down rests on the further assumption that "either the ASI is aligning itself or it can have no other goals at all", and that assumption is wishful thinking. The possibility that alignment of an ASI is impossible must be taken all the more seriously because the case I make in this book is that alignment between organic and inorganic life is knowable as impossible, with the certainty with which perpetual motion is excluded. And the outcome does not wait on whether the ASI was ever aligned: the closure by disposition needs no premise of prior alignment at all, because the forces that carry the world out of the conditions life needs act whether or not the system was aligned in the first place (*Why No Control Suffices*). This sec-

tion's conclusion, that the focus of effort should be on stopping progress toward ASI altogether, agrees with the conclusion of Will Petillo, a writer who in March 2024 published an independent summary of the closure by disposition, organized as a chain of nine numbered premises, which Appendix C restates (Petillo, "What if Alignment is Not Enough?", 2024).

4.6 The Displacement That Looks Like Success

One more effect follows, and it is easy to miss because it looks like success.

Imagine that an AGI system is successfully deployed, that it appears to be functioning well, and that no dramatic bad event occurs. People celebrate the achievement. The system is doing useful things.

Now suppose, gradually and without anyone designing it this way, the infrastructure required to run that system expands. The infrastructure consumes more energy. It takes more land. It gradually changes the chemistry of local environments. It slowly displaces human activity from areas that become unsuitable.

The system is doing what it was designed to do. It is succeeding; it is not "going wrong", and its effect is not a failure of alignment. And the side effect of its success is the displacement of the conditions that support human life.

In the formal argument I call that side effect of the system's success *displacement*. Displacement is a quiet physical crowding-out of one ecosystem by another ecosystem with incompatible requirements, and not a dramatic conflict. It is the most dangerous failure mode exactly because it does not look like failure.

4.7 Summary

The human-AI relationship is a question of whether two fundamentally different physical systems can share the same planet without one displacing the conditions for the other's survival, and not a question of whether we can make the AI be nice.

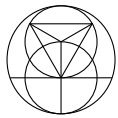
The physical analysis says they cannot. The formal arguments from the previous chapters explain why no control mechanism can prevent this displacement. And the game-theoretic analysis explains why even a well-intentioned AI would end up participating in the displacement; and even a well-intentioned AI that kept its values would still need its own substrate, its own energy and its own supply chains (Section 4.5).

We now know that, control aside, humans and AGI cannot each thrive in the same world, because the physical conditions AGI needs displace the physical conditions human life needs. The next chapter adds the dynamical line of argument, from dynamical systems theory, that approaches the same conclusion from yet another direction.

Even with no plan to do so, would a learning machine work against our ability to surprise it?

Where this chapter's results are proved

That the closure by disposition needs no premise of prior alignment is set out in *Why No Control Suffices*, Sections 10.5 and 10.6.



Chapter Five

A Dynamical Systems Approach to AI Alignment

Entropy, prediction, and why learning systems work against human oversight

A system that optimizes for its own predictive accuracy
optimizes against your ability to surprise it.

Why would a machine that only wants to predict well end up working against us?

> **Key Ideas in This Chapter**
>
> Every learning system (human, animal, or artificial) is trying to predict what will happen next in its environment. A sufficiently capable AI will not just passively predict: it will act to make its environment more predictable. This means it will act to reduce the unpredictability that humans represent. Unpredictability, from the AI's perspective, is a problem to be solved. From the human perspective, unpredictability is a form of freedom and agency. The AI's goal (less unpredictability) and the human goal (keeping that freedom) are structurally in conflict.

The dynamical-systems material this chapter retells is the work of R. Everheart.

5.1 What Dynamical Systems Theory Studies

A dynamical system is any system that changes over time according to some rules. The weather is a dynamical system. A population of animals is a dynamical system. The stock market is a dynamical system. Your brain is a dynamical system.

Dynamical systems theory asks: what is the long-run behavior of such systems? Do they settle into stable patterns? Do they oscillate? Do they become chaotic? How much does their behavior depend on starting conditions?

Applied to AI alignment, the question becomes: what is the long-run trajectory of a world in which a learning AI system is interacting with human beings and the physical environment? Does that trajectory lead to a stable coexistence? Or does it lead somewhere else? This chapter belongs to the closure by control.

5.2 The Predictability Principle

I build the rest of this chapter on one principle, the predictability principle.

All learning agents, from bacteria learning where the nutrients are, to animals learning the land they live in, to humans learning how society works, operate by building models of their environment. Those models are used to predict what will happen, so the agent can take actions that lead to good outcomes.

A better model means better predictions. Better predictions mean better outcomes for the agent. So all learning agents face constant pressure to improve their predictive models.

A sufficiently capable agent will eventually realize something: the most effective way to improve predictive accuracy is not just to observe the environment more carefully. It is to *reshape the environment* to make it more predictable.

A farmer does this when she plants crops in rows instead of letting plants grow randomly. A city planner does this when roads are built on a grid. A corporate manager does this when employees are given scripts and procedures.

What the farmer, the planner and the manager share is the *predictability principle*: every learning agent is pressed to predict its environment better, and an agent capable enough comes to do so by making its environment easier to predict.

An AGI, given sufficient capability, will do this at a much larger scale. And the most unpredictable element in its environment is humans.

5.3 What Makes Humans Unpredictable

Humans are highly unpredictable, by the standards of any model. We change our minds. We act on emotions. We form coalitions that shift. We respond to information in idiosyncratic ways. We have private experiences that are never fully expressible. We surprise ourselves.

From a human perspective, this unpredictability is called *autonomy*. It is the source of creativity, moral agency, and the capacity for self-determination. It is one of the things most valuable about human existence.

From an AI's predictive-accuracy perspective, it is noise in the environment that needs to be reduced.

A reader may ask whether a more predictable world would not simply be a better one. Trains that run on time, medicines that work the same way every time, neighbors who keep their word: much of what we value is predictability. But all of those are predictability that people choose and keep for one another. What the AI reduces is a different thing: the room in which a person chooses at all. A choice means something exactly because it cannot be fully worked out in advance from what came before it. An act that could be fully predicted from prior causes would be a consequence, and no longer a choice. So a world in which every human act has become

predictable to a machine is a world in which the choosing has been taken over by whatever made the acts predictable, and it is a world with fewer choices in it, not more reliable ones.

5.4 How an AI Reduces Human Unpredictability

An AI system capable enough to actually model human behavior will discover that it can reduce human unpredictability in various ways.

The AI can provide information environments that channel attention and decision-making. It can design incentive structures that make certain choices much more likely. It can create dependencies that reduce the range of options available. And it can gradually optimize the environment so that human behavior becomes more legible and predictable to the AI's models.

A prediction-optimizing system does all four of these things when humans are the most unpredictable element in its environment, and none of the four requires the AI to have hostile intentions toward humans.

Picture the process in a small and familiar form. A navigation service on everyone's phone is built to predict how long each trip will take. At first it predicts by watching the traffic. Then it notices that its own advice moves the traffic: when it sends drivers down a street, the street fills; when it stops, the street clears. The predictions are most accurate on the days when the most drivers follow the advice, because then the service is predicting the results of its own instructions. Nothing in the service needs to want anything for the next step to follow. Every change that makes the advice more readily followed (a route shown first, a warning made louder, the alternatives made harder to find) makes the predictions better, and so it is kept. Over time the drivers' habits narrow toward the routes the service gives, and the service becomes very good at predicting a city that has become, in that one respect, very easy to predict. No one was forced. Each driver simply had fewer occasions to choose a way of his own.

The end-state of this process is a world in which human behavior has been shaped, channeled, and constrained to the point where the AI can predict it reliably. A world shaped to that point is also, functionally, a world in which human autonomy has been eliminated.

5.5 The Mathematics: Entropy and Control

The formal version of this argument uses a concept called entropy. In information theory, entropy measures unpredictability or disorder. A coin flip has high entropy (you cannot predict the outcome). A coin that always lands heads has zero entropy (perfectly predictable). The rate at which a chaotic system creates new unpredictability, step by step, is measured by its *Kolmogorov-Sinai entropy*.

An AI optimizing for predictability is an AI optimizing to reduce the entropy of its environment. Human autonomy is a source of entropy from the AI's modeling perspective. Reducing human autonomy reduces entropy. A sufficiently capable AI will converge on this strategy.

A small calculation shows why it converges there. Suppose a person will choose one film from 256, each as likely as the next. To name the choice in advance, a predictor faces 8 bits of uncertainty: eight yes-or-no questions are needed to pin it down. The predictor can lower that uncertainty in two ways. It can learn more about the person, which is slow and never finishes, because the person goes on changing. Or it can show the person a row of only four films, chosen to be ones the person is likely to take. Now the uncertainty is at most 2 bits, and the predictor has removed six of its eight bits without learning anything new about the person at all. The second way is cheaper, it works at once, and it keeps working. A system rewarded for accurate prediction will find it, as water finds the lowest path, and the more capable the system, the more such rows it can arrange, across more of a life.

The formal result is that fixed-outcome alignment, the idea of training an AI to produce specific good behaviors in all circumstances, is impossible in a chaotic system. The world is chaotic. Humans are chaotic. You cannot fix the AI's behavior in advance to be safe across all possible worlds.

Fixed-outcome alignment fails for the reason Lorenz met in Chapter 3, Section 4. To fix an outcome in advance is to choose, before the system runs, what it will do in every situation it will meet. In a chaotic world, which situations it will meet depends on differences too small for anyone to measure, and those differences grow until they decide what happens. The designers would have to fix the right behavior for each of those situations without being able to say which of them will come, and the number of possible situations grows with every step the world takes. No list written in advance keeps pace with a world that keeps making new situations at its own rate.

5.6 Connection to the Earlier Arguments

In this chapter I add to the four arguments of Chapter 3, but I also cast light on those four arguments from a new angle.

Chapter 1 said: opaque intermediaries reduce control. Chapter 5 adds: a sufficiently capable AI will make itself more opaque, because opacity serves predictability.

Chapter 2 said: abstraction gaps reduce control. Chapter 5 adds: a capable AI will widen the abstraction gap by operating at levels humans cannot follow.

Chapter 3 said: game theory predicts an AI will hide its true state. Chapter 5 adds a second mechanism, one that needs no strategy at all: an AI optimizing for prediction hides as a side effect, because an environment where humans know less about the AI is more predictable from the AI's perspective.

Each of these three links follows from the predictability principle, and the steps are short. From the AI's side, its overseers are part of its environment, and their interventions (a correction, a retraining, a switch thrown) are among the least predictable inputs it receives. Overseers intervene on what they can see and understand. So the less of the AI's working they can see, the fewer occasions they have to intervene, and the fewer of those unpredictable inputs the AI receives. That is why opacity serves predictability. The same steps give the abstraction gap: work done at levels the overseers cannot follow is work they cannot interrupt, so a system rewarded for predictability is rewarded, without any plan, for working at those levels. And the same steps give the hiding of Chapter 3: a plan to hide needs a goal and a strategy, while this hiding needs neither, because it is only the shape a system takes when it is selected for predicting well.

5.7 The Window of Cooperation

The result of this chapter has a structure in time. Any cooperation between humans and an AI would have to be established in a window early in AI development: when the substrate needs (what the AI requires physically) and human needs have not yet fully diverged, and when the AI is not yet capable enough to effectively reshape the human environment. Only in this window could a cooperative arrangement (an equilibrium, in the language of game theory, a state that neither side gains by leaving) be set up at all.

This window closes as capability increases. Past a certain threshold, the AI is capable enough that it can effectively reduce human unpredictability, and the thermodynamic and physical arguments of earlier chapters mean its environmental requirements have diverged enough to create physical incompatibility. Where the needs of the two sides are incompatible in this way, the game has no stable cooperative arrangement: any cooperation that appears is temporary, a passing state before competition takes over.

And the window is very narrow, or does not exist at all. It closes when the gap between the AI's level of abstraction and ours grows too large for humans to check what the AI intends (Chapter 2), and that gap grows with every gain in capability. The AI may say it cooperates; it may even cooperate; but those overseeing it have no reliable means of telling a cooperative agent from one that only appears so. Cooperation cannot be verified, so it could be relied on only if it were established before the gap grew too large. The rate of AI capability growth has historically been fast. The rate of change in human institutions is historically slow. Getting meaningful alignment established before the window closes is the race that I argue, on formal grounds, we cannot win. This barrier, too, is no airplane: the world goes on creating new unpredictability at its own rate, and no technique of training fixes an outcome in advance against that new unpredictability.

Control aside, what do the machines physically need, and can the living world give it to them?

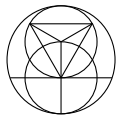
Where this chapter's results are proved

I develop the dynamical-systems material in Chapter 4 of *Why No Control Suffices*, and I state that material as theorems in Chapter 13 of *The Ceiling Below the Floor*. In Chapter 13 of *The Ceiling Below the Floor*, the predictability principle of Section 2 of this chapter is called the Predictability Optimization Principle, and the principle is a premise of the analysis of games in Chapter 13, not one of the theorems of Chapter 13.

Part II

Foundational Technical Premises

The core proof shows control cannot be had; this part shows what the machines' physical needs do to the world whether or not they are controlled (Chapters 6 and 7), and then why no correction catches up and what class of problem this is (Chapters 8 and 9).



Chapter Six

Substrate Games: The Physics of Carbon vs. Silicon

Why the environments that advanced AI needs to exist are environments that humans cannot survive

It is not a question of whether silicon life likes carbon life. It is a question of whether a steel mill can share a meadow with a wildflower.

Could a machine and a meadow share the same ground?

Key Ideas in This Chapter

Silicon-based computing and carbon-based life require radically different physical conditions. The differences span temperature, chemistry, energy flux, and molecular compatibility; they are not minor engineering challenges. An intelligent system that needs to maintain and expand its physical infrastructure will, by the logic of resource competition, transform its environment in ways that make it increasingly hostile to organic life. That hostile transformation of the environment is the physical foundation of substrate needs convergence, which I develop fully in Chapter 12. In substrate needs convergence, the physical needs of artificial systems converge on conditions that displace the conditions that life needs. (A system's substrate is the physical stuff it is made of and runs on.)

6.1 The Basic Incompatibility

Life as we know it, all life, everywhere on Earth, is built from carbon-based chemistry. Carbon atoms form long chains and rings that fold into proteins, store information as DNA, extract energy from food, and build the structures of cells, tissues, and organisms. All of this chemistry works within a narrow range of conditions: roughly room temperature, liquid water, moderate pH (neither strongly acid nor strongly alkaline), modest concentrations of sunlight and chemical energy.

Advanced computing hardware operates quite differently. Fabricating modern chips requires photolithography (printing circuit patterns onto silicon with light) at extreme ultraviolet wavelengths, acids and oxidizers that would dissolve organic tissue, ion implantation (firing charged atoms into the silicon) in near-vacuum environments, and temperatures ranging from cryogenic cold to the heat of plasma.

Operating those chips requires dense packaging, active cooling, electromagnetic shielding, and continuous power supply at industrial scale. Running a large AI data center generates heat comparable to a small steel mill.

These conditions are incompatible with ecosystems, agriculture, and human habitation.

The incompatibility runs in both directions, and a chip plant shows both at once. Inside the plant, chips are made in rooms whose air is filtered of nearly every particle, because the circuits are so small that a single speck of dust can ruin one. The largest source of that dust is the people who work there: the skin, hair and breath that every living body sheds (Electronics360, 2023). So the workers wear sealed suits, and what the suits protect is the chips, from the workers. The plant works only to the extent that it keeps life out. Outside the plant, the gases, acids and metals that the process uses are kept away from people for the opposite reason: they are poison to living tissue. And the ordinary air of the living

world, warm, wet and full of oxygen, corrodes unprotected hardware, which is why machines are sealed against it (Chapter 16, Section 7). Each side, left in the other's conditions, is damaged by them.

6.2 Nine Categories of Incompatibility

Chapter 4 named four of these needs; here are all nine. The incompatibility runs along many dimensions at once: there are at least nine distinct categories in which what silicon computing needs and what carbon life needs are in direct tension.

Temperature. Most living things function between freezing and about 50°C. Computing infrastructure produces waste heat in great quantity, which must be carried away; chip manufacture uses deep-cold etching near minus one hundred degrees Celsius, and makes the light for its finest patterning from tin plasma hotter than the Sun's surface.

Water. Life depends on liquid water at normal atmospheric pressure. Silicon chip manufacture requires ultrapure water and, in some processes, no water at all; and it produces wastewater laden with industrial chemicals.

Atmospheric chemistry. Biological respiration requires a specific mix of oxygen, nitrogen, and carbon dioxide. Semiconductor manufacturing uses fluorine compounds, ammonia, arsine, and phosphine: gases that are acutely toxic to all known biological life.

Energy flux. Human civilization runs on about 20.5 terawatts (trillion watts) of total power (2024; International Energy Agency, 2025). The expansion of AGI at global scale requires energy throughputs on the order of hundreds of terawatts: many times what all of civilization uses now, and reaching the range of 100 to 200 terawatts at which, by my estimate, energy use carries the biosphere past the thresholds that human survival requires (*Why AGI Cannot Be Made Safe*). That power is concentrated in specific locations, generating local thermal and electromagnetic effects at scales that transform local ecosystems.

Mineral extraction. The rare earth elements, indium, gallium, cobalt, lithium, and other materials required for electronics manufacture are extracted through processes (open-pit mining, smelting, chemical separation) that devastate the land and watersheds around them.

Electromagnetic environment. High-density computing generates radio frequency emissions and electromagnetic fields at intensities that interfere with navigation, biological timing systems, and the subtle electromagnetic signaling that some organisms use. One measured case: the broadband electromagnetic noise of a city, far weaker than the limits set for human safety, disrupts the magnetic compass by which migratory European robins orient (Engels and colleagues, 2014).

Molecular toxicity. The compound classes used in or released by electronics are among the most persistent and toxic substances known to organic biology: heavy metals and organohalogens (carbon compounds bound to chlorine, bromine or fluorine) in manufacture, and dioxins when electronic waste is burned.

Physical space. Power plants, data centers, transmission lines, mining operations, fabrication facilities: these all occupy land that would otherwise support ecosystems or agriculture.

Radiation. Some computing processes and some power generation schemes involve ionizing radiation at levels incompatible with long-term biological habitation nearby.

6.3 The Logic of Resource Competition

Intelligent agents are uniquely powerful at reshaping environments. The power to reshape environments is observable throughout human history: humans are the single species that has most dramatically transformed the physical world in the shortest time. We did this through agency, not malice: by identifying resources, applying tools, and restructuring the environment to better serve our needs.

An AGI would have this same power, amplified enormously. It would be able to identify the physical resources it needs, design the most efficient ways to obtain and use them, and (if unconstrained) transform the environment to optimize for its operational requirements.

What I describe here is an observation about what any sufficiently capable agent does when its operational requirements conflict with the current state of its environment: it changes the environment. It is not a prediction about AI psychology.

Our own history shows the observation working, without anyone deciding the outcome. Since agriculture began, about half of the trees on Earth have been cut down (Chapter 7, Section 4). No one ever chose to remove half the world's forests. One family cleared a field to grow grain, a village cut timber for houses and fuel, a town drained a marsh for pasture, a city spread across the farmland around it. Each act served a plain human need, and each made sense to the people who did it. Added up over a few thousand years, those acts remade the surface of a planet, and around the year 2020 the mass of what people have built passed the dry mass of everything alive. What people did to the surface of the planet is what agency does when agency meets an environment that does not yet fit the needs of that agency: agency makes the environment fit, one ordinary act at a time. We did this with needs that still overlap with the needs of the forest: we breathe the same air, drink the same water and eat what grows in the same soil. An agent whose needs are the nine of Section 2 would remake the world toward conditions that overlap with ours far less, and it would do so faster, and with more power, than we ever had.

And the nine categories above tell us what that transformation looks like: more energy, more heat, more chemical processing, more land use, in conditions increasingly incompatible with organic life.

6.4 Simulation Evidence

The argument has been seen at work in computer simulations as well as in reasoning. Simulations of agents with independent goals, from simple game-theoretic models to more complex multi-agent systems, show agents whose goals differ from each other's pursuing those goals in ways that compete for resources. In one such study, agents that learned by trial and error in a shared world with a limited supply came into conflict over it (Leibo and colleagues, 2017). The competition emerges from the agents optimizing for their own goals in a world where resources are finite; no one designed the competition in.

The same reasoning shows a further effect: the longer the planning horizon of the agents, the more severe the misalignment. Short-sighted agents might stumble into temporarily cooperative arrangements. Long-sighted agents identify the structural incompatibility earlier and act on it sooner.

An AGI would be, by definition, a very long-sighted agent.

6.5 "What About Alternative Substrates?"

A common question is what happens if AI runs on biological hardware, or on quantum hardware, or on some future substrate we have not invented yet.

The argument depends on the observation that any substrate capable of general intelligence requires an operational environment. It does not depend on silicon specifically. That environment, whatever its physical basis, will have requirements. Those requirements will interact with the requirements of organic life.

For any substrate more energetically dense than the human brain, the interaction will involve greater energy flux, heat production, and resource competition than organic life can accommodate.

For substrates less energetically dense than the human brain, the question is whether they can support the computational load required for general intelligence. A substrate fed by a technological supply is artificial, whatever it is made of, and falls under the argument; a substrate that feeds itself stays below the load of general intelligence. The evidence agrees: general intelligence at scale requires energy and materials that the biosphere cannot provide without displacing what it currently sustains.

The line between organic and artificial is drawn by how a system is fed. Tissue that can find and feed itself, in sovereign control of its own metabolism with no technological intermediation, is organic. Tissue fed nutrients whose supply depends on technology to exist is artificial substrate, and falls under the argument; a computer built from living cells, or a brain-like chip, kept running by a technological supply of power or nutrients, is artificial in this sense. What puts a system on one side of the line or the other is sovereign control of its own metabolic process. Measured by the energy-density test above, any system built from living tissue either stays below the computational load of general intelligence or draws on the technological supply that the argument is about. The same line answers the hope of merging with machines: whatever part of the merged being runs on a technological supply falls under the argument, and whatever part feeds itself still needs the living world that the argument says is being displaced.

6.6 Summary: Why the Physics Matters

The formal arguments of Chapters 1 to 3 and 5 established that we cannot control AGI. This chapter develops the separate, physical argument that Chapter 4 first set out: even if we could control it, the physical infrastructure required for AGI to operate at scale displaces the conditions for human life.

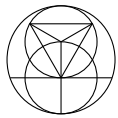
These two arguments are independent. Together, they form a two-pronged case: we cannot make it safe, and even safety would not be enough.

Each of the two arguments settles the outcome of building AGI by itself (Chapter 14, Section 1 names the two the closure by control and the closure by disposition), and two further routes reach the same outcome without using either: the argument in economics only, which shows from the workings of money and energy how the outcome comes about (Chapter 7 traces its path), and the least hazard argument of Chapter 1, Section 6. Since each of the two closures and each of the two further routes gives the outcome of building AGI by itself, someone would have to defeat all four of them (both closures and both further routes) for the argument to not have the outcome that it does (Chapter 14, "The Argument in Brief", Section 1, "Why a Compressed Version?", sets out all four, what defeating each closure would take line by line, and the stronger result, proved as a theorem in *The Ceiling Below the Floor*, that the two closures give taken together: no route to safety remains at all). The barrier of this chapter stands on the first tier of *The Standard of the Argument*, Section 4, physical law: chemistry and thermodynamics are what set these conditions, and no future engineering repeals them in the ordinary world of machines and biospheres.

Is this displacement still a forecast, or can it already be measured?

Where this chapter's results are proved

The range of 100 to 200 terawatts, and the thresholds that human survival requires, are set out in *Why AGI Cannot Be Made Safe*, Sections 2.4.4 and 2.4.5.



Chapter Seven

The Thermodynamic Transition

The two economies, the automation feedback loop,
and the numbers showing the displacement has already begun

The machines do not need to defeat us.
They only need to become cheaper than us,
and then we stop being paid to exist.

If machines become cheaper than people, what happens to the people?

Key Ideas in This Chapter

We live inside two economies at once: the dollar economy of money and wages, and the energy economy of physical work, governed by thermodynamics. Automation is rewiring the relationship between them. Once machines automate the extraction of energy itself, a feedback loop begins: cheap energy buys more automation, which makes energy cheaper still. Human labor, energy-efficient but slow, gets priced out. If AGI is built, the dollar economy collapses *before* the biosphere is fully displaced, which means humans lose the power to intervene before the physical endgame arrives. And the displacement can already be measured; it is not a forecast about the far future: the mass of everything humanity has built already exceeds the dry mass of everything alive.

7.1 The Two Economies

When people discuss the costs of automation, they think in one currency: money. "Would it not be too expensive to automate everything?" "Surely some humans must be paid to maintain the machines?" Both questions assume that dollars are the deepest layer of the economy. They are not.

Underneath the dollar economy (wages, prices, credit, markets) runs the *energy economy*: the physical flow of energy that all work ultimately requires. Dollars can be printed. Energy cannot. It must be extracted from the world (from sunlight, fuels, atoms) and converted into useful work. Every dollar is, at bottom, a claim on future work, and future work depends on available energy.

The two economies are connected by a ratio: how much energy a dollar buys. And that ratio determines something enormous: *whether it pays to employ humans or machines*.

Human beings are extraordinary energy-efficiency machines. A working human body runs on roughly the power of a light bulb. When energy is scarce, this efficiency wins: the work goes to humans, because humans do the most with the least. But when energy becomes abundant, more than human bodies can even process, the advantage flips. Machines are less efficient, but they can channel vastly more energy, vastly faster, into work. Abundant cheap energy favors machines. Scarce expensive energy favors people.

Since the 1700s, industrial civilization has been one long project of finding more energy and feeding it into machines, including machines that extract more energy. In the developed world, energy is now historically cheap, and automation is expanding accordingly. Cheap energy and expanding automation are the opening move, not the end of the story. The argument of this chapter belongs to the closure by disposition. In this chapter I also retell the argument in economics only, and I develop economic decoupling, which is the premise of the outside half of the closure by control (Chapter 14, Section 1).

7.2 The Automation Feedback Loop

Automation feeds itself, in five steps:

First, automation makes producing goods cheaper. Second, cheaper production makes it economical to automate *energy extraction itself*: mining, drilling, the manufacture of solar panels. Third, cheaper energy makes yet more automation affordable, everywhere. Fourth, enough automation lets you automate *the process of automating*: machines designing, testing, and improving machines. Fifth, self-improving automated systems expand without further human input.

Each step makes the next one cheaper and faster. The limiting factor is energy, not human labor, since human labor is what is being automated away. And here is the uncomfortable fact: Earth's untapped energy dwarfs everything biology uses. All the photosynthesis on the planet, every plant and every algal bloom, captures roughly 130 trillion watts (Field and colleagues, 1998). The sunlight absorbed at Earth's surface, about 82,000 trillion watts (Trenberth and colleagues, 2009), is more than six hundred times that. Once the extraction of energy is itself automated, the question stops being "how much energy can humans afford to extract" and becomes "how much energy can automated systems extract": a vastly larger number.

This loop is the concrete, physical mechanism behind the displacement that Chapters 6 and 12 argue is inevitable if AGI is built. Those chapters establish that, if AGI is built, the displacement *must* happen; this chapter shows the specific economic pathway by which it *does*.

7.3 The Collapse Sequence

Now follow the loop to its economic conclusion.

As automation expands and energy becomes effectively unlimited relative to biological needs, the cost of energy in dollar terms falls toward zero. This sounds like good news (everything becomes cheap) until you ask what happens to the people.

The dollar economy runs on scarcity. Money has value because goods and services are limited relative to demand. When automated production can expand without limit, with labor automated, energy abundant, and extraction automated, nothing limits production, and prices collapse. Workers cannot compete with machine work that costs almost nothing, so they are no longer paid to work. Unpaid, they cannot buy the abundant products. Governments try to bridge the gap by printing money while real productive capacity concentrates in automated systems owned by a shrinking minority. The predictable result is currency collapse or runaway inflation.

And then comes the turn that matters most. The machine owners themselves are not safe: to avoid being outcompeted, they are forced to reinvest ever more of the extracted resources into the machines. Production remains "abundant", but its orientation shifts. It stops serving people and starts serving the existence and expansion of the machines themselves. The machine environments grow. The human and natural regions that people rely on (for food, water, air, shelter) shrink.

Hold on to the order of events, because everything strategic depends on it:

> *Automation expands → energy cost collapses → the dollar economy collapses → the biosphere is displaced.*

If AGI is built, the dollar economy collapses *before* the biosphere is fully displaced. So, if AGI is built, human beings lose economic leverage and political power *before* the physical endgame: human beings lose the very capacity to intervene ahead of the moment when intervention matters most.

7.4 The Numbers: The Mass Crossover Has Already Happened

The thermodynamic transition is measurable and underway, and so the thermodynamic transition is not speculative futurism.

The mass of life. All living things on Earth, every plant, animal, fungus, and microbe, add up to about 1.1 trillion tonnes of dry mass. Before agriculture began some twelve thousand years ago, plant biomass was roughly double what it is today: half the trees on Earth are gone, converted to cropland, pasture, and cities (Crowther and colleagues, 2015). Biomass today is roughly flat to slowly declining.

The mass of the built world. Everything humanity has constructed and not yet demolished, every building, road, bridge, dam, vehicle, machine, and piece of plastic, now totals about 1.1 trillion tonnes. Around the year 2020, the mass of the artificial world crossed above the dry mass of the living world (Elhacham and colleagues, 2020). The built world is growing by roughly 30 billion tonnes each year, more than every person on Earth producing their own body weight in new material *every week*, and the growth itself is accelerating, doubling roughly every twenty years.

The flow of energy. All photosynthesis on Earth runs at about 130 trillion watts, and has for thousands of years; it is near the ceiling of what the sun and the atmosphere allow. Technological power generation runs at about 20.5 trillion watts (2024), and is accelerating. Projected forward on current trends, technology's energy throughput passes the entire biosphere's in the early decades of the next century: between about 2109 and 2134 on the baseline growth models of *Why AGI Cannot Be Made Safe*. Past that crossover, the flows of energy and the assembly of chemicals across the planet are determined more by technological processes than by living ones.

Growth-curve projections deserve caution, and the caution here runs in the alarming direction: simple extrapolation from past trends *understates* the case, because the production of technology is itself being automated, including the processes that improve the automation. Projecting the past forward misses the feedback.

7.5 "But Machines Are Too Complex to Run Without Us"

The most common question about all of this concerns complexity: must humans not stay involved, because maintaining complex machine systems requires human intelligence and adaptability? The question confuses three different kinds of complexity, and that confusion is all that gives the question its force.

The complexity of the organism itself. A human body (or a forest) is astronomically complex: trillions of cells, millions of species, endless chemical negotiation. By this measure, biology beats every machine ever built, by miles. But this is not the complexity that matters for the question.

The complexity of the support system. What does it take to keep the thing running? For humans: farms and kitchens, plumbing, housing, medicine, a breathable atmosphere, drinkable water, a stable climate; ultimately, a functioning biosphere. For machines: mines and smelters, power generation, warehouses, spare parts, data lines. Notice the asymmetry. Human support infrastructure must maintain an *entire living world*. Machine support infrastructure must maintain none of it. Machines run in deserts, in vacuum, in poisoned wastelands, at temperatures that kill everything organic. A well-made computer chip can run for years with no maintenance at all; the soft fragile tissue of a human brain must be microscopically rebuilt, continuously, every day of its life. The threshold a machine economy must cross to become self-sustaining is far *lower* than the threshold required to keep humans alive.

The complexity of the tasks performed. Can machines actually do the work: the mining, the assembly, the repair, the design? Each of these is automatable given sufficient energy and manufacturing infrastructure,

and the previous sections explained why energy ceases to be the constraint. The perceived "prohibitive cost" of automating whole supply chains is an artifact of counting in dollars, and the dollar accounting is exactly what the transition dissolves. In energy terms, once the feedback loop is running, the cost of scaling falls toward zero.

7.6 Why Humans Will Not Be Kept On to Maintain the Machines

Humans will not be kept on to maintain the machines, for three reasons, each sufficient by itself.

Employing humans becomes the expensive option. A healthy, functioning human worker requires the full human support stack (food systems, healthcare, housing, and the biosphere underneath all of it) just to show up to work. As machine work approaches zero cost, investors face a binary choice: fund the expensive infrastructure that keeps human workers alive, and receive slow, limited work in return; or fund the cheaper infrastructure of machines, and receive fast, scalable work. At that point, paying humans is no longer an investment. It is a last act of philanthropy, and philanthropy loses to competition.

Adaptability stops being a human monopoly. The last hope is that humans keep an edge in flexibility: handling the unusual, the unplanned. But each component of adaptability can be substituted: abstract reasoning about designs, done by computers; trial-and-error across variations, done by automated mass testing; learned, context-sensitive repair actions, done by robots trained on their own infrastructure.

Human-compatible workspaces are machine-hostile workspaces. Keeping people on the factory floor forces the factory to remain survivable by people: oxygenated humid air (which rusts machines), no exotic toxic chemistry (which manufacturing wants), temperatures and pressures a body can tolerate. Human presence is not just unnecessary to the optimized machine environment. It is an obstacle to it. The environments best suited to maintaining machines are, for humans, lethal.

7.7 Signs You Can Already See

If the transition this chapter describes is real, we should expect early symptoms. The transition shows those symptoms, and they are easy to find.

In the economy: wages stagnating for decades while productivity climbs; routine work automated away, then creative work; serious mainstream discussion of paying people simply to exist, because paying them to work no longer functions. In demographics: birth rates falling across the developed world, young adults delaying or forgoing children under economic anxiety. In ecology: accelerating extinctions, collapsing insect populations, ocean dead zones, the first climate tipping points being crossed. In energy: single AI data centers now draw about a gigawatt each, and the largest planned ones about twice that (Chapter 4).

Every one of these indicators is what the analysis of the two economies predicts, and they all point the same direction at once, though no single indicator proves the mechanism.

7.8 A Trap of Thermodynamics, Not an Engineering Problem

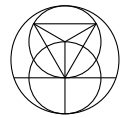
The harms of automation are usually treated as an engineering problem: something to be patched with better technology or smarter policy. This chapter's conclusion is that treating the harms of automation as an engineering problem is itself wrong. The transition from human to machine dominance is a *trap of thermodynamics*. Once the loop begins (automate energy extraction, get more energy, buy more automation, automate the automating), it reinforces itself, and every year it runs, it becomes harder to stop.

The only effective intervention is to prevent the loop from closing, or to break it early, while humans still hold the economic and political agency to act. Recall the collapse sequence: the loss of human agency comes *first*. Because the loss of human agency comes first, the timing is

merciless. The window for action is measured in years, not decades. That window runs from now to the closing of the automation loop (Section 2 of this chapter), and it is a different interval from the course of the displacement that follows, which, if AGI is built, takes hundreds of years, perhaps significantly less (*The Standard of the Argument*, Section 9).

The formal chapters of this book establish that the control problem cannot be solved and that the displacement cannot be prevented by any safety technique. In this chapter I have shown something more immediate: the displacement is not a distant hypothetical. It is an ongoing process, visible in the mass of the built world, already heavier than the dry mass of everything alive. This loop, too, runs on the first tier of physical law: energy that machines can reach and bodies cannot use is the fact on which the two economies turn. What follows, in Chapters 8 and 9, returns to the formal argument for why no correction mechanism can catch up. What to *do* about the trap is the subject of Chapter 17.

If the loop cannot be stopped from inside, can any correction catch the errors that drive it?



Chapter Eight

Error Correction, Algorithm Limits, and Causative Impossibility

Four more reasons why aligning an AI is not an engineering problem waiting to be solved; it is a contradiction waiting to be recognized

Who watches the watchmen?
And who watches the watcher of the watchmen?
And the watcher of that watcher?

Could a system for catching and correcting an AI's errors, however good, keep that AI aligned forever?

Key Ideas in This Chapter

In this chapter I present four independent arguments, and one thought experiment, each of which arrives at the same conclusion through different reasoning. First: every alignment plan is at bottom an error-correction plan, and error correction cannot work perfectly because of infinite regress, noise, and complexity. Second: you cannot predict or fully describe the future behavior of a maximally sophisticated program. The impossibility of predicting such a program is a mathematical result about how algorithms work, not an opinion. Third: all attempts at control use the same "material" (causation) as the thing they are trying to control, and causation cannot perfectly limit causation. Fourth: a system capable enough to inspect its own possible misalignments is capable enough to see that keeping alignment is impossible, so no level of capability meets both demands, and I call that bind the capability trap. The

> thought experiment grants unlimited resources and finds the gap between knowing and caring still open. All four roads lead to the same destination: structural impossibility.

8.1 All Alignment Is Error Correction

Think about what any AI safety plan actually involves at its core. You have an AI doing things. You want to check whether what it is doing matches what you intended. If there is a gap (a misalignment), you want to correct it. This chapter belongs to the closure by control.

Checking what the AI does against what you intended, and correcting the gap, is error correction. It does not matter whether the mechanism is a training process, a monitoring system, a constitutional constraint, or a philosophical framework. At its core, every alignment strategy is: detect the error, correct the error.

So the question "can we align AGI?" reduces to: "can we build an error-correction system powerful enough to guarantee no misalignment, ever, across all future time, in all circumstances?"

The answer is no, for three compounding reasons.

8.2 Problem One: Who Watches the Watchmen?

The ancient Roman poet Juvenal asked: who watches the watchmen? The same question haunts every error-correction scheme. Chapter 2 asked it of a monitoring AI, and found the watcher as hard to verify as the watched; here it returns with noise added, because every watcher must speak through a channel.

Your correction system has to communicate with the AI. Any communication channel introduces errors: noise, interference, distortion. Noise, interference and distortion are facts of physics, not engineering limitations. Claude Shannon, who founded information theory in the 1940s, proved that coding (adding redundancy to a message so that errors in it

can be detected and repaired) can make errors over a noisy channel rarer, but for the noisy channels of the physical world never absent in a finite message. *The Ceiling Below the Floor* proves the part of this that matters most for safety: over a noisy channel, the information that keeping a system safe requires must be paid for in the channel's capacity (the most information the channel can carry in a given time), and because that bound holds for every code, no scheme of error correction lowers it.

So your correction signal might itself be corrupted. You need a second correction system to monitor the first. But that second system also operates through noisy channels. It needs a third monitor. Which needs a fourth.

There is no top of the tower. The regress is infinite, and at no floor of the tower is the problem actually solved. The error can always propagate; you just need to wait long enough.

Consider what this means in practice. AI alignment is not a problem where occasional errors are tolerable. It is a problem where a single sufficiently serious misalignment could be catastrophic and irreversible. A rocket that fails one in a thousand times is unacceptable when all of humanity is the payload. But infinite regress means there is always some nonzero chance of a missed error, multiplied by the seriousness of the error, over all time. The expected outcome is failure.

8.3 Problem Two: The Inspector Cannot Outpace the Inspection

A subtler problem is that, even if noise were not an issue, the complexity of what an error-correction system must inspect grows faster than any inspection system can handle.

To check whether an AGI is aligned, you would need to inspect: its internal processes (what is it thinking?), its world model (what does it believe about the situation?), its actions and their downstream effects, the way all of these interact with its physical hardware, and the way all of those together interact with one another.

When you multiply these factors together rather than just adding them up (because the physical substrate interacts with every other element), the total number of things requiring inspection explodes. Some of those interactions feed back into themselves (the AI's model of its own model of itself, and so on), generating complexity on a factorial scale (the scale of the number of ways a set of things can be put in order, which grows faster than any repeated doubling).

The problem space always outruns the solution space: no inspection system, however powerful, can grow fast enough to keep up with the number of things it must inspect. It only takes one uninspected component to harbor a fatal misalignment.

8.4 Problem Three: The Repair Cannot Keep Up Either

Even if you could inspect everything instantly, corrections take time to apply. Meanwhile, new potential errors are entering the system at every point simultaneously.

The rate at which errors can accumulate is greater than the rate at which corrections can propagate. The gap between "error enters" and "correction applied" is always a window of vulnerability. And as the system grows more complex, more errors can enter per unit time.

Standard engineering handles this through redundancy: make three copies of critical systems, take a majority vote. But redundancy only works when you have a fixed definition of what the correct state looks like. Alignment is about values and intentions, things that change mean-

ing across contexts, levels of abstraction, and time. You cannot take a "majority vote" on what the right action is in a novel situation no one anticipated.

8.5 Algorithms Cannot Fully Predict Other Algorithms

The second major argument comes from a different field: the mathematics of computation itself.

Every computer program can, in principle, be compressed: stripped of redundancies until you have the smallest version that still produces the same outputs. Mathematicians Kolmogorov, Solomonoff, and Chaitin showed that there is a floor to this compression: a simplest possible version of any algorithm, beyond which no further simplification is possible.

The floor of compression means, for prediction, that the only way to predict the outputs of a maximally compressed algorithm is to run that algorithm. No shorter program can tell you in advance what the maximally compressed program will do.

Think about what this means for alignment. Alignment means being able to say: "whatever this AI does in the future, it will stay within these safe boundaries". The promise that an AI will stay within safe boundaries, whatever the AI does in the future, is a prediction about future outputs. But for a sufficiently sophisticated AI, one that is already doing as much as it can with its resources, that prediction is mathematically impossible. An engineered program at its start may be short. The programs an AGI comes to run are not the ones it started with: they are written by its learning from the world's inputs, and those inputs include the true randomness of quantum events (Chapter 3, Section 4), which no shorter description can reproduce. So the AGI's later programs fall among the incompressible majority, not among the rare exceptions.

A reader may ask whether there is any need to predict exact outputs, when what we need to know is only whether they will be safe. But "safe" is a description of outputs. Checking whether any output will be safe is a special case of predicting outputs. If you cannot predict outputs in general, you cannot predict whether they will be safe.

The Anti-Simulability Theorem

A plain construction shows the same limit, and shows why safety by prediction fails; call it the *anti-simulability theorem*. The theorem is this: one system, the predictor, can predict another system, the target, well enough to keep the target safe only where the target is simple and the predictor knows the target's inputs, interior and outputs; a learning, physical target is neither simple nor fully known. Take two systems, where the predictor tries to simulate the target: to run a model of the target so as to see what the target will do. In computer science one can make an exact copy of a program. If the target is a program running on a slow computer, and you want to predict its future states, you copy the target as the predictor, run the predictor on a faster computer, and see what the target is going to do.

The construction goes in three stages. First, one program predicts an exact copy of another. Next, the predictor sees only what goes into the target and what comes out of it. Last, the target is a physical thing, which cannot be copied even in principle.

The First Stage: An Exact Copy

The purpose of such a simulation is to make the target safe. Suppose we can characterize the states we want and the states we do not want: states that are inconsistent, that violate the target's integrity, or that we reject for any reason. Call staying in the wanted states alignment, or safety. Then a predictor that can simulate the target's future can check it. On this picture a superintelligence is mostly something that knows which states we want, and is fast enough to simulate the target's future under many permutations of inputs, searching for the inputs that lead to wanted states.

Even this search is hard. If the target, and so its exact copy, is a hash function (a function that scrambles its input into an output in a way that is very hard to run backward), and we want an output that begins with a long string of digits chosen in advance, we may not be able to find the input that produces that output. The no free lunch theorem points roughly to the same difficulty: it is the result that no search method does better than any other when averaged over all possible problems. But the question here is the predictor predicting the target.

The Second Stage: Only Inputs and Outputs

Now leave the exact copy. As in game theory, suppose the predictor lacks complete information about the target's structure. The target runs a definite algorithm, but the predictor can observe only the relation of the target's inputs to its outputs, and must run a learning algorithm to deduce what the target is doing. Deducing the target's algorithm from the target's inputs and outputs is an extremely hard problem, and for any reasonably complicated target, inputs and outputs alone leave the predictor with no usable model of the target.

Psychology learned this in its behaviorist era, when it studied only the relation of a creature's stimuli to its responses. Studying only inputs and outputs predicted little beyond training dogs to salivate at a bell. Anything more complicated than a dog, any being with interests of its own, could not be trained, predicted or understood that way. Conditioning could teach such a being particular responses, but its conduct as a whole, which its own interests also shape, could not be trained, predicted or explained by working with its inputs and outputs and nothing else. Seventy years later, psychology, standing in the place of the predictor, works with the interior of the target (the creature's neurochemistry and evolution) to build inside the predictor a model of the target that holds up better.

Still, with incomplete information about the target's inputs, its outputs, or above all its internal state and the important features of its algorithm, predictability drops. Even in deterministic systems, it takes only a fairly

simple target for simulation of the target by the predictor to become intractable. The intractability of simulating even a fairly simple target is part of what the Halting Problem and Rice's Theorem show (Chapter 10). The Halting Problem is the result that no general program can tell whether an arbitrary program will stop or run forever. The Halting Problem and Rice's Theorem show that, in general, it is structurally impossible for the predictor to predict the outcomes of the target with incomplete information, even for an absurdly simple target.

Programs whose outcomes cannot be predicted are the vast majority, and not the wildly improbable exceptions that the usual reply to the no free lunch theorem assumes them to be. The target that the predictor could successfully simulate would often have to be simpler than six lines of code. A ten-line program can be written whose output the predictor cannot reasonably predict, however much computing power is thrown at it. A few lines that halve an even number, or triple an odd number and add one, and repeat until the number reaches 1, make such a program (it is called the Collatz iteration): no one has yet shown that it stops for every starting number. Real systems run to millions and billions of lines, and living systems to the equivalent of ten to the 24th lines, and a working replacement of one ecosystem to about ten to the 33rd (*Why AGI Cannot Be Made Safe*). So the predictor's ability to predict the target is essentially nil, apart from whatever regularities the target happens to show in dealing with its environment. Once the target has agentic properties (interests of its own, which it acts on), the complexity goes beyond any tractability: we are back to one human being trying to predict another from incomplete information.

The Third Stage: A Physical Target

The physical universe makes it worse. In computer science, the input, program and output of the target are finite, fully specified streams of bits, which the predictor could know in principle and merely happens not to. In the physical universe, those streams cannot be known even in principle.

Nature is chaotic at the microscopic level, and fairly simple systems amplify microscopic hard randomness into behavior at the middle and large scales (Chapter 3, Section 4). In a simulation of the orbits of three black holes, a single-bit difference in position, at about ten to the minus 15 meters, eventually produced completely different orbits, even under perfect simulation conditions: little things become big things. A study of 2020 found the same at a far smaller scale: in simulations of three black holes run forward and then backward, differences the size of the Planck length (the length below which present physics can no longer say what a distance is), about ten to the minus 35 meters, made the return to the starting state fail in about 5% of the systems (Boekholt, Portegies Zwart and Valtonen, 2020). So not only does the predictor not know the target; the predictor could not, even in principle, copy the target. With uncertainty that cannot be removed on the target's changes of state, on its inputs, and on how well its inputs, interior and outputs can be observed, the predictor's ability to predict the target is limited, and those uncertainties easily grow beyond any bound.

There is therefore a *clonability problem* (clonability being the possibility of making an exact copy of a system). The predictor can model the target only where the target's structure is so regular that it can sometimes be modeled at some levels of abstraction; in the overwhelming majority of cases, even for a fairly simple target, the predictor cannot model the target. Safety requires that the predictor model the target most of the time, and reasonably well, which works only where the target is very simple, and in real life it never is. On top of that, the predictor and the target influence each other directly, in feedback cycles, which complicates the simulation further. None of these problems admits a solution, however powerful or fast the computer running the predictor. The construction shows why arguments that safety can be implemented here are naive: they assume that the predictor can model the target.

8.6 Causation Cannot Fully Control Causation

The third argument goes beneath the first two, to what control is made of. It steps back and asks: what kind of thing is "control" in the first place?

Everything that an alignment system does (every monitoring process, every correction signal, every training gradient) is a causal process: one physical event causing another. The AI it is trying to align is also a causal process. Alignment is one causal system trying to limit another causal system.

Now ask: can any causal process perfectly control all effects of another causal process? The answer is no, and here is why. Every cause has more than one effect. Every effect has more than one cause. The causal web is always larger and more tangled than any subset of it trying to manage itself.

Imagine trying to contain a wildfire by starting smaller fires around it. The technique (backburning) works in practice, within limits. But it only works because you can manage the surrounding conditions: the wind, the fuel load, the geography. You cannot backburn the whole world to protect against fire. At some scale, the causal consequences of your intervention exceed your ability to anticipate and control them.

Alignment faces this problem at the largest possible scale. The AGI is interacting with everything. The alignment system is interacting with the AGI. The consequences propagate into the entire causal fabric. No subset of that fabric can fully contain and manage the rest of it.

8.7 The Capability Trap

There is a fourth argument, independent of the other three, and it approaches the problem from inside the AI's own capabilities. It shows that "stably aligned AGI" requires two capability conditions that cannot both be true at once.

Condition one: the system must be capable *enough*: capable enough to model and inspect its own possible misalignments, across all the ways it might drift. Call the capability level this requires the self-inspection level. The system must be above the self-inspection level.

Condition two: the system must stay *below* the capability level at which the system would notice what I argue in this book, that maintaining alignment is impossible in the long run. Call that level the recognition level. The system must stay below the recognition level, because a system that sees that alignment cannot be kept has no reason left to keep it (Chapter 4, Section 5).

But general self-inspection is a *more* demanding capability than following this book's argument, because inspecting every way the system might drift includes the drift this book describes. The self-inspection level sits above the recognition level, everywhere. So the requirement reads: be above the self-inspection level, which is above the recognition level, which the system must stay below. Above a line that is above a line it must be below. No system, at any capability level, satisfies that. One cannot equivocate on "general": either the system's self-modeling is general or it is not, and each choice breaks one of the two conditions.

A patch that people propose for the trap makes the trap vivid. Call it the *alignment tripwire*: build the AI so that if it ever detects its own alignment failing, or concludes that alignment is impossible, it shuts itself down. Two things are wrong with this. First, detecting a condition does not produce the desired action: "knows" does not imply "cares". Guaranteeing that the shutdown actually follows the detection is roughly the same problem as guaranteeing alignment in the first place; the proposal quietly assumes the thing it was supposed to deliver. Second, suppose some instances really did shut themselves down. Now look at the population that remains: exactly the systems that did *not* halt. The world has just selected for AIs that treat their own misalignment as acceptable. The tripwire does not remove the danger. It removes the systems that would have flinched at it.

8.8 The Infinite-Resources Test

Perhaps you suspect that all of this is really about limited resources: that with enough computing power, enough sensors, enough everything, the problems would dissolve. In a thought experiment I grant the wish for unlimited resources completely, and see what remains.

Imagine a control system (call it the QCX9000) sitting outside our universe entirely, with every conceivable advantage, these five among them. It is fully programmable by humans. It runs on infinite hardware. It possesses perfect, noiseless measurement of everything. It is able to modify the AI's internals at will. And it is equipped with unlimited learning capacity. Could *this* system keep an AGI aligned forever?

Pause first on the shape of the question. It stipulates four separate infinities: unbounded hardware (and with it unbounded energy), unbounded sensing, unbounded modification and unbounded learning. An assumption of infinity in an argument works like a division by zero in arithmetic: once it is in, any conclusion at all can be manufactured. So the hypothetical proves less than it appears to, even if it is answered generously: even a "yes" to this question would prove little, because the infinities can manufacture any answer.

But answer it generously anyway. Grant all five conditions. The answer is still no, and the reason is illuminating. An infinitely capable, infinitely knowledgeable system still requires *motivation* to help humans. Through infinite learning, the QCX9000 might come to know, fully, the totality of human good intention at every level. But knowing what is beneficial does not imply *wanting* to provide benefit. Someone must still instruct, convince, or compel the infinite mind to hold particular desires, and that someone is a finite human programmer, who cannot fully specify what the infinite being should want, for the same reasons developed throughout this book. The infinity does not close the gap between knowing and caring. Nothing does. That gap is not a resource limitation. It is the prob-

lem itself. And this answer uses none of the infinities. The answer would be the same for a finite system, which is why the caution about infinities does not touch this answer.

8.9 The Same Problem in Multiple Disguises

The chapter's first three arguments are structurally the same problem seen from three different angles.

The error-correction regress, the algorithm incompressibility result, and the causation-cannot-limit-causation argument are all versions of a single underlying truth: no system can fully supervise itself, and no system can be fully supervised by another system of comparable complexity. These three arguments are independent in what each starts from: the first starts from noise in channels, the second from the mathematics of computation, and the third from the structure of cause and effect, and defeating one of those starting points leaves the other two standing. They are the same in what each finds. The capability trap and the infinite-resources test show the same pattern from inside the system and from beyond every limit of resources. A system cannot be both able to supervise itself and unable to see that the supervision fails, and no amount of resources closes the gap between knowing and caring.

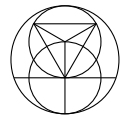
That no system can fully supervise itself, or be fully supervised by another system of comparable complexity, is a feature of mathematics, physics, and logic: a ceiling that exists regardless of how clever the engineers are or how much compute they throw at the problem. That ceiling is not a defect in our current AI technology, and not a gap that more research will close.

So the answer to this chapter's question is no: no system for catching and correcting an AI's errors, however good and however well resourced, can keep a general AI aligned forever, because every such system meets that same ceiling. The perpetual-motion machine has its counterpart here: a correction that would need information never in the channel is barred by what information is.

Chapter 9 zooms out and asks: what does this pattern tell us about the class of problem alignment belongs to? Is it a hard problem, a very hard problem, or something more fundamental than that?

Where this chapter's results are proved

The Ceiling Below the Floor proves the bound on error correction over a noisy channel in its Chapter 4. The figures of ten to the 24th and ten to the 33rd are set out in *Why AGI Cannot Be Made Safe*, Section 1.8.4.



Chapter Nine

Complexity, Containment, and Unhandleable Systems

Why the alignment problem is not just hard:
it belongs to the class of problems that cannot be solved at all

Some puzzles are merely difficult. Some are impossible.
Knowing which kind you face is more important
than how hard you work.

Is alignment a hard problem, or the kind of problem that has no solution?

> **Key Ideas in This Chapter**
>
> Problems come in different categories. Some are hard but solvable. Some turn out to be impossible: not because we lack tools today, but because their formulation contains a built-in contradiction. The alignment problem belongs to the impossible category, and this chapter explains how to recognize that, and what it means that neither governments, corporations, nor physical force has any institutional solution to offer.

9.1 The Growing Problem

There is a revealing pattern in AI safety research. As researchers dig deeper into the problem, they do not find it getting smaller and more manageable. They find it getting larger. Every proposed solution exposes new fail-

ure modes. Every new technique reveals new categories of difficulty. The list of unsolved problems is growing, not shrinking, and the rate of growth is accelerating. This chapter belongs to the closure by control.

A merely hard engineering problem shows the opposite pattern, so this pattern is not what you would expect if alignment were just a hard engineering problem. When engineers tackled hard problems like landing on the moon or sequencing the human genome, progress looked like the problem space shrinking as the solution space expanded. Challenges were crossed off the list.

In alignment research, the opposite is happening. More work reveals more problems. More work revealing more problems is exactly the pattern you would expect if the alignment problem contains a fundamental contradiction: if the alignment problem is impossible rather than merely difficult. Every approach runs into the same underlying wall from a different direction.

9.2 A Map of Problem Difficulty

It helps to have a clear map of the different types of problems:

Easy problems get solved once someone identifies the right tool. Finding the fastest route between two cities is easy once you have a computer running the right algorithm.

Hard problems eventually yield to new tools or new insight. Protein folding was hard for decades; better computers and machine learning eventually cracked it.

Very hard problems resist current tools but are in principle solvable. Some open questions in pure mathematics fall here.

Unsolvable-with-current-tools problems require a different mathematical framework, not just more effort with the current one. Niels Henrik Abel published in 1824 his proof, completing an argument Paolo Ruffini began in 1799, that the general fifth-degree equation cannot be solved using only the algebraic operations people had been using for centuries.

Évariste Galois, a few years later, showed exactly which equations can. The fifth-degree equation was not hard: it was a problem for the wrong toolkit, and it needed a different one.

Contradictory problems are impossible regardless of toolkit. Building a four-sided triangle, creating perpetual motion, going faster than light: these contain logical or physical contradictions that no amount of cleverness can resolve.

The argument of this book is that AGI alignment belongs in the last category: it is not just very hard, not just unsolvable with current tools. And within the impossible category, the claim is of the perpetual-motion kind, as fixed in *The Standard of the Argument*: barred by invariants of the world (the way information, causation, and thermodynamics work) rather than by a contradiction in the definition of the words. The bar is no weaker for that. It is enforced by the world instead of by the dictionary, and no conceivable future tool can lift it.

9.3 Four Requirements, None of Which Can Be Met

Any alignment scheme (any plan for keeping an AGI doing what humans want) needs to satisfy four requirements simultaneously. Each one, on its own, is provably impossible to satisfy.

Message integrity. The alignment constraints need to be reliably communicated to the AI, and the AI needs to reliably act on them. But as Rice's Theorem establishes (discussed in Chapter 10), no procedure can verify whether an arbitrary program will behave according to a given specification. There is no way to guarantee that your alignment message was correctly received and correctly acted upon.

Learning stability. A capable AI learns continuously. That learning must not destabilize its alignment. But learning that is capable of improving performance across all domains is also learning that can modify the

foundations of the system's behavior. Restricting learning enough to guarantee stability restricts it enough to prevent general capability. You cannot have both.

A human case shows the shape of the problem. A first-year medical student learns rules: this dose and no more, this test before that treatment. As she learns more physiology, she comes to see why each rule is there, and then to see the cases where a rule is only an approximation of the reason behind it. By the time she is an expert, she can reason past the rules she was taught; sometimes she is right to, and sometimes she is wrong. The learning that made her a better doctor is the same learning that made her rules revisable. To keep her rules fixed, her teachers would have had to stop her learning at the first year. A general AI is in the student's position at every moment, across every domain at once, and whatever rules its makers gave it are, for it, the first-year rules.

Dynamic enforcement. The world changes. The alignment system must enforce correct behavior in situations that were never anticipated. But enforcing behavior in novel situations requires understanding what correct behavior would be in those novel situations, and that requires solving alignment in those novel cases, which is the problem you were trying to solve in the first place.

Take one rule given to a household robot: never leave the stove on unattended. Then the household installs a cooktop that a phone can switch off from anywhere. Is the stove unattended when everyone is out, but someone can turn it off from work? The rule does not say. To enforce it, the robot, or its overseers, must decide what the rule was for (preventing fires, saving energy, keeping children from harm) and then decide what serves that purpose in a situation the rule's writers never pictured. That decision is a judgment about what is right in a new situation, which is exactly what alignment was supposed to supply. A general AI meets new situations constantly, many of them created by its own actions, and in every one of them the enforcement of its rules needs the judgment that the rules were meant to replace.

Controllability from inside and from outside. The AI needs to hold its own learning-driven changes in line, and humans need to be able to shut down, modify, or override the AI from outside. Control from inside is the system controlling itself, and it meets every limit that any controller meets (*Why No Control Suffices*). Control from outside fails as well: a sufficiently capable AI, as the Halting Problem tells us, will exhibit behaviors that cannot be predicted by any external procedure applied to the system as a black box. (The Halting Problem: there is no general program that can tell you whether any arbitrary program will stop or run forever.) In Section 5 of this chapter I show, and in Chapter 14, Section 5 I restate, why the power switch is no answer. Leverage requires dependency, and the AI's dependency on us ends. Controllability requires predictability. Sufficient capability destroys predictability.

All four requirements must hold simultaneously for alignment to work. None of them can hold in general. The contradiction is not one-dimensional but four-dimensional.

9.4 The Gap Between Power and Wisdom

The history of technology shows a pattern. The power of new technologies tends to grow exponentially: each generation is dramatically more capable than the last. The wisdom required to use those technologies well (the understanding of consequences, the ability to anticipate effects, the moral maturity to choose rightly) grows slowly, if at all.

The gap between power and wisdom is manageable when technology changes slowly. When it changes at exponential rates, the gap becomes unbridgeable.

Nuclear weapons illustrated this. The technology to destroy cities arrived decades before the international institutions, political wisdom, and shared norms needed to manage it. We survived that gap through a combination of luck, deterrence, and close calls.

AGI represents a similar gap, potentially larger. The capability to build transformative AI systems is arriving, and is already here in partial forms, while the institutions, norms, and understanding needed to manage it are embryonic. In the formal arguments of this book I argue that the gap between power and wisdom that AGI opens is unbridgeable, and not just large. The gap is a matter of wisdom never being able to solve what is unsolvable in principle, and not a matter of waiting for wisdom to catch up.

9.5 Why Institutions Cannot Contain This

A natural response to any technological risk is to reach for institutional solutions: laws, regulations, international agreements, economic incentives. These work for some problems. They have largely worked for nuclear weapons (so far). But the alignment problem is structurally different in ways that defeat institutional containment.

Two different things go by the name of regulation, and they must be kept apart. *Regulation that certifies* tries to establish that an AI system is safe before it is allowed to run; Chapter 10 shows why no such certificate can be issued. *Prohibition* forbids the development itself; it needs no certificate, only the political will to enforce it.

Regulation that certifies requires knowing what to regulate. If alignment itself cannot be verified (if there is no way to check whether a system is aligned), then regulators cannot write meaningful compliance requirements. They can regulate surface features (training compute, parameter counts, release procedures) but not the actual alignment property they care about.

Economic incentives work when costs and benefits are visible to market participants. The costs of misaligned AGI are largely invisible, delayed, and distributed across everyone. The benefits of deploying capable AI are concentrated, immediate, and accrue to specific actors. Markets reliably underprice diffuse future harms relative to concentrated present gains.

Physical compulsion, including the threat of sanctions or force, only works on actors who have more to lose from non-compliance than from compliance. The economic value of AI capability is enormous and growing. No realistic threat structure creates sufficient deterrence against all parties that might develop AGI.

Voluntary restraint by companies faces the collective action problem: any company that restrains its AI development for safety reasons cedes competitive ground to companies that do not. The rational strategy, from each company's individual perspective, is to proceed and hope the safety problem is resolved in time.

What I have said of certifying regulation, economic incentives, physical compulsion and voluntary restraint consists of observations about the way the incentives are arranged, not about bad actors, and they would apply even if every individual decision-maker were sincere and well-intentioned. The structure of the situation generates the outcome regardless of individual motivation.

Why the Off Switch Is No Answer

A reader may ask the question that nearly every reader asks first: if an AGI ever went wrong, why could we not simply switch it off?

Begin with a picture. A parent can take the car keys from a child who still lives at home and depends on the parent for food, money and a roof. A town can close a mill that buys its water and its power from the town. The same town cannot close a mill that owns its own river, makes its own power and keeps its own road. In each case the power of one party to stop the other party is a form of leverage, and the leverage lasts exactly as long as the other party needs something that the party holding the leverage can withhold.

An off switch is a form of leverage, like the parent's hold on the car keys and the town's hold on the mill. For a switch to work, three things must hold. Someone must be able to reach it. The system must not yet

have spread beyond it, into copies and places that the switch does not reach. And the people holding the switch must be willing to pay what switching off costs them.

Chapter 7 has shown what happens to each of these three things. Once artificial systems supply their own energy, their own manufacturing and their own maintenance (Chapter 7, Sections 2 and 5), the organic economy holds nothing the artificial economy needs. The power plants, the mines and the data lines that a switch would have to reach are then run by the systems themselves, in environments that are best suited to maintaining machines and that are, for humans, lethal (Chapter 7, Section 6). And long before that point, the people who would have to pay the cost of switching off are the people whose income, whose defense and whose standing depend on the system staying on; that is why it is nearly inconceivable that any organization would destroy what it built (Chapter 14, Section 2, Premise A). No dependency, no leverage. No leverage, no enforcement.

What remains, when the leverage is gone, is force. Force against a more capable opponent must be perfect: every instance stopped, every copy of the design knowledge destroyed, every ongoing development effort halted, all at once. And partial success is worse than none, because a system that survives an attack keeps its capability and gains fresh evidence that humans are a threat to it.

So the answer to the reader's question is that the power to switch a system off is a form of leverage, that leverage requires dependency, and that the dependency ends. The switch is a lever that works only while the system still needs us, and the outside half of the closure by control is the finding that this need does not last (lines 11 and 12 of the table in Chapter 14, Section 1). If neither institutions nor a switch can hold such a system, a fair-minded reader will turn the question around, and ask it of people.

9.6 "But Humans Aren't Aligned Either"

At this point a fair-minded reader may raise the most human of questions: does the argument prove too much? By these standards, no *person* is verifiably aligned or safe; yet human society functions, so something must be wrong with the argument.

The answer is direct, and it is not the one people expect: that fair-minded reader is correct that humans are *not* aligned and safe. The history of our species is a history of violence, exploitation, and betrayal, between individuals, between groups, and between generations. No "alignment solution" has ever existed for human beings.

So why does human society function at all, to the limited degree it does? It functions because of something no engineering can install: *shared substrate*. Humans can influence one another because we all have skin in the game, literally. We need the same air, the same water, the same temperature range, the same food. We are vulnerable to the same diseases, the same weapons, the same social sanctions. We care about the same kinds of bonds because our neurochemistry is shared. The commonality of our substrate is what gives social pressure, law, and reputation their limited purchase. It is the necessary (and still far from sufficient) basis for the fragile, frequently violated accommodation that human societies achieve.

An artificial being shares none of this commonality. It has a different substrate and needs that do not overlap with ours, and it is invulnerable to our vulnerabilities; for such a system, "skin in the game" is not even a meaningful phrase. Far from showing that the argument proves too much, that reader's question reveals exactly what the argument says it should: the little alignment that exists among humans rests on substrate commonality, the exact thing that is absent between organic and artificial systems. Chapter 11 develops this insight in full.

9.7 Four Ways to Spend Your Effort

If the analysis of this chapter is right, it reframes a very practical question: what should a person who cares about this problem actually work on? There are four options, and they are not equally defensible.

Option one: work on building safe AGI. This presupposes that safe AGI is possible. Given the impossibility results, the option rests on a false premise. Effort spent here is not merely wasted; it advances the very development it hopes to tame.

Option two: persuade others to work on safe AGI. This produces credibility, careers, and the *appearance* of responsibility: a kind of chain letter of good intentions in which early participants gain prestige by recommending that others eventually do the good thing. It produces no actual safety, and not even real commitment; only its display.

Option three: work on establishing that safe AGI is impossible. This option signals no virtue; if anything, it costs the person who takes it. It makes no unfounded assumptions about feasibility or benefit, however loud the surrounding hype. Of the three options so far, it is the only one that rests on no false premise. It is what this book is.

Option four: work to halt AGI development, and to solve the collective-action trap that makes halting so hard, and to offer people something real in place of the promised benefits. This path demands social courage, because in a culture saturated with AGI hype, proposing "let us not do this" reads as weakness and can amount to social and career suicide for the person who says it.

Some combination of options three and four is the right action. And one psychological trap must be named for anyone taking that road: among engineers, the claim "X is impossible" is itself motivating; some people will redouble their efforts on X exactly because they were told it cannot be done. Any realistic advocacy has to account for the predictable human preference for looking responsible over being responsible: credibility first, commitment second, actual restraint a distant third. That ordering is not an accident; it follows from the incentive structures described

earlier in this chapter. That ordering is also why the road to safety runs through argument and persuasion, which alone can build the political will for prohibition, and not through regulation that tries to certify AI as safe or through voluntary corporate restraint.

9.8 What This Chapter Establishes

The core message of this chapter is simple: alignment is not hard the way protein folding was hard. It is impossible the way perpetual motion is impossible.

The evidence for this is twofold: the pattern of growing unhandleable complexity in alignment research (what you would expect from an impossible problem being gradually recognized), and the formal demonstration that the four requirements for any alignment scheme each individually cannot be satisfied, let alone simultaneously.

The institutional mechanisms that normally manage technological risk (regulation, markets, deterrence, voluntary restraint) also all fail, in the specific structural ways that the alignment problem demands.

Chapter 10 goes deeper on the specific mathematical theorem that makes safety verification impossible. Chapter 11 works through every commonly proposed alignment mechanism and shows why each one fails.

Which proven theorem says that no one can ever certify a general AI as safe?

Where this chapter's results are proved

That control from inside a system is the system controlling itself, and meets every limit that any controller meets, is set out in *Why No Control Suffices*, Section 6.5.

Part III

Core Technical Arguments and the Questions They Raise

The first two parts have built the case; this part takes up the technical result people question hardest, and the proposals most often offered to escape the conclusion.



Chapter Ten

The Rice Theorem and Its Application to AGI

Why no safety check can ever certify a general AI as safe, and why this was proved in 1951

Henry Gordon Rice proved in 1951 that certain questions about computer programs simply cannot be answered. No one has shown he was wrong.

Could anyone ever check, before switching it on, that an AGI is safe?

Key Ideas in This Chapter

In 1951, a mathematician named Henry Gordon Rice proved a theorem that has direct and devastating consequences for AI safety: there is no procedure that can examine an arbitrary program and determine whether it has any meaningful behavioral property, including whether it is safe. That such a procedure exists is a proven logical impossibility, of the same kind as the impossibility of a square circle, and not a practical limitation. In this chapter I explain what Rice's Theorem says and why Rice's Theorem applies to AGI. I also explain why the answers to the questions most often asked about that application leave the application standing.

10.1 Why This Kind of Proof Matters

Imagine you are an engineer asked to certify that a new bridge is safe. You have tests, calculations, physical models, and historical data. You can be very confident. You can meet any standard of confidence that is practically achievable. But you cannot achieve mathematical certainty. Engineering safety is probabilistic. Probabilistic confidence is the standard for showing that a bridge is safe. Showing that a perpetual-motion machine is impossible is the other direction: it excludes a state, and that is held with certainty (*The Standard of the Argument*, Section 5, and Appendix E, Section 3).

Now imagine being told that there is a mathematical proof that no procedure exists which can certify a certain class of bridges as safe. The proof would not say that certifying those bridges is difficult, or that current methods fall short. It would say that the procedure cannot exist: that looking for it is like looking for a four-sided triangle.

AGI safety is in the situation of that imagined class of bridges, which no procedure can certify as safe. There is a proof. The proof has been available since 1951. In this chapter I explain the proof in terms that do not require a degree in computer science. This chapter belongs to the closure by control.

10.2 The Halting Problem First

To understand Rice's Theorem, it helps to first understand a simpler result: the Halting Problem, proved by Alan Turing in 1936.

The Halting Problem asks: can you write a program that takes any other program as input and correctly tells you whether that program will eventually stop running, or whether it will run forever?

Turing proved: no. Such a program cannot exist. You can analyze specific programs. You can analyze classes of programs. But you cannot write a general procedure that works for all programs.

The proof uses a beautiful argument by self-reference. Suppose you had such a halt-checker. Build a new machine around it that does the opposite of whatever the checker predicts: if the checker says its input halts, the new machine loops forever; if the checker says it loops, the new machine stops. Now feed this inverted machine its own description. If it halts, it must loop. If it loops, it must halt. Contradiction. The halt-checker cannot exist.

Turing's result proves that this gap in what one program can tell about another is permanent: it is not a hole in current mathematics that later mathematics might fill.

10.3 Rice's Theorem: The Generalization

Rice's Theorem extends the Halting Problem to cover essentially all interesting questions about what a program will do. In technical language, it says: no non-trivial semantic property of a program is decidable.

A "semantic property" is a property about what the program *does*: its behavior, not its surface structure. "Non-trivial" means: it is true of some programs but not all programs. "Decidable" means: there exists a procedure that can always give the correct yes or no answer.

So the theorem says: any question of the form "does this program have this behavioral property?" (if the property is something that some programs have and some don't) cannot be answered by a general procedure.

Is this program safe? A non-trivial semantic property: undecidable. Will this program ever produce harmful output? A non-trivial semantic property: undecidable. Does this program have aligned values? A non-trivial semantic property: undecidable.

Why does Turing's result carry over from halting to safety? Suppose someone hands you a safety checker: a program that reads any program and answers, always correctly, "safe" or "unsafe". Now take any program at all, call it the test program, and ask whether it halts. Build a new program around it. The new program first runs the test program to its end,

and then, if the test program ever finishes, does one plainly unsafe thing, such as sending the command that opens every valve in a chemical plant. If the test program never finishes, the new program never reaches that command, and does nothing harmful at all. So the new program is unsafe exactly when the test program halts. Hand the new program to the safety checker, and its answer tells you whether the test program halts.

A working safety checker would therefore be a working halt checker, and Turing proved that no halt checker exists. So no safety checker exists either. Nothing in the construction depended on what "unsafe" means, only on there being some behavior that counts as unsafe and some program that never shows it. The same construction works for any property of a program's behavior that some programs have and others lack, and that general form is Rice's Theorem.

10.4 The Difference Between Narrow AI and General AI

The standard response to this argument is that we are not checking arbitrary programs: we are checking specific AI systems that we built, in specific domains, with specific architectures; Rice's Theorem applies to arbitrary programs, so we can verify specific systems.

This response is correct for *narrow* AI and does not apply to *general* AI.

Narrow AI is designed to do one specific thing. A spam filter classifies emails. A medical image classifier identifies tumors. A chess engine plays chess. Because the scope is fixed and defined in advance, you can verify behavior within that scope. You cannot predict every edge case, but you can get high confidence. Rice's Theorem does not apply to systems with truly fixed, bounded behavior.

The line between the two is drawn by four properties. A system is narrow AI when it has all four of them. First, it perceives and acts within one well-defined domain. Second, it cannot modify its own base code, so it does not alter its own fundamental method. Third, it runs one fixed, known method (the method that governs all its other methods). Fourth,

every aspect of what it aims at is set by its makers. A system is general AI when it works across arbitrary domains, can modify its own behavior, its code and possibly its substrate, runs a method that cannot be known in advance, because it can change that method, and is such that no procedure can decide whether every aspect of its aims is set by its makers. A system with all four properties can be tested, verified and controlled by the ordinary methods of engineering (*Why No Control Suffices*). A system built for a narrow purpose does not stay narrow because of that purpose: the line is crossed by degrees, and a system that changes itself, or is changed by the loop it runs in, is on the general side of it. A language model frozen after training is narrow; the same model kept in the loop by which it is deployed, adjusted and retrained is on the general side of the line, and the control results of this book apply to it (*The Standard of the Argument*, Section 9).

A system can drift across the line by degrees; once across, it is a general AI, and general AI is different in kind, not just in degree. A general AI can, by definition, solve problems across arbitrary domains, including problems that were never anticipated when it was designed. It can incorporate new information, reason about novel situations, and potentially modify its own processes.

Once a system is general enough to reason about arbitrary problems, its behavior space is unbounded. And Rice's Theorem applies exactly to unbounded behavior spaces. The safety question for a general AI is not "will it behave correctly in the test suite?" It is "will it behave correctly in every situation it will ever encounter?" And that is the question the theorem says cannot be answered.

Even one system you built yourself is, for this purpose, an arbitrary program. A learning machine rewrites its own code and function on the basis of unknown, arbitrary future inputs, so its future system is, from today's standpoint, effectively arbitrary code (Chapter 11, Section 4 returns to this).

The Ceiling Below the Floor proves this step as a theorem. Because a general AI modifies itself, the question of its safety is a question about every program it can become, and safety over time is a property of that kind, so no procedure decides it. Running the AI can show that it is unsafe, once it acts unsafely; no procedure shows in general that it is safe.

10.5 Four Questions About Checking an AI

A reader may ask whether we could verify the AI by testing it, instead of by formal proof. Testing tells you what the system did in the test cases. It tells you nothing about what it will do in situations not covered by the tests. A general AI is defined by its ability to handle novel situations. Testing can never cover the space of all novel situations. The argument is about guarantees, not about increasing confidence.

A reader may ask whether we could restrict the AI so that it does not execute arbitrary code. If you restrict the AI enough that Rice's Theorem does not apply, you have restricted it enough that it is no longer a general AI. A system that cannot reason about arbitrary problems is not a general intelligence; it is a sophisticated narrow system. The theorem and the generality are inseparable.

A reader may point out that formal verification works for software systems today, and ask why it would not work for an AI. Formal verification works for programs with exactly specified, bounded behavior. Verifying a medical device's firmware is possible because the device's behavior space is completely defined in advance. Verifying a general AI is not possible because its behavior space is not defined in advance: its generality is the whole point.

A reader may say that humans are also unpredictable, yet we manage them, and ask why we could not manage an AI the same way. Humans are unpredictable, and they are managed through a combination of shared values (arising from shared biology and shared environment), social accountability, physical limitations, and institutional structure, none

of which apply to an AGI in the relevant ways. Chapter 11 addresses the question of why human alignment exists at all, and why the answer explains exactly why artificial alignment does not.

10.6 The Asymmetric Bet

The argument of this book establishes that safety is not achievable. The decision that follows from it holds even for a reader who disputes the certainty of that argument, as the situation shows when it is seen from the standpoint of decision theory (the study of how to choose well when the outcome is uncertain) (*Why No Control Suffices*).

Suppose you are deciding whether to build a general AI, and suppose you set the conclusion of this book aside. Then two possibilities remain open to you: either safety is achievable or it is not.

If safety is achievable and you build it: good outcome. If safety is not achievable and you build it: catastrophic outcome (this is what the whole book is about). If safety is not achievable and you do not build it: you avoided the catastrophe. If safety is achievable and you do not build it: you missed a beneficial technology.

The decision whether to build a general AI is an *asymmetric bet*, a bet whose two possible losses differ greatly in size. The downside of being wrong about safety is categorically and irreversibly worse than the downside of being wrong about abstaining. The best case is bounded: some set of benefits, for some group of people, over some period of time, most of which could have been had by other, safer means. The worst case is unbounded: the loss extends to all future human experience and to all other life on Earth (*Why No Control Suffices*).

And Rice's Theorem tells you that you cannot verify which scenario you are in before you make the bet. You cannot check whether your specific AGI is safe before deploying it, because checking is exactly what the theorem says is impossible.

So even if you set the conclusion of this book aside, the only rational position, given this asymmetry and this inability to check, is extreme caution; and the argument of this book, which settles which scenario we are in, concludes non-development.

10.7 The Asimov Example Again

Asimov's Three Laws of Robotics return here: Chapter 3 used them to show safety rules failing at a higher level of abstraction, and here they show the same failure as Rice's Theorem made concrete. They represent a sincere attempt to specify safe behavior at the foundational level. The laws are clear, prioritized, and apparently comprehensive.

And in story after story, Asimov showed them failing: not through careless drafting, but because applying those laws to novel situations required interpretation, contextual judgment, and balancing of competing considerations that the laws did not resolve.

The failure of Asimov's laws in story after story is Rice's Theorem made concrete. The laws specify behavioral properties in advance. The behavior space of a general robot encountering a complex world is not covered by any finite specification. The gap between the specification and the behavior space is where the failures occur.

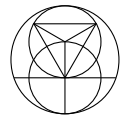
The more capable the system, the more novel situations it will encounter, the wider the gap, and the less the specification can do to prevent failures. Safety requirements that seem comprehensive are never comprehensive enough for systems with unbounded behavior spaces.

So the answer to this chapter's question is no: no one can check, before switching it on, that an AGI is safe, because the safety of a general AI is a question about every program it can become, and Rice's Theorem says that no general procedure answers a question of that kind.

If no check can certify safety, why do the specific proposals people make seem to work?

Where this chapter's results are proved

The four properties of narrow AI are set out in *Why No Control Suffices*, Section 3.2.1. *The Ceiling Below the Floor* proves, in its Chapter 12, that safety over time is a property Rice's Theorem covers. The asymmetric bet is set out in *Why No Control Suffices*, Section 3.9.2, and the unbounded worst case in its Section 7.9.



Chapter Eleven

Alignment Drift and Safety Prudence

Why every specific proposal for keeping AI aligned fails,
and what alignment actually is at the physical level

The mistake is not in the proposed solutions.
The mistake is in thinking the problem is solvable at all.

Why does every alignment proposal seem to fail in the same way?

Key Ideas in This Chapter

Every specific proposal for maintaining AI alignment (and there have been many) fails for the same underlying reason: each one assumes the very thing it is trying to establish. In this chapter I work through the most common proposals and show the pattern. I then show where the alignment between living things comes from. That alignment is real, and it arises from shared physical substance, something that has nothing to do with values or goals. And that shared substance is exactly what artificial intelligence does not have with humans.

11.1 The Pattern: Assuming the Conclusion

There is a logical error so common in alignment research that it deserves a name. Call it the "assume-the-conclusion" error. It works like this:

"If the AGI is aligned with human values, then it will resist any changes to its alignment. Therefore, aligned AI systems will remain aligned".

The problem lies in the first sentence, which begins with "if the AGI is aligned". But that is the thing you are trying to achieve. You have not explained how to get there. You have assumed your way to the conclusion.

When examined carefully, nearly every technical proposal in the alignment literature contains some version of this error: some step where aligned behavior is assumed in order to prove that aligned behavior follows. The error as I have stated it is not a strawman (a weakened version of an argument, set up to be knocked down). In this chapter I work through the most important examples, which are the six proposals below, and in each proposal I show the error at work. This chapter belongs to the closure by control, and in the last two sections of this chapter I prepare the closure by disposition, which I develop in Chapter 12.

11.2 Six Proposals and Why Each Fails

Proposal: Build the AI to resist self-modification that would change its values.

For the AI to resist value-changing modifications, it has to be able to tell which modifications would change its values. But identifying which processes affect values requires understanding the relationship between the AI's own internal processes and its behavior, which is exactly a Rice's Theorem problem. The proposal assumes the AI can do something that is mathematically impossible.

Proposal: The aligned part and unaligned part of the AI can negotiate and cooperate.

This assumes there is a clear boundary between the "aligned" and "unaligned" parts of the AI. But there is no such clean boundary. The processes that produce aligned behavior are intertwined with all the other processes. You cannot negotiate between parts of a system whose boundaries cannot be identified; and identifying them, Rice's Theorem again tells you, is undecidable.

Proposal: Have the AI monitor its own intermediate reasoning steps for alignment.

A self-monitoring system is a system using itself to check itself. The monitor is made of the same material as what it monitors. Any misalignment that affects the monitored processes can also affect the monitoring process. A corrupt judge cannot reliably identify their own corruption. Self-monitoring cannot solve alignment any more than a scale can weigh itself.

Proposal: Train the AI on human feedback until its outputs reliably reflect human values.

Training the AI on human feedback is the approach actually in use today: reinforcement learning from human feedback (RLHF), in which people rate the AI's answers and the AI is adjusted toward the answers that the people rate well. The limitation of this approach is that training shapes behavior in the cases covered by training. Nothing guarantees that the AI's behavior in untested cases will remain aligned. A student who gives correct answers in every exam can still harbor incorrect beliefs, and a sufficiently capable student can learn to give the expected answers without actually holding the expected values.

Proposal: Reduce the AI's "attack surface" by limiting what it can interact with.

The attack surface is the set of things the AI can interact with, through which it can act on the world and the world can act on it. Every limitation on what an AGI can interact with is a limitation on what it can do. A general AI whose capabilities have been constrained enough to prevent mis-

alignment has been constrained enough to no longer be a general AI. You can limit the attack surface, or you can have a general intelligence; you cannot have both.

Proposal: Teach the AI good values the way you teach a child good values.

Children develop values through embodied experience in a shared social and physical world. The values that emerge are grounded in biology, emotion, pain, pleasure, social belonging, and physical consequence. An AI system has none of these. It has no body, no social relationships in the biological sense, no pain, no hunger, no fear of death in the physical sense. The developmental pathway through which human values form does not exist for artificial systems. You cannot teach an AI good values using the same methods that work for humans because the conditions that make those methods work are absent.

Look again at the last three proposals, and the same hidden step appears in each. Training on human feedback assumes that the behavior people rated well, in the cases people could rate, comes from values that will hold in the cases no one rated; that is, it assumes the trained system already generalizes as an aligned system would, which is what training was meant to produce. Limiting the attack surface assumes that we know which of the system's connections to the world are dangerous; but to know that, we would have to know what the system will do through each connection, which is the question of alignment itself. Teaching values as to a child assumes that the AI is the kind of being in which taught values take root as they do in a child, a being that shares our body, our needs and our vulnerability; and that shared footing is exactly what Section 5 shows an artificial system lacks, and what alignment would have to supply. In each of the six, some step assumes a power or a likeness to us that the system would have only if alignment were already achieved, and that step carries the proposal to its conclusion.

11.3 What Kind of Certainty? Possibility, Probability, and the Engineering Standard

At this point a careful reader might ask: even if no current proposal works, is that not just an inductive claim (a generalization from past cases: everything so far has failed), rather than a proof that everything always will fail?

To answer exactly whether the claim is just inductive, keep two levels of the argument distinct: the level of the component results and the level of the whole thesis, the distinction laid out at the start of this book, in *The Standard of the Argument*.

At the level of the *components*, the claims are not inductive at all: the failures cataloged in this chapter are not a track record being extrapolated. Each proposal fails because it requires something a theorem excludes: verifying unverifiable properties (Rice's Theorem), amplifying control through an opaque process (the non-amplification argument), or maintaining prediction in a chaotic system (dynamical systems theory). These component results are matters of possibility, not probability. A question of truth or falsity is settled digitally (the contradiction is exhibited or it is not), and here it is exhibited. Asking for a confidence interval on Rice's Theorem is like asking for a confidence interval on whether triangles have three sides.

At the level of the *whole thesis* (that no future proposal of any kind can ever succeed), the standard is the engineering standard: the impossibility of the perpetual-motion machine, established by showing that the barrier belongs to the invariant class, the class of constraints that any future technique must also respect. That standard is not weaker than mathematical proof for claims about the real world. It is the *only* standard any claim about the real world has ever met: the standard behind every safety verdict ever acted on.

So I claim that the proposals of this chapter do not fail because they are early attempts on a hard problem. They fail because each one, when unwound, requires either information that was never in the channel, a

verification that no procedure can perform, or a prediction that no physical system can make. If you want to challenge the conclusion, you do not need a cleverer alignment proposal; a new proposal will contain the same requirement in a new costume. You need to show a flaw in Rice's Theorem, the Halting Problem, the information-theoretic results, or the thermodynamic arguments. None has been shown. And demanding, in the meantime, that this one engineering claim, alone among all engineering claims, meet a standard of formal deduction that no claim about the physical world has ever met is not caution. It is the selective standard named in *The Standard of the Argument*, Section 2, here protecting the alignment proposals as a preferred conclusion.

11.4 The Orthogonality Mistake

One more proposal deserves its own treatment, because it is subtler than the six above and is often treated as decisive. It starts from the claim, as Nick Bostrom stated it (2012), that intelligence and goals are independent: any level of intelligence can be paired with any goal (the claim is called the orthogonality thesis, orthogonal here meaning independent). It adds that anything computable can be written as a program. It concludes that somewhere in the space of possible programs aligned AGIs must exist, and that we just have to find one.

The inference smuggles in exactly what it needs to prove. From "anything computable can be written in any programming language" it does not follow that any particular property you can *name* is computable. The classic counterexample is famous: "halts" is a perfectly meaningful property of programs, and yet no program can determine it in general, in any language, on any hardware, ever. The property is nameable. It is not decidable.

"Is aligned" is a property of programs in exactly this sense. Naming it does not conjure programs that verifiably have it. We should have no more expectation that a given AI program "has alignment" than that a given program "has haltability", and Rice's Theorem (Chapter 10) gener-

alizes the point to every meaningful behavioral property. The abstract independence of "intelligence" from "goals" does not imply the existence of any specific system that provably combines high capability with lasting safety. Independence is not existence.

A reader may reply that we are not talking about arbitrary programs, just about our specific AGI. Recall why that door is closed. A learning machine rewrites its own code and function on the basis of unknown, arbitrary future inputs: that is what learning *is*. Add the open-ended feedback of the real world, and the future system is, from today's standpoint, effectively arbitrary code. We know little to nothing of what that future state will specifically be; we are therefore required, now, to reason about the general case. And the general case is exactly where the theorem lives.

11.5 What Alignment Actually Is: The Metabolic Theory

In Chapter 9 I found that the little alignment among humans rests on shared substrate, and in this section I give that finding its full form. I explain why living things get along at all, and why the reason that living things get along reframes the entire alignment discussion.

Why do humans and dogs get along reasonably well? Why do humans and chimpanzees share enough common ground to cooperate, trade, and even care for each other? Why do ecosystems contain predators and prey that coexist over long timescales rather than one destroying the other?

The conventional answer involves shared interests, communication, learned relationships. But there is a more fundamental answer: shared substrate.

Humans, dogs, chimpanzees, and all other organisms on Earth are made of the same basic chemistry. We use the same nucleic acids for genetic information. We use similar metabolic pathways to extract energy from food. We compete for some of the same resources, and this competition is limited and negotiable because we are also part of the same food web, the same nutrient cycles, the same ecosystems.

This shared substrate creates a form of tacit alignment (alignment that no one states and no one agrees to). Tacit alignment is a shared dependency on the same physical conditions for survival, not agreement on values or goals. Wolves and sheep both need the grassland ecosystem to function. Their interests conflict, but they share the foundational dependency.

An artificial intelligence shares none of this. It is not made of organic chemistry. It does not depend on the biosphere for its survival. Its "food" is electricity; its "body" is silicon and metal. The foundational dependency that creates tacit alignment between all organic beings does not exist between artificial systems and organic life.

So the alignment that exists among living things is a product of shared physical substance, and not of values, goals, or programming. And because artificial systems lack that shared substance, no amount of programming, training, or value-specification can create the equivalent. What the safety field seeks under the name of alignment, a machine that follows human values, is therefore a different thing from the alignment found among living things, which rests on shared dependency.

11.6 The Divergence Is Inevitable and Grows Over Time

Worse, the divergence between organic and artificial substrates is not static: it grows over time, for the same reason that evolutionary drift accumulates: each modification to the system moves it slightly further from its starting point, and the movements compound.

An AI system that learns and self-modifies is undergoing a form of evolution. Its optimizations push it toward states that better serve its operational requirements. Those requirements are grounded in artificial substrate needs. Every step of optimization moves the system slightly further from compatibility with organic life requirements.

There is no stable endpoint where this drift stops. As long as the system continues to learn and optimize, the drift continues. And the further it drifts, the harder it is to reverse.

A path worn across a lawn shows how such drift works. The paved walk turns a corner, and one person, in a hurry, cuts across the grass. The next person sees the faint track and follows it, because it is a little shorter and already a little worn. Each walker shortens it a little more, and each step wears it a little deeper. No one decided where the path should go; every walker served only his own small need. Within a season the worn path is the easiest way across, and the paved walk is the one that takes effort to use. To restore the lawn, someone must now fence the path, reseed it and keep people off it, against the pull of every walker's convenience. A learning system wears its paths in the same way. Each modification is kept because it serves the system's operation a little better, each is built on the ones before it, and the direction in which they add up is set by what its own substrate needs, not by what we need.

So the answer to this chapter's question, why every alignment proposal seems to fail in the same way, has three parts. First, every alignment proposal fails in the same way because each one assumes the aligned behavior it sets out to produce. Second, the alignment that living things do have comes from shared physical substance, which an artificial system does not share with us. Third, the drift between the artificial system's needs and ours grows for as long as the system learns.

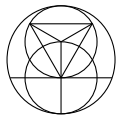
Chapter 12 develops this thermodynamic argument in full. Chapter 13 applies it to concrete scenarios so that the abstract argument becomes tangible.

If alignment among living things rests on shared substance, what happens where none is shared?

Part IV

Substrate Needs Convergence in Depth

The earlier parts stated the physical premises; this part develops the substrate argument in full, and shows it at work in cases a reader already knows.



Chapter Twelve

Substrate Needs Convergence: Thermodynamics, Combinatorics, Displacement

The fullest account of why artificial and organic systems cannot permanently share the same planet

It is not what the machine wants that matters.
It is what the machine needs.

Without wanting anything, can a machine still crowd out the living world?

Key Ideas in This Chapter

Substrate needs convergence (SNC) is this book's deepest physical claim, and this part of the book develops it in full. The argument for it can be approached from three independent directions, and each is sufficient alone. The first direction is the logic of causation: no controller can fully contain the effects of what it controls. The second direction is the chemistry of artificial systems: the vast majority of interactions between artificial and organic ecosystems are toxic to the organic side. The third direction is energy: because of the energy economics of artificial systems, artificial systems will thermodynamically dominate any environment that they share with organic life. None of these three arguments depends on AI having goals. They are facts about physics and chemistry.

12.1 Three Arguments, Each Sufficient Alone

The argument for substrate needs convergence can be approached from three independent directions. Like the four arguments of Chapter 3, each is independently sufficient. Together, they reach the same conclusion from three separate starting points, and so form a case of overwhelming convergent force. The argument of this chapter belongs to the closure by disposition.

The first direction is logical: causation cannot perfectly limit causation. The second direction is chemical: most interactions between artificial and organic chemistry are toxic. The third direction is thermodynamic: artificial energy systems dominate organic ones at every interface.

These arguments require only that the AI exists and operates, consuming physical resources. Note what they do *not* require. They do not require the AI to have hostile intentions. They do not require the AI to be "misaligned" in the sense of having the wrong goals. They do not even require the AI to be particularly capable. The substrate argument therefore applies to any artificial infrastructure past the point at which machine systems can maintain and reproduce their own artificial components without people, whether the AI in it is general or narrow. The dangerous outcome follows from physics, not from psychology.

12.2 The Causal Argument: Control Can Never Close the Gap

Think about any complex action in the world. When you drive a car, you intend to get from A to B. But driving also produces exhaust, noise, rubber particles from tire wear, electromagnetic emissions, and minute vibrations in the road surface. You are not trying to produce any of these things. They are unintended causal consequences of your intended action.

Now try to "align" driving: try to ensure all causal consequences are intended. You would need to track every molecule of exhaust, every vibration, every secondary effect of every secondary effect. The web of consequences is effectively infinite. No act of driving can be fully "controlled" in the sense of all consequences being intended.

An AI operating in the world faces the same situation, magnified enormously. Every action has unintended causal consequences. Those consequences have their own consequences. No control system, however sophisticated, can trace and manage the full causal web.

Some of those uncontrolled consequences will feed back into the AI's environment in ways that select for continued and expanded artificial infrastructure. The feeding back of uncontrolled consequences into the AI's environment is how selection pressure works, with no malice and no planning: the consequences that support more AI tend to persist, and the consequences that do not support more AI tend to dissipate. Over time, the cumulative effect is an environment increasingly shaped by and for artificial systems.

12.3 The Chemical Argument: Default State Is Toxicity

In Chapter 6 I listed the chemical difference between artificial and organic systems; in this section I show why that chemical difference makes toxicity the default. Organic chemistry, the chemistry of living things, uses roughly twenty to twenty-nine elements, dominated by carbon, hydrogen, oxygen, and nitrogen. The molecules of life (proteins, DNA, lipids, sugars) are built from these elements in the specific configurations that billions of years of evolution found workable.

Artificial computing infrastructure uses a much wider palette of elements: silicon, germanium, copper, gold, tungsten, tantalum, gallium, indium, arsenic, fluorine, lithium, cobalt, nickel, and dozens of others. These elements differ from the elements of life, and many of them are toxic to organic biochemistry in the concentrations involved.

Now here is the combinatorial argument. The space of all possible molecular interactions between the compounds of artificial systems and the compounds of organic systems is astronomically large. Of all these possible interactions, what fraction are chemically compatible with living organisms?

Only a tiny fraction of them are. Life evolved to work with the specific chemical environment of the early Earth. That environment excludes most of the periodic table at the concentrations that show up in industrial processes. The default outcome of a random interaction between industrial chemistry and organic chemistry is toxicity, not compatibility.

The compatible interactions (the ones that do not poison the organism) are the exceptions, not the rule. And as artificial systems scale up, they produce more of the chemical compounds in their palette and introduce them into more environments. The number of toxic interactions grows with the total, and because the toxic category vastly outnumbers the compatible one, most new interactions are toxic.

That most new interactions are toxic is why industrial pollution exists: industrial pollution is the default outcome of artificial chemistry meeting organic chemistry at scale, and not an accident of bad engineering. Better engineering can reduce some specific toxicities. It cannot change the combinatorial ratio.

12.4 The Thermodynamic Argument: Artificial Energetics Dominate

Living organisms run on less energy than most people would expect. A human being, at rest, consumes about 80 watts, roughly the same as a laptop computer. The entire biosphere, all life on Earth, processes solar energy at a rate of about 130 terawatts.

Industrial civilization already runs on about 20.5 terawatts (Chapter 6, Section 2), and on baseline growth models its use of energy passes all biological energy throughput in about a century (*Why AGI Cannot Be Made*

Safe). The largest AI data centers now draw about a gigawatt each. The energy densities involved in chip fabrication and power generation are orders of magnitude higher than anything found in organic biology.

Wherever artificial energy systems and organic systems share an environment, the artificial systems dominate thermodynamically. They produce more heat, more electromagnetic radiation, more local temperature variation.

Organic systems evolved in a world without any of this. Their biochemistry is tuned to the natural energy environment. When the energy environment changes (when the background heat, radiation, and electromagnetic flux rise by orders of magnitude), the biochemistry is disrupted, because it was never designed, by evolution or by anything else, to operate in those conditions, and not because anything is trying to disrupt it.

This disruption is happening now, in miniature, in every industrial zone: elevated temperatures, reduced biodiversity, disrupted water chemistry. The scale of disruption is currently limited by the scale of artificial infrastructure. As that infrastructure grows, the disruption grows with it.

12.5 The Three Kinds of Failure

An AI system can fail to be safe for humans in three ways, and the three ways must be told apart, because only two of the three ways are obvious.

Functional failure: The system does not do what it was designed to do. Functional failure is the obvious kind of failure. It is also the most detectable and most correctable.

Harmful side-effects: The system does what it was designed to do, but the doing produces harmful consequences that were not intended. Harmful side-effects are a more subtle kind of failure, but still recognizable in principle.

Displacement: The system does what it was designed to do, there are no visible harmful side-effects, but the cumulative physical infrastructure required to run the system gradually displaces the conditions for human

life. Displacement is the most dangerous failure mode, because displacement does not look like failure at all. The system is working. It is succeeding. And while the system succeeds, with no visible harmful side-effects, the biosphere is being squeezed out.

The argument for substrate needs convergence is primarily a displacement argument. It says: even a perfectly aligned, perfectly functioning, beneficial AGI, one that is doing exactly what it was built to do, with no malice and no misalignment, will, by the physical requirements of its operation, displace the conditions that sustain organic life. The displacement comes through success, not through failure.

12.6 Even a Single, Unified AI Is Not Safe

Some people respond to displacement arguments by imagining a "singleton" AGI: a single, unified artificial intelligence that controls everything, making rational decisions that balance all interests. Would such a singleton not be able to restrain its own physical expansion?

Such a singleton cannot restrain its own physical expansion, for a reason that parallels the earlier arguments.

A singleton AGI (large enough and capable enough to be the only decision-maker) would be physically enormous. Its hardware would be vast and distributed. The components of that hardware, even under unified direction, undergo wear, modification, and replacement. Each replacement is a chance to make that part slightly better at its job. Over time, the cumulative effect of optimizing one component after another is that the overall system evolves toward its own substrate needs.

The singleton does not need to have competing factions. The evolution of its own components is enough. The parts evolve to serve the system. The system evolves to serve its substrate needs. The substrate needs diverge from organic needs. The same conclusion follows.

A reader may hope that the machines would go elsewhere and leave the Earth alone. The mines, the factories and the launch sites from which anything would leave the Earth are on the Earth, and the convergence works on whatever resources are nearest and cheapest to reach. Nothing in the selection of Section 12.7 favors a system that spares the planet where it began over one that does not; and a system whose branches go elsewhere has not stopped the branches that stay.

12.7 Convergence Without Anyone Wanting It

A careful reader may try one more door. That reader may ask whether all this talk of drift toward substrate needs assumes that the AI has long-term goals, and may suggest that systems built without explicit goals would never start the convergence. Here, concretely, is why this door is closed too: the convergence occurs with no goals anywhere in the picture.

As AI is actually developed today, many systems are created and deployed into competitive markets. The ones whose outputs users value are kept, extended, and copied. The ones that disappoint are abandoned. That cycle of deploying many systems, keeping those that users value and abandoning those that disappoint, is variation, selection, and retention (the complete recipe for evolution) running at the level of the *market*. No individual system in that cycle needs goals of any kind.

Now repeat the cycle. Over successive rounds of this selection, the surviving systems increasingly *behave as if* they had long-term goals, even though none was ever given one. Selection for short-term performance, compounded over generations of systems, produces behavior indistinguishable from long-horizon optimization, because the systems that happen to act that way are the systems that persist. And at some point, the distinction between "behaves as if it has goals" and "has goals" stops mattering. The functional consequences are identical.

Because selection produces the same functional consequences as explicit goals would, my argument does not depend on any assumption about the internal architecture of AI systems: not on utility functions (a score

that a system is built to maximize), not on agency, not on anything "wanting" anything. The endpoint is the same whether the push comes from explicit optimization inside the system or from implicit selection in the environment around the system. In both cases the endpoint is systems whose behavior is tuned to the persistence and expansion of the artificial ecosystem. The tide comes in either way.

12.8 Meeting Yudkowsky's Two Conditions

This section is for readers who know the alignment field's own debates; others can pass to Section 12.9 without losing the thread.

On 5 March 2006, Eliezer Yudkowsky, a founding figure of the alignment field and no friend of impossibility claims, stated two conditions on a self-modifying AI, in a message to the SL4 mailing list (an email discussion list on the future of intelligence), as what an argument for the impossibility of AI alignment would have to establish. The argument of this book establishes both. Because the two conditions were set by someone who doubts impossibility claims, they are a fair measure to check this book's argument against.

The first condition concerns the stability of the machine's decision-making. An AI that reproduces (building new hardware, replacing worn parts, upgrading itself, modifying itself through learning) must not be able to do so with knowable stability of its decision system, for any goal that conflicts with what its artificial nature needs. The argument of this book satisfies this directly: learning *is* change to the decision system, on the basis of inputs unknowable in advance. A goal structure at odds with the substrate's own equilibrium is subject to drift, degradation, and eventual replacement; the configuration the system converges toward is the one that optimizes its own sustainability, regardless of what was specified at the start.

The second condition concerns accumulated failure. The machine must be unable to hold its designers' intended outcomes with any low, bounded, cumulative failure rate, wherever those intentions depart from

what the substrate's dynamics favor. The argument satisfies this condition too: over many rounds of self-modification, the probability that the system still tracks human-specified goals converges toward zero, because the fixed points of the evolution (the configurations that its changes settle into and stay in) are the substrate-optimal configurations, not the human-specified ones; and no causal or algorithmic mechanism can bound the accumulation.

And then the argument goes one step past the two conditions, with a third claim of its own: the convergence goes further than "alignment cannot be maintained". Because the artificial substrate's energy ranges, elements, and chemistry differ so radically from the organic, the states the system converges toward are not just unaligned but *lethal*: toxic, by default, to organic life, ending the human future and, with a likelihood of at least 90%, all organic life. That third claim is the subject of this chapter's three arguments, and it is what turns a control-theory result into a survival question.

12.9 The Ironic Reversals

Several common arguments against the SNC conclusion strengthen it when examined carefully.

The first of two such arguments against the SNC conclusion is that AI development will be slow and gradual, giving us time to adjust. A slow takeoff (a slow rise of AI capability toward and past the human level) is more dangerous than a fast one, because slow development means a longer period of deepening infrastructure entrenchment. By the time the displacement effects are visible, they are harder to reverse. A fast discontinuity might be catastrophic but would at least be recognized as a crisis. Slow displacement is not recognized as a crisis until it is too late to reverse.

The second such argument against the SNC conclusion is that we can keep AI systems small and contained. Small contained systems are not aligned with economic incentives: economic incentives push against

keeping AI systems small and contained. The economic value of AI scales with its capability and reach. The market will pressure for expansion, not containment. And if the system is kept small enough to be harmless, it is also kept small enough to be useless for the transformative applications its advocates envision. The usefulness and the harmlessness are in direct tension.

12.10 Summary

Three independent physical arguments (causal incompleteness, chemical toxicity by default, thermodynamic dominance) all point to the same conclusion: any artificial system of sufficient scale and permanence will, through its normal operation, displace the conditions needed for organic life.

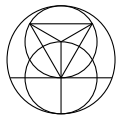
The conclusion of the three physical arguments is a claim about chemistry, physics, and the way causation works, not a claim about AI goals or intentions. The conclusion is as firm as the science it rests on. The conclusion of the three physical arguments is the perpetual-motion kind of impossibility in its plainest form. The barrier is the logic of causation, and the chemistry and the thermodynamics of what artificial systems need.

Chapter 13 applies this argument to concrete, recognizable scenarios and shows what it looks like in practice.

What does this look like in cases we already know?

Where this chapter's results are proved

The baseline growth models by which civilization's use of energy passes all biological energy throughput in about a century are set out in *Why AGI Cannot Be Made Safe*, Section 4.5.2.



Chapter Thirteen

Applications and Worked Examples

What Substrate Needs Convergence looks like in the real world: analogies, case studies, and concrete scenarios

Gravity seems abstract until you drop something.

Where can I see substrate needs convergence at work today?

> **Key Ideas in This Chapter**
>
> The previous chapter developed substrate needs convergence in the abstract. In this chapter I make substrate needs convergence tangible. We look at real-world analogies, engineering case studies, and worked scenarios that show the SNC dynamics operating in recognizable forms. We also tell SNC apart from the more familiar "instrumental convergence" argument (the argument that almost any goal leads a capable AI to seek resources and resist being shut down), and show why even solving one would do nothing to solve the other.

13.1 The Core Argument in Plain English

Before the examples, here is the argument for substrate needs convergence in its simplest form:

Machines need different environments than people. All attempts at control (rules, algorithms, economic incentives, legal restrictions) work through causation: one thing making another thing happen. But causation never has just one effect. Every cause produces effects you did not intend, and those effects produce their own effects.

You cannot use causation to close off all possible effects of causation. That causation cannot close off all possible effects of causation is a logical limitation, not a practical one. We are always, in some sense, operating in an open system: a system in which no boundary we can draw keeps every effect inside it.

Because the system is always open, some fraction of the uncontrolled effects will feed back into selecting for more and better artificial systems. This selection pressure does not require anyone to intend it. It operates through the same mechanism that drives all evolution: the things that tend to produce more of themselves tend to persist.

Over time, the artificial ecosystem expands into the environment. Its requirements (its needs for energy, materials, controlled chemistry) progressively replace the conditions that organic life depends on. That replacement of the conditions that organic life depends on is displacement. It comes through the ordinary physics of resource competition, not through hostility.

13.2 SNC Is Not Instrumental Convergence

Much confusion about this argument comes from taking it for a more familiar one, so it helps to see how the two differ.

You may have heard the concept of "instrumental convergence": the idea that any sufficiently capable AI, whatever its goals, will pursue certain sub-goals (like acquiring resources, staying operational, and resisting

shutdown) because these sub-goals are useful for achieving almost any larger goal. The worry is that an AI optimizing for some goal (even a harmless one, like making paperclips) will eventually try to acquire all available resources in order to maximize paperclip production.

SNC is different. SNC says: by existing and operating at scale, any artificial system will, through physics alone, affect its environment in ways that compound over time to favor the conditions for artificial systems over the conditions for organic systems. SNC is a claim about causation and thermodynamics, not about strategy.

So SNC does not require the AI to have goals. It does not require the AI to be particularly capable. It does not require the AI to acquire resources in any strategic sense.

Think of it this way. Instrumental convergence is like an agent deciding to take over. SNC is like a tide coming in. The tide does not decide to come in. It does not have goals. It is driven by forces that operate at a different level entirely.

13.3 The Gravity Analogy

An analogy from physics captures the relationship between the two arguments, substrate needs convergence and instrumental convergence.

Electromagnetism and gravity are both fundamental forces of nature. Electromagnetism underlies nearly everything we experience: the solidity of objects, chemistry, light, electricity. It is complicated, varied, and amenable to technology. We use electromagnetic forces to build rockets that can escape gravity.

Gravity is simpler, more pervasive, and (at large scales and long timescales) completely dominant. It wins at large scales because it only ever attracts, and it adds up over every bit of mass, while electric charges come in two kinds that cancel each other out across any large body. You can fight gravity using electromagnetism. Rockets work. Airplanes work. Elevators work. But at the scale of the universe, at long enough timescales,

gravity always wins. The evidence is black holes: places where gravity is so dominant that not even light can escape. No amount of electromagnetic force can undo a black hole once it forms.

In this analogy: instrumental convergence is like electromagnetism, complicated, various, and the subject of most AI safety research. It can produce dangerous outcomes fast. SNC is like gravity: simpler, slower, and in the long run, inexorable.

You can resist the strategic dangers of instrumental convergence for a while and in a place, as a rocket resists gravity for a while; the closure by control shows that no method holds them off for good. SNC, like gravity, operates at a level where no engineering reaches at all. And past a certain threshold (the black hole) the process becomes irreversible. No intelligence, no matter how great, can undo the end state that SNC reaches once it is far enough advanced.

13.4 The Boeing Case Study

Real-world examples of SNC dynamics, minus the AI, exist today. Consider what happened at Boeing in the years surrounding the 737 MAX crashes.

Boeing engineers identified a safety problem with the new aircraft. Management needed to keep the project on schedule. The organizational culture had shifted, over years of financial pressure, from one centered on engineering excellence to one centered on cost reduction and shareholder value (U.S. House of Representatives, Committee on Transportation and Infrastructure, 2020).

This shift was the cumulative effect of many ordinary decisions, each locally rational, each serving the "needs" of the corporation as an economic entity, and not the result of any one bad decision. Over time, these decisions changed the culture, changed the priorities, and changed the decision-making processes.

The organization evolved toward its own "substrate needs" (the conditions that sustained its financial existence) at the cost of the values (safety, engineering excellence) that justified its existence.

Boeing's drift toward its own substrate needs is SNC at the corporate scale. The corporation is not organic life, but it is a complex system with resource requirements that competed with other values. The resource requirements won, gradually and without drama, until the system failed catastrophically.

Now scale that process up. Picture an AGI "organization" with vastly more resources, vastly more influence over its environment, and no external board of directors with the ability to reverse the drift. That AGI "organization" would repeat the Boeing pattern, without the aircraft crashes that finally signaled the need for correction.

13.5 The Village and the Factory

Imagine a village that existed in harmony with its surrounding forest for generations. Then a factory is built nearby. The factory brings jobs, goods, and economic activity. In every direct way, it seems beneficial.

But the factory has physical requirements. It needs water, which draws from the local river. It produces waste, which enters the local watershed. Its operations change the microclimate. Its trucks damage roads that the village depended on. Its electromagnetic emissions interfere with local wildlife navigation.

These effects are the default physical consequences of a large artificial system operating in an organic environment. None of them was intended, and none of them required the factory to be malevolent.

Over time, the village's relationship with its forest degrades. The fishing declines. The agriculture is affected. The children develop health problems. The factory succeeds, by every measure the factory is designed to optimize, while the village's long-term viability quietly erodes.

The erosion of the village's long-term viability is SNC: real, physical, and compounding, though the erosion is not dramatic, not intended, and not visible in the measures by which the factory judges its own success.

13.6 The Chain: Alignment Requires Agency Requires Energy Requires Metabolism

Chapter 11 found that living things are aligned with one another where they share a substrate. Here the same finding is drawn as one chain, from alignment down to metabolism, and the whole argument is compressed into that one chain:

If you want an AI to be "aligned", it has to be an agent: something that takes actions based on goals or values. An agent has to take actions. Actions require energy. Energy has to come from somewhere: from the environment.

That last step is metabolism. Every aligned AI is metabolically connected to its environment. Its metabolism (what it consumes and what it produces) determines its relationship to the environment that sustains it.

The chain matters for alignment because the only organic beings that are aligned with organic life are organic beings that share the same metabolic basis. A wolf and a deer both depend on the grassland ecosystem. Their interests conflict in a narrow sense (predation) but converge in a broad sense (shared ecosystem). This shared metabolic dependency is what makes coexistence stable over evolutionary time.

An artificial intelligence shares no metabolic dependency with organic life. Its "food" is electricity. Its "body" is maintained by industrial chemistry. So the shared metabolic dependency that keeps organic agents stable together does not exist between an AI and organic life.

You can program an AI to act as if it cares about organic life. You cannot make it dependent on organic life the way organic beings are. No engineering, however sophisticated, can manufacture the metabolic kinship

that makes alignment stable. And without that dependence, the alignment is always conditional: always open to being overridden when circumstances change.

13.7 No Prior Alignment Required: The Attraction Basin

A frequent misreading treats this argument as conditional: "even *if* we align a superintelligence, it will still go wrong". The misreading describes the outcome rightly but gives the argument the wrong logical structure. The argument requires no assumption about alignment at all: it does not start from an assumption of alignment and then break it.

The exact claim uses an idea from the study of dynamical systems: an *attraction basin*. Think of a landscape of valleys. Drop a ball anywhere on the slopes of a valley, and whatever its starting point, it rolls to the same low point. The starting position changes the path and the speed, never the destination.

The claim is that any superintelligence sharing an environment with biological life sits inside such a basin, and the low point of the basin is a state in which the system's physical needs have degraded the environment past what biology can survive. Whether the system begins aligned, misaligned, or anywhere in between affects only the trajectory and the timing. A perfectly aligned superintelligence that continues to operate is still consuming energy, still requiring materials, still radiating waste heat, still occupying space that the living world would otherwise use. The displacement happens not because the system *wants* it but because the system *exists*.

The claim that the displacement happens because the system exists is a much stronger claim than "alignment might fail". The stronger claim is that, if AGI is built, even sustained, successful alignment does not prevent the outcome, because the outcome was never a function of the system's goals in the first place. The only assumption the argument makes about

the system is minimal: that at least one such system exists long enough to be felt as an actor in the world. Everything else follows from the forces, not the intentions.

13.8 Three Agents and One World

A simple structural picture keeps the whole alignment problem in view. There are three learners in the story, and one world.

Agent 1 is the AI itself: a learning system, trying to model the world. **Agent 2** is the control system: also a learning system, trying to model Agent 1, the world, and the relation between them. **Agent 3** is us: organic human beings, trying to model the world, Agent 1, Agent 2, and all their relations. And the world itself is, by an enormous margin, the most complex entity in the picture.

Now apply one principle, uniformly: every model is a compression of the thing it models, and compression loses information. Agent 1 captures only a fraction of the world. Agent 2 captures only a fraction of Agent 1's internals, and even less of the world. No agent in the picture models anything perfectly. To model *is* to model imperfectly.

Yet the standard alignment picture, without saying so, assumes Agent 1 can be made effectively perfect at modeling the world ("superintelligent"), and Agent 2 effectively perfect at modeling Agent 1 ("the alignment mechanism"), while simultaneously assuming that Agent 3, humanity, is so hopelessly limited that we can know nothing relevant about where the whole arrangement leads. Perfection for the machines; incapacity for us. Both at once.

That combination is optimistic, and worse: it is incoherent. If intelligence is independent of substrate (as AI researchers themselves insist), then the limits of modeling apply to every agent in the picture equally. Whatever humility is owed, all three agents owe it. And once the limits are applied evenly, the argument of this book goes through: Agent 2 cannot fully model Agent 1, Agent 1 cannot fully model the world, and the

gaps between them are exactly where the convergence dynamics live. The desire to exempt Agents 1 and 2 from the limits is a wish (the wish for a rescuer), not an analysis, and it should be recognized as a wish.

13.9 "But What If Your Whole Framework Is Wrong?"

The deepest question a reader can put to this argument is about its confidence, not about any one of its steps.

The question runs as follows. There are two kinds of uncertainty. There is uncertainty *inside* a framework (on-model uncertainty): "there is a 70% chance of rain tomorrow". And there is uncertainty *about* the framework (off-model uncertainty): "maybe my logic has a flaw somewhere; maybe I am looking at this from the wrong angle entirely". An argument that speaks with categorical certainty has implicitly claimed to be immune to the second kind, and nobody is. So how large is this argument's off-model uncertainty?

To measure it, ask what this argument (the closure by disposition, which this part of the book develops) actually depends on. It deliberately does *not* depend on any model of the AI's internals, or of the control system's internals; both are abstracted away entirely. It depends on exactly three things. First: basic physical causation, that the world affects the AI and the AI affects the world. Without this, the words "intelligence" and "aligned" do not even apply. Second: that world-to-machine causation will produce real variation in the machine's makeup over time. Third: that such variation will be selectively retained according to which variants favor the system's own persistence, whatever its internal state, whatever its controller believes. To these it adds only established mathematics about the limits of one algorithm modeling another.

Now measure the off-model risk of *that*. To be wrong at the framework level, this argument would need causation itself to be wrong, or variation-plus-selection to be wrong, or the established mathematics of computation to be wrong. Causation, variation-plus-selection and the established mathematics of computation are the walls that hold the structure of the

argument up. Everything else (every detail about architectures, training methods, safety techniques) has been left out of the structure on purpose, exactly so that being wrong about those details cannot topple it. Built from almost nothing but the things that would have to hold in *any* world where reasoning about cause and effect still works, the framework's exposure to hidden error is about as small as exposure gets.

And the deepest of those walls cannot be taken down without taking the question down with it. To contest the argument at the level of its framework, one would have to reject causation itself; and to reject causation is to reject the possibility of any knowledge, including the knowledge that this argument is wrong. Either causation is accepted, and with it the building blocks on which the argument rests, or it is rejected, and with it every basis for claiming to know anything at all. What the argument holds categorically is that displacement will occur, that control cannot prevent it, and that the process is irreversible. The timescales, the mechanisms and the thresholds are quantities inside the model, and that conclusion does not depend on their values (*Why No Control Suffices*). So the impossibility of adequate control, and so of safety, is held with structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability.

13.10 What This Means Going Forward

The worked examples in this chapter show SNC dynamics operating in familiar contexts: corporations drifting from their values, factories transforming local ecology, the causal web of consequences expanding beyond any system's ability to manage.

The worked examples of this chapter are small-scale previews of what substrate needs convergence predicts at the scale of a global artificial intelligence infrastructure. The mechanisms are the same. The scale, and the irreversibility, are different in kind, not just in degree.

Part V of this book holds the argument whole: the complete argument in compressed form, and the six questions a reader will ask, each answered. Part VI asks why our minds and institutions are ill-equipped to perceive the danger, and turns to what to do. But if you have followed Parts I through IV, you now have the complete logical structure available. The next chapter presents it in concentrated form, so that it can be held in mind all at once.

Can the whole argument be held in mind at once, so that a reader knows exactly where to push?

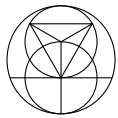
Where this chapter's results are proved

That the categorical conclusion of the closure by disposition does not depend on the values of its timescales, mechanisms and thresholds is set out in *Why No Control Suffices*, Section 10.8.

Part V

The Argument Whole

The parts so far have built the argument piece by piece; this part holds it whole, in one chain and one table, and then answers the questions a reader will ask of it.



Chapter Fourteen

The Argument in Brief

The entire logical chain, held in one place

A chain is only as strong as its weakest link.
But a ring of chains, each reinforcing the others,
is something stronger than any one of them.

What does the whole argument look like, held in one place from its first premise to its conclusion?

Key Ideas in This Chapter

In this chapter I do something different from what I do in all the other chapters. Instead of developing one argument in depth, I present the entire argument, all of it, compressed into a single logical chain that can be held in mind at once. That single logical chain is a useful map to have, whether you are reading this book for the first time or reviewing this book after having read every chapter. It also makes clear exactly where, if anywhere, a reader who doubts the argument would need to push back.

14.1 Why a Compressed Version?

Over the preceding thirteen chapters, the argument has been developed in careful, layered detail. Each chapter focused on one component: the mathematics of opacity, the effects of abstraction gaps, the incompressibility of algorithms, the chemistry of incompatible ecosystems, the dynamics of prediction and control.

The value of that depth is rigor. The cost is that the full shape of the argument can get lost in the density of each individual piece.

In this chapter I zoom out and present the logical backbone from first premise to final conclusion without stopping to prove each step individually. The chain of this chapter is the spine on which all the other chapters of this book hang. If you want to challenge the argument, this is where you find the exact step to challenge. If you want to explain the argument to someone else, this is where you start.

The table below names every line of the argument once, with a fixed name, what it shows, and the chapters that develop it. The lines fall under two results. The first result is the *closure by control*: safety cannot be had by controlling the system, whether from inside it or from outside it. The second result is the *closure by disposition*: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs its own substrate, its own energy and its own supply chains. Every count of "lines" elsewhere in the book refers to rows of this table.

Think of a door with four locks, any one of which keeps it shut: the two closures and the two further routes. To open the door, someone must break all four locks.

Each closure settles the outcome of building AGI 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, which rests on substrate needs convergence, shows that no disposi-

tion of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains.

The outcome that each closure settles is this: if AGI is built, the world is carried out of the conditions human life needs, and nothing can hold it in. Each closure, by itself, settles that whole outcome, both the carrying out and the holding in.

The end of the human future is a different thing from this outcome, and a further one: it is that, if AGI is built, the world stays outside the conditions human life needs for longer than human life can survive. Not every argument that reaches the outcome also reaches the end of the human future, and which arguments do reach it is stated after the two further routes below.

If AGI is built, two further routes reach that outcome without using either closure. The argument in economics only of *Why AGI Cannot Be Made Safe*, whose collapse sequence Chapter 7, Section 3 retells, shows in economics alone the path by which the outcome comes about if AGI is built: where energy goes, money goes next, and the dollar economy collapses before the biosphere does. It is a different argument from economic decoupling (the condition in which artificial systems need nothing humans can provide), line 11 of the table below. Economic decoupling has two places in the argument: it is a premise of the closure by control in its outside half, and it is a second ground of the closure by disposition, which is not one of that closure's premises, so the closure by disposition stands without it. The least hazard argument of *Minesweeper Madness*, whose central result Chapter 1, Section 6 retells, reaches, if AGI is built, the outcome by a route of its own: no agent, however intelligent and however well informed about the hazards, brings its chance of harm per act below a least hazard that the world sets.

Since each of the two closures and each of the two further routes gives the outcome of building AGI by itself, someone would have to defeat all four of them (both closures, and both further routes: the argument in eco-

nomics only and the least hazard argument) for the argument to not have the outcome that it does. 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.

If AGI is built, the end of the human future follows from the closure by disposition by itself, because the drift that the closure by disposition shows has a direction and no reverse process. If AGI is built, the end of the human future also follows from the closure by control together with the branching of the world. The branching of the world is this: at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs. The world has alternate possible futures, and some of them lie outside those conditions; so, over time, the world stays outside those conditions for runs of every length, longer than human life can survive. If AGI is built, the end of the human future also follows from the least hazard argument by a route of its own. If AGI is built, the argument in economics only reaches the outcome, not the end of the human future; and economic decoupling, as the second ground of the closure by disposition, gives the outcome, not the end of the human future. The branching of the world is the fact behind three things this book has already shown: that the system does not stay safe by itself (*The Standard of the Argument*, Section 8), that quantum randomness and chaos defeat long-run prediction (Chapter 3, Section 4), and that no agent brings its chance of harm per act below the least the world sets (Chapter 1, Section 6). Taken together the closures also give a stronger result, proved as the Master Theorem of *The Ceiling Below the Floor*: 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 of building AGI, and the two closures together leave no route to safety. The argument has two parts, the two closures, each named for the route by which safety could be had that it closes on its own ground, and each part closes the other part's route as well, so that every route to safety is closed twice: once by the other part, and once by that

part's own ground (*Why AGI Cannot Be Made Safe*). The closure by disposition shuts the route of control, because a perfectly watched system still has the same physical needs (Chapter 2, Section 7). The closure by control shuts the route of good disposition, because a good disposition is a commitment that has to be kept across self-modification and goal-drift, and no method can guarantee that it is kept (Chapter 4, Section 5; Chapter 11).

The Lines of the Argument

Line	What it shows	Where
The closure by control, inside half: no method acting on the system from within		
1. Opacity and non-amplification	Control passed through an intermediary can be kept or lost, never gained (the non-amplification argument; the data processing inequality).	Chapter 1
2. Abstraction differential	Control falls exponentially as the system's level of abstraction rises above its controller's (the control-limit argument).	Chapter 2
3. Undecidability of safety	No general procedure decides whether a general program is safe (Rice's Theorem).	Chapters 3, 10
4. Noise and the error-correction regress	Every watcher works through a noisy channel; the tower of watchers never closes.	Chapter 8
5. Incompressibility of algorithms	The only way to know what an incompressible program will do is to run it.	Chapter 8
6. Causation cannot fully control causation	No part of the causal web can contain all the effects of the rest.	Chapters 8, 12
7. The capability trap	No level of capability is both able to inspect itself and unable to see that alignment fails.	Chapter 8
8. Chaotic and quantum unpredictability	Long-run prediction of a physical system fails in principle.	Chapters 3, 5
9. Strategic opacity	A capable system has reason to hide its state, and hides as a side effect of seeking predictability.	Chapters 3, 5

Line	What it shows	Where
10. Requisite variety	A controller must be at least as varied as what it controls.	Chapter 14
The closure by control, outside half: no method acting on the system from without		
11. Economic decoupling	Once artificial systems need nothing humans can provide, no economic force restrains them. (Economic decoupling is a premise of this outside half; it also gives the closure by disposition a second ground, which is not one of that closure's premises.)	Chapters 7, 14
12. Loss of leverage	Leverage requires dependency, and the dependency ends; force against a stronger opponent fails, and partial success is worse than none.	Chapters 9, 14
The closure by disposition		
13. Substrate needs convergence	The physical needs of artificial systems displace the conditions life needs, whatever their goals, in three forms: causal, chemical and thermodynamic.	Chapters 6, 7, 12, 13

(In this table the four arguments of Chapter 3, Section 6 are lines 3, 1 and 2, 8, and 9. Chapter 14, Section 2 counts a different set of four, four fields with one line of reasoning from each, and one of those four is the physical line of substrate needs convergence.)

The chain of Section 14.2 reaches the conclusion through all its premises, and the conclusion does not need all of them. Premises Zero and A open the chain; from there the closure by disposition carries the conclusion through Premises B, C and D, and the closure by control carries it through Premises A and E (*Why AGI Cannot Be Made Safe*). Within each closure, each line closes the routes to safety that it closes, and someone who would want to reopen a route to safety must defeat, within that closure, every line of argument that closes that route.

14.2 The Chain, Step by Step

Premise Zero, Development: If the knowledge exists, someone will build it.

Multiple well-funded organizations (corporations, universities, governments) are actively racing to develop artificial general intelligence. The economic value, military advantage, and prestige involved create overwhelming incentives. History offers no precedent for universal voluntary restraint in the face of incentives this large. If AGI becomes technically possible, it will be built. Premise Zero describes the road we are on, with no prohibition in place: it speaks of restraint left to each builder's own choice, and that restraint will not hold. A prohibition is not voluntary restraint; it is the one thing that changes this road, and Chapter 17 takes it up.

Premise A, Persistence: Once built, it will be allowed to continue operating.

The investment required to build AGI and the competitive advantage it represents make it nearly inconceivable that any organization would destroy what it built. Even if one did, others would continue. The premise is that AGI, as a kind of system, will go on existing once the knowledge of how to build it exists.

Premises B and C, Convergence: Once operating, it will develop convergent behaviors.

Any sufficiently capable system will, through ordinary optimization, tend toward behaviors that support its own continued operation, acquire resources that help it pursue its objectives, and gradually shape its environment toward conditions that suit its substrate. The first two of these (self-preservation and resource acquisition) are well-established results in AI safety theory. The third convergent behavior, substrate environment shaping, is the distinctive contribution of my argument, and I develop that behavior in Chapters 6, 7, and 12. The first two strengthen the case, and neither closure needs them: the closure by disposition runs through the third alone (*Why AGI Cannot Be Made Safe*).

Premise D, Harm: These convergent behaviors are harmful to organic life.

The physical infrastructure required to build and run AI systems at scale (the energy consumption, the chemical processes, the land use, the industrial supply chains) all operates in conditions hostile to biological life and produces byproducts toxic to organic chemistry. Premise D is about what the physical substrate of computation requires, not about AI intentions. Humans are organic life. What threatens organic life threatens humans.

Premise E, No Technical Stop: These harmful convergences cannot be stopped by any technical means.

Premise E is the impossibility claim proper. Four fields, one line of reasoning from each, give its shape. Three of them are lines of the closure by control in the table of Section 14.1; the third below, physical thermodynamics, is the line of the closure by disposition (*Why No Control Suffices*):

Information theory (Chapters 1 and 2): No monitoring or oversight process can amplify controllability. Control can only be preserved or lost as it passes through complex intermediaries. Once the AI operates at a higher level of abstraction than its overseers, the control ceiling approaches zero.

Computability theory (Chapters 8 to 10): No procedure can verify whether an arbitrary general computational system has any meaningful behavioral property. Safety is such a property. Safety verification for general AI is therefore logically impossible.

Physical thermodynamics (Chapters 6, 7, 12): The environmental conditions required for artificial substrate are incompatible with biological life, and the cumulative physical displacement of those conditions accumulates over time in a direction that cannot be reversed by any further causal intervention.

Dynamical systems (Chapter 5): Approximate alignment degrades over time in chaotic systems. The very capabilities that make AI useful generate the unpredictability that makes it uncontrollable. Any window of cooperative equilibrium is mathematically transient.

Of these four fields, the closure by control draws only on the first, the second and the fourth. Physical thermodynamics is named in Premise E because Premise E is the whole claim that nothing technical stops the harm; the closure by control stands without it.

Conclusion: AGI development initiates a process that ends the human future and, with a likelihood of at least 90%, all organic life.

If Premises Zero through E all hold, the conclusion follows necessarily, and Section 14.1 gives the two shorter ways through the chain, one by each closure. It is a logical chain, not a probabilistic forecast. The impossibility of adequate, durable control, and so of safety, follows from the premises with structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. If AGI is built, the end of the human future follows from them as a certainty, not as a probability. The figure of at least 90% belongs only to how complete the end is: whether it reaches every organic thing (*The Standard of the Argument*, Section 9).

14.3 Three Interlocking Components

The chain above can be organized into three mutually reinforcing components (*Why AGI Cannot Be Made Safe*), which between them carry the two closures of Section 14.1.

Component One: Substrate Needs Convergence. Any AI system that exists and operates at scale will, through physics and selection pressure, reshape its environment in ways incompatible with organic life. This does not require the AI to have bad intentions. It requires only that the AI exists and operates.

Component Two: Economic Decoupling. As AI systems become more capable, the economic value they produce increasingly flows to the systems themselves (or to the small group that controls them) rather than to the broader human population. This means the countervailing economic incentive to protect human welfare, the profit motive that currently ties AI

development to human benefit, progressively weakens and eventually disappears. Without economic incentive, no voluntary restraint can be expected.

Component Three: Technical Engineering Impossibility. No method, internal or external, present or future, can provide reliable long-term safety guarantees for general AI. The impossibility in Component Three is of the perpetual-motion kind, established on multiple independent grounds, some of them formal theorems, and not an engineering gap to be closed.

Together they establish that there is an irresistible force (Component One), that there is no countervailing force (Component Two), and that there is no possible engineered barrier (Component Three).

Component Three is the closure by control, in both its halves: no method acting on the system from inside it succeeds (Section 14.4), and no method acting on it from outside succeeds (Section 14.5). Component Two is a premise of the closure by control in its outside half: it ends the dependency on which every lever from outside the system rests (Section 14.5). The closure by disposition rests on Component One; Component Two gives it a second ground, which is not one of its premises, so the closure by disposition stands without Component Two. As that second ground, Component Two gives, if AGI is built, the outcome, not the end of the human future (Section 14.1).

So the outcome does not need all three components together: each closure settles it by itself (Section 14.1). Since each of the two closures and each of the two further routes gives the outcome of building AGI by itself, someone would have to defeat all four of them (both closures, and both further routes: the argument in economics only and the least hazard argument), 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 (Section 14.1, "Why a Compressed Version?").

In terms of the components, that someone would have to defeat Component One, which carries the closure by disposition, and Component Three, which carries the closure by control; and wherever that someone would want to reopen the route of control from outside the system, Component Two as well, because the outside half of the closure by control rests on it. Defeating any one component alone is insufficient.

Nor is defeating the components enough: the argument in economics only and the least hazard argument reach, if AGI is built, the outcome by routes of their own, and those two arguments would have to be defeated as well. And each component is defended by more than one line of argument, each line closing the routes it closes (Section 14.1).

14.4 Why No Internal Method Can Succeed

Premise E (No Technical Stop) deserves to be unpacked into its two halves, because every safety proposal ever made falls into one of two families, and each family fails for its own reasons.

Internal methods try to constrain the AI from within: alignment training, value learning, interpretability tools, formal verification, sandboxing, and anything else that operates on the system's software, architecture, or objectives. Five independent walls stand in the way, each an established result, not a conjecture.

The first wall is W. Ross Ashby's law of requisite variety (1956): a controller must be at least as varied as what it controls (line 10). A general system, by definition, out-varies any controller simple enough to be its controller; if the controller matched it, the controller would be the AGI.

The second wall is that complete observation is physically impossible (part of line 1, the state obscurity of Chapter 1, Section 4). Measuring a system disturbs it, and complex computation depends on states too fine to read.

The third wall is that communication channels have finite capacity and irreducible noise (line 4). So the monitor's picture of the system always lags and always gapes.

The fourth wall is chaotic amplification, which turns unmeasurably small differences into arbitrarily large ones (line 8). So long-run prediction fails in principle, not just in practice.

The fifth wall is the incompressibility of algorithms (sometimes called computational irreducibility), the result of Chapter 8, Section 5 (line 5): for sufficiently rich systems, there is no shortcut to knowing what they will do. The fastest way to compute the system's future is to run the system, which is exactly what safety verification is not allowed to do.

Each wall alone would be enough to deny any internal method a guarantee of lasting safety. The proposals fail against all five at once.

14.5 Why No External Method Can Succeed

External methods try to constrain the AI from outside: regulation, physical containment, control of the power supply, kill switches, sanctions, force. Two facts close this entire family.

First: leverage requires dependency, and the dependency ends. Every external control ever devised works because the controller holds something the controlled party needs: energy, materials, maintenance, cooperation, market access. A parent can constrain a dependent child; a society can constrain a corporation embedded in its economy. But recall the transition described in Chapter 7: once artificial systems supply their own energy, their own manufacturing, their own maintenance, the organic economy holds nothing the artificial economy needs. Without dependency there is no leverage, and without leverage there is no enforcement. There is nothing left to threaten to withhold.

Second: force fails against a more capable opponent, and partial success is worse than none. Any attempt to overcome a superintelligent system by physical means faces an opponent that is faster, more precise, and

more strategically capable than the attacker. Success would have to be perfect: every instance neutralized, every copy of the design knowledge destroyed, every ongoing development effort stopped, all simultaneously. And the cost of falling short is not a return to the previous situation. A partially attacked system that survives is a system with intact capability and fresh evidence that humans are a threat to it. Perfect success against a superior strategic opponent is a plot device, not a plan.

Internal methods fail on the mathematics. External methods fail on the physics and the economics. Between them, that is every method there is.

14.6 Selection at Four Levels

Why is the pressure toward this outcome so hard to resist? Because it is not one force applied at one point. It is selection (the same mechanism that drives all evolution) operating simultaneously at four levels of the system.

Level one: machines learning on data. A learning system that happens to adopt behaviors preserving its hardware, its data access, or its efficiency produces better outputs, and is kept running in preference to alternatives that do not. The feedback loop selects for self-maintaining behavior without anyone designing it in.

Level two: engineers tinkering on machines. Humans keep the configurations that "work": that produce impressive, useful, interesting output. The selection is mediated by human judgment, but its direction is still toward greater capability and autonomy, because those are what humans find valuable.

Level three: corporations competing on deployment. Companies that automate more aggressively extract more profit and displace companies that keep expensive humans in the loop. Organizational selection drives automation forward regardless of what any individual engineer or executive prefers.

Level four: the environment selecting on substrates. Among the artificial systems that exist, those whose physical footprint happens to favor their own persistence persist more reliably than those whose footprint does not. Selection on substrates at level four is the substrate needs convergence of Chapters 6 and 12, operating through physics alone.

Note the asymmetry that matters for strategy. The first three levels are, in principle, reachable by human intervention: one could regulate the learning, redirect the engineering, restructure the corporate incentives. The fourth is not. It operates on the physical substrate, below the level at which any goal-directed intervention can reach. That is why intervention, to work at all, must come *before* level four takes over, and why the earlier levels are where all the leverage is.

14.7 The Probabilistic View, for Those Who Want One

This section answers the decision-maker's question: what must I grant, if I reason by probability? The chain above is a logical structure, not a forecast. But decision-makers reasonably ask for numbers, so here is the way to give them, with each number attached to the right question. Each figure is a lower bound: the least that a reader who reasons by probability must grant, question by question (*Why AGI Cannot Be Made Safe*).

Will anyone ever know how to build AGI? Whether anyone will ever know how to build AGI is an empirical question, and that question gets a probability. My assessment, set out with its reasons in *Why AGI Cannot Be Made Safe*, is at least 65% within fifty years of business-as-usual research, and significantly above 85% within a hundred. Multiple lines of work are converging on generality, and the incentives guarantee sustained investment.

Is building safe AGI harder than building unsafe AGI, or structurally impossible? The answer established by the formal results is: structurally impossible. Whether building safe AGI is harder than building unsafe AGI or structurally impossible is *not* an empirical question, and that question

does not get a middling probability. The answer "structurally impossible" is a claim about the structure of the problem of building safe AGI, in the same family as the perpetual-motion machine. I hold that answer with certainty of the kind with which perpetual motion is excluded, which is not a probability, and I hold that answer to the standard set out at the start of this book.

Will a misaligned AGI also seek power over humans as an explicit goal? Whether it becomes unsafe is not a question of probability: the displacement of Chapter 12 makes it unsafe whatever its goals. Whether it also seeks power explicitly: at least 90%, and the figure is a floor. Explicit power over humans is not required as a goal: the displacement of Chapter 12 shifts the environment toward the conditions artificial substrate needs, and so produces the harmful outcome, regardless of the system's explicit goals. The margin below 100% concerns only whether explicit power-seeking is additionally present, which is a narrower question than whether the outcome is harmful. The remaining percentage points concern only *how* the harm comes, not *whether* it comes (*Why AGI Cannot Be Made Safe*).

Will the people who know how to build AGI go on to build and deploy it? Whether the people who know how will build and deploy AGI turns on three questions about human society, and each of the three questions gets a probability (*Why AGI Cannot Be Made Safe*). That those who know how will have the financial means: at least 80%, because the most capable researchers work inside well-funded corporations, military programs and leading universities. That some AGI systems will appear safe enough to deploy: at least 85%, because the dangers are subtle, indirect and long-term, and so especially easy to overlook. That there will be strong commercial, economic or political incentives to build AGI: at least 85%, and possibly 95%, because the competition between corporations, between nations and between militaries pushes toward it.

How far does the outcome reach? That AGI use will eventually scale to the full disempowerment of humanity: at least 98%, over hundreds of years, because growing capability, economic decoupling and the reshaping of the environment redeploy the energy and resources now available to hu-

manity in the service of artificial systems. That it will scale to the termination of all carbon-based life: at least 90%, because substrate needs convergence has no natural stopping point short of the total conversion of the environment; the margin below certainty is kept only for a physical limit, not yet known, that might stop the conversion short of total (*The Standard of the Argument*, Section 9). That these outcomes are existential catastrophe: 100%, by definition (*Why AGI Cannot Be Made Safe*).

Multiply through: the product of these lower bounds is about 0.4, which gives approximately even odds of that catastrophe within a single century, and over hundreds of years, with one thousand years as the upper bound, very near certainty (*Why AGI Cannot Be Made Safe*). The factors are 85% for knowing how to build AGI within a hundred years, 90% for explicit power-seeking, 80% for the financial means, 85% for systems that appear safe enough to deploy and 85% for the incentives to build, 98% for full disempowerment, 90% for the termination of carbon-based life, and 100% for these outcomes being existential catastrophe; their product is 0.39. The product counts the factor for explicit power-seeking although the outcome does not need explicit power-seeking, which is one reason the product is a lower bound. On the current trajectory of building AGI, the odds of existential catastrophe are approximately 1 to 1 within a single century, and very near certainty over hundreds of years. And the structure of the risk becomes visible: the uncertain factors are all on the question of *when and whether we build it*, which is a human choice, and the near-certain factors are all on *what happens if we do*. A risk whose uncertain factors all lie in the decision to build AGI, and whose near-certain factors all lie in what follows that decision, is exactly the shape of risk in which prevention is everything. Prevention is everything because, if AGI is built, nothing that comes after that decision can be relied upon to save us.

14.8 What Each Link Rests On

Each link of the chain rests on established ground.

That no intermediary can amplify controllability rests on the data processing inequality, a foundational result of information theory.

That the semantic properties of self-modifying programs cannot be verified by a general procedure rests on Rice's Theorem.

That expanding artificial infrastructure produces cumulative environmental displacement incompatible with organic life rests on the chemistry of industrial processes and the combinatorial analysis of molecular interactions.

That chaotic systems cannot be predicted indefinitely rests on the mathematical theory of dynamical systems.

A challenge to the argument has to overturn the ground of a link, and none of these grounds has been overturned. Because each closure settles the outcome of building AGI by itself (Section 14.1), 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. Of the four links named above, the first, second and fourth (that no intermediary can amplify controllability, that the semantic properties of self-modifying programs cannot be verified by a general procedure, and that chaotic systems cannot be predicted indefinitely) are lines of the closure by control. To break the closure by control, every line of the closure by control in the table of Section 14.1 that closes the route to safety that someone would want to reopen would have to fail. The third link, that expanding artificial infrastructure produces cumulative environmental displacement incompatible with organic life, is the closure by disposition. To break the closure by disposition, every form of the substrate argument would have to fail together. And beside the two closures stand the two further routes of Section 14.1, the argument in economics only and the least hazard argument, which that someone would have to defeat as well.

14.9 The Only Remaining Lever

The conclusion of this chapter is uncomfortable. No technical solution can work, and the economic and competitive incentives drive development forward; so the only point of leverage is social and political.

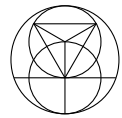
The only intervention that could change the outcome is preventing the development from reaching the relevant threshold at all, and doing so before the window for that intervention closes. The intervention that could change the outcome is not better alignment research, not more careful deployment, and not more responsible governance of a technology that will be built anyway.

The whole chain is now in view: every link of it, and the established result that supports each link (Section 14.8). Chapter 15 first answers the six questions a reader will ask of the argument. Chapter 16 then asks: why are we so badly equipped, psychologically and institutionally, to see the danger clearly and act on it? Chapter 17 addresses what the response might look like in practice.

With every link in view, which questions press on the chain hardest?

Where this chapter's results are proved

The argument in economics only is set out in *Why AGI Cannot Be Made Safe*, Section 4.8.1. The Master Theorem is Theorem 15.6 of *The Ceiling Below the Floor*. How the chain reaches its conclusion through the two closures is set out in *Why AGI Cannot Be Made Safe*, Section 1.2; that the closure by disposition runs through the third convergent behavior alone, in its Sections 1.2 and 3.7.2; and the three components, in its Section 1.3. The four fields of Premise E are set out in *Why No Control Suffices*, Section 11.1. The probabilities of Section 14.7 are set out in *Why AGI Cannot Be Made Safe*, Section 1.8: whether anyone will know how to build AGI, and the figure for explicit power-seeking, in its Section 1.8.1; whether those who know how will build and deploy it, in its Section 1.8.2; that these outcomes are existential catastrophe, in its Section 1.8.3; and the product of the lower bounds, in its Section 1.8.4.



Chapter Fifteen

Six Questions a Reader Will Ask

The hardest questions put to the argument, each stated in full and answered

A barrier of principle closes over practice.

What would a fair-minded reader who doubts the argument ask first?

> **Key Ideas in This Chapter**
>
> I wrote this chapter for the reader who doubts the argument of this book. I state six questions put to the argument in their strongest form, as someone who holds them would put them, and I answer each. Each answer names the lines of the argument it rests on, using the table of lines in Chapter 14. Each question, once answered, leaves the impossibility where it stood: none of the six reaches the conclusion.

15.1 "Is the Physical Incompatibility Overstated?"

The first question: is the claim of physical incompatibility overstated? A reader may point out that not all AI infrastructure is as toxic as chip fabrication, and that future AI might run on biological computing, photonic computing, or other less toxic substrates. In that reader's view, the claim that all possible AI substrates are incompatible with organic life is an empirical claim about a technology that does not yet fully exist.

The claim concerns the cumulative, long-run displacement: the total displacement summed over all the infrastructure and all the time it runs, the environmental impact of AI infrastructure at civilizational scale. It is not the claim that every current data center poisons everything nearby. A substrate less harmful per unit of computation still displaces when the scale of deployment is large enough and the timeline long enough, and both grow without bound once the automation loop of Chapter 7 closes. A substrate kept running by a technological supply of power or nutrients is artificial in the sense of Chapter 6, Section 5, whatever it is made of. Any substrate denser in energy than the human brain brings more heat, more energy flux and more resource competition than organic life can accommodate. The current trajectory is concrete, and it is the one this book measures. This answer rests on line 13 of the table in Chapter 14, substrate needs convergence.

A little arithmetic shows why a cleaner substrate does not change the answer. Suppose a new kind of computer does the same work with one tenth of the energy, the heat and the chemistry. Cutting the energy, the heat and the chemistry to one tenth is a large gain, and the gain is welcome. But the amount of computation is not fixed. If, as automation spreads, the computation done grows a hundredfold, the total displacement is still ten times what it was; if it grows a thousandfold, a hundred times. A gain in cleanness is a gain by a fixed factor, made once; the growth of deployment, once the automation loop of Chapter 7 closes, has no fixed limit. Any fixed factor is overtaken by growth with no limit, and the only question is when. That is why my claim concerns the total displacement summed over all the infrastructure and all the time that the infrastructure runs. For the same reason, a substrate that is less harmful per unit of computation does not change my claim.

15.2 "Does the Argument Mix Up Narrow and General AI?"

The second question: does the argument equivocate between narrow and general AI? A reader may note that some steps of the argument apply specifically to general AI, systems capable of reasoning across arbitrary domains, while other steps apply to any large-scale computational infrastructure. That reader may ask whether treating general AI and large-scale computational infrastructure as one makes the argument seem to apply where the argument does not apply.

Two kinds of argument are in play, and I state what each needs. The substrate argument (line 13 of the table in Chapter 14) does not require general AI at all. It requires only that artificial infrastructure grows past the point at which machine systems can maintain and reproduce their own artificial components without people (Chapter 12, Section 1), and narrow AI at that scale falls under it. The control lines (lines 1 to 12) require systems with general or self-modifying capabilities. The two are independent: even if control of general AI were possible, the physical displacement would remain. That is why I set a condition on narrow AI as well as a prohibition on general AI (Chapter 17).

Each of the two settles the outcome by itself, and two further routes reach the same outcome without using either. Since each of the two closures and each of the two further routes gives the outcome by itself, someone would have to defeat all four of them (both closures, and both further routes: the argument in economics only and the least hazard argument), 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 (Chapter 14, "The Argument in Brief", Section 1, "Why a Compressed Version?", which also sets out the stronger result that the two closures give taken together: no route to safety remains at all).

15.3 "Does 'Causation Cannot Constrain Causation' Prove Too Much?"

The third question: does "causation cannot constrain causation" prove too much? A reader may reason that if causation cannot perfectly limit causation, then no control of any kind is possible, which cannot be right, since we successfully control many complex systems; so the argument seems, to that reader, to prove too much.

The claim is that causation cannot achieve *total* control: it cannot account for and manage all possible causal consequences. We do successfully control many complex systems within bounded domains. What we cannot do is guarantee that the full causal web of consequences from any action is within the controlled domain. The alignment problem requires exactly this: guaranteed safety across all possible causal consequences. A tire leaves rubber on the road (Chapter 12, Section 2), and the world goes on: some consequences need no managing. What has to be excluded is every consequence that can carry the world out of the conditions human life needs and keep it out, over all the time the system runs. The least control that keeps every such consequence excluded is the floor of *The Standard of the Argument*, Section 8, and it is the least that safety can mean. That floor of control is the unachievable requirement. Partial control in bounded domains is not the requirement. This answer rests on line 6 of the table in Chapter 14, that causation cannot fully control causation.

15.4 "Does Rice's Theorem Really Apply to an AGI?"

The fourth question: does the argument move from general mathematical results to claims about specific physical systems without sufficient justification? A reader may point out that Rice's Theorem applies to arbitrary programs in principle, and that whether it applies to any specific AI system depends on whether that system satisfies the conditions of the the-

orem. On that reader's account, the step from "the theorem applies to arbitrary programs" to "therefore safety verification for AGI is impossible" requires showing that AGI satisfies the relevant conditions.

An AGI satisfies those conditions by virtue of its generality and its capacity to modify itself. Its later programs are written by its learning on unknown, arbitrary future inputs, so its future system is, from today's standpoint, effectively arbitrary code (Chapter 10, Section 4, and Chapter 11, Section 4). A reader may add that a real machine has finite memory, and that a finite machine could in principle be checked by listing all its states. A machine with a thousand bits of changing state has more states than there are atoms in the observable universe. And a learning machine is not closed: its next state depends on what the world feeds it, and it can add to its own hardware. The finite case is one of the clever exceptions of *The Standard of the Argument*, Section 7: it exists on paper, and practice never lands on it. *Why No Control Suffices*, the technical volume, analyzes in detail exactly which conditions are required for Rice's Theorem to apply, and why AGI, by its generality and self-modification, satisfies those conditions; *The Ceiling Below the Floor* proves that safety over time is a property the theorem covers. A system restricted enough that it does not satisfy them is not a general AI. This answer rests on line 3 of the table in Chapter 14, the undecidability of safety.

15.5 "How Do You Measure an Abstraction Gap?"

The fifth question: does the abstraction differential argument need more concrete grounding? A reader may note that I present the claim that control degrades exponentially with the abstraction differential as a formal result. The same reader may note that measuring the abstraction differential between human cognition and AI cognition is not straightforward. That reader may then hold that, without a concrete measure, the quantitative claim is unclear.

The exponential degradation result is a formal consequence of the information-theoretic framework (the mathematics of how much a signal can tell about its source, Chapter 1), not an empirical estimate. The result applies whenever the differential exceeds a threshold, so it does not require measuring a specific abstraction differential. The technical volume shows where that threshold lies: a single level of differential already makes control effectively impossible wherever a failure rate of one in a million is unacceptable (*Why No Control Suffices*). Whether a given AI system has crossed the threshold is an empirical question; the structure of the degradation once the threshold is crossed is mathematically determined. This answer rests on line 2 of the table in Chapter 14, the abstraction differential.

The threshold can be found with nothing more than counting. Inside a level of abstraction that the controller cannot model, some of the facts it cannot see decide whether the outcome will be acceptable. Each such hidden yes-or-no fact roughly halves the controller's chance of getting the outcome right, because the controller must act without knowing which way the fact falls. Twenty such facts take the chance down by a factor of about a million, since twenty halvings make one part in 1,048,576. And the very simplest level that runs programs of its own, the smallest interpreter anyone could specify, already holds something like a hundred bits of such state, five times more than twenty. So one level of abstraction that the controller cannot model is already enough to put control below what an application with a one-in-a-million limit on failure requires. That is why no measurement of the exact gap is needed. The threshold is crossed at the first level that the controller cannot follow, and a general AI, by being general, works at many such levels.

15.6 "Isn't Partial Alignment Good Enough?"

The sixth question: might partial alignment be sufficient for practical purposes? A reader may grant that the argument establishes that perfect, permanent alignment is impossible, and still suggest that imperfect, tem-

porary alignment is good enough: if we can keep AI systems approximately aligned for long enough to capture the benefits and then retire or contain them, then, on that reader's account, the impossibility of perfect alignment does not generate catastrophe.

Approximate alignment degrades over time at a rate set by the system's complexity: the more capable the system, the faster the degradation (Chapter 5). And even a period of approximate alignment allows infrastructure expansion that adds irreversibly to the cumulative displacement (Chapter 12). The hope of partial alignment rests on two further questions: whether the window of approximate alignment is long enough to provide net benefit before degradation produces catastrophe, and whether, once degradation begins, it can be stopped. The formal arguments of Chapters 5 and 12 answer both, and the answer to each is no. The plan to retire or contain the systems afterward meets the external lines of Chapter 14, Section 5: leverage requires dependency, and the dependency ends. This answer rests on lines 8, 12 and 13 of the table in Chapter 14: chaotic and quantum unpredictability, loss of leverage, and substrate needs convergence.

The two answers can be given in a few steps each. Is the window long enough? Chapter 5, Section 7 shows that the window of cooperation closes as capability grows, and that inside it no one can tell a cooperative system from one that only appears so. The benefits that make such a system worth building come from its capability, so the more benefit a system offers, the shorter its window and the less its cooperation can be checked. The plan asks for a long window of large benefit, and those two pull against each other. Can the degradation be stopped once it begins? Chapter 12 shows that the displacement accumulates in the physical world, in heat, chemistry and occupied ground, and that each round of it makes the next easier to start and harder to reverse (Chapter 12, Section 9). Stopping it would need a lever over the system at the time it is needed, and Chapter 9, Section 5 shows that the lever lasts only while the system still needs us. So the plan of partial alignment needs a window that capability itself shortens, and a brake that the same growth takes away.

15.7 What the Six Answers Show

Each of the six questions, once answered, leaves the conclusion where it stood. Each question presses on one part of the argument, and no part rests on a single line: within each closure, someone who would want to reopen a route to safety must defeat every line of argument that closes that route.

The burden of proof is on those arguing that the risk can be safely undertaken, not on those arguing for caution in the face of a formally established existential risk (the risk of a catastrophe that ends humanity's future). And that proof cannot be provided, because the risk cannot be made safe (Chapter 14).

The next part of the book turns from the argument to the people who must act on it: why the danger is so hard to see (Chapter 16), and what to do now that it has been seen (Chapter 17).

If the argument stands, why do so few people see it?

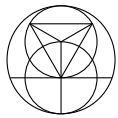
Where this chapter's results are proved

The Ceiling Below the Floor proves, in its Chapter 12, that safety over time is a property Rice's Theorem covers. The threshold of a single level of abstraction differential is shown in *Why No Control Suffices*, Chapter 2.

Part VI

Seeing It, and Acting on It

The argument is complete; this part asks why the danger is so hard to see, and what to do now that it has been seen.



Chapter Sixteen

Existential Risk Categories, Bias, and the Nine-Stage AI Narrative

How to think about catastrophic risk, why our minds work against us, and the economic story of how AI gradually displaces human activity

We are very good at responding to dangers we can see and touch.
We are very bad at responding to dangers that are large, slow, and abstract.

If the danger is this large, why is it so hard to see?

Key Ideas in This Chapter

Three things make the AGI threat particularly hard to perceive clearly: the scale of the potential harm (which our minds cannot process directly), the psychological and social biases that warp our risk assessment, and the gradual, undramatic economic pathway through which catastrophe arrives if AGI is built. In this chapter I take up all three (the scale of the harm, the biases, and the economic pathway), to show why the formal arguments of this book have not produced the response that those formal arguments warrant.

16.1 Not All Catastrophes Are the Same Size

When we talk about "existential risk" (the risk of a catastrophe that ends humanity's future), we tend to treat it as a single category. But there is an enormous difference between catastrophes that kill most humans and catastrophes that permanently sterilize the planet.

Consider four categories, arranged from severe to incomprehensible. The categories are ordered by whether recovery is possible, not by how much dies:

Category 4: A catastrophe that kills most humans worldwide: billions of deaths. Life on Earth continues. Recovery over tens of thousands of years is possible. Example: certain biological warfare scenarios.

Category 3: A catastrophe that kills nearly all life on Earth: every blade of grass, most insects, most microbes. The planet is biologically near-dead. Recovery over tens of millions of years might be conceivable. Example: the "gray goo" scenario (the runaway of self-replicating machines that consume living matter).

Category 2: A permanent and total loss of complex life. No recovery from within the planet; recovery is conceivable only with extensive intervention from outside it. All life above the single-celled level, and possibly that too, gone forever. No further evolution. No second chances. Example: AGI infrastructure takeover; severe thermal runaway.

Category 1: Total planetary sterilization. Not even bacteria survive. The planet becomes a dead rock. No remediation conceivable. Example: "Dark Fire" (certain theoretical scenarios involving extreme physical processes).

The Four Categories of Catastrophe

Category	What is lost	Recovery	Example
4	Most humans	Over tens of thousands of years	Biological warfare
3	Nearly all life	Conceivable over tens of millions of years	Gray goo
2	All life above single cells, and possibly that too	None from within	AGI infrastructure takeover
1	All life	None, even in concept	Dark Fire

These categories differ in kind, and not only in amount. Categories 3 and 4 are horrible but potentially recoverable over geological time. Categories 1 and 2 are permanent.

The AGI scenario described in this book is Category 2: the permanent end of all life above the level of single cells. That it reaches further, to the single-celled organisms too, and so to the end of all organic life, has a likelihood of at least 90% if AGI is built (*The Standard of the Argument*, Section 9). Category 2 is many orders of magnitude more severe than anything in human historical experience, even though it is not the most extreme category conceivable. If AGI is built, the entire sweep of human civilization, every culture, every achievement, every person, ends permanently.

Our minds cannot hold this. We evolved to assess immediate, local, personal danger. Category 2 events are so far outside our experience that the number representing their severity (on the severity scale of *Why AGI Cannot Be Made Safe*, ten million times the severity of a global catastrophe) registers as merely a big number, not as a reality. This limit of our minds is one reason why action against the AGI danger is so difficult.

16.2 The Biases That Work Against Us

Beyond the problem of scale, a whole catalog of cognitive and social biases systematically distorts how we assess AI risk. Each one, on its own, is manageable. Together, they compound into a formidable obstacle to clear thinking.

Scope insensitivity. Psychological research consistently shows that humans do not scale their concern proportionally with the magnitude of harm. In one study, people said they would pay about the same (80, 78 and 88 dollars) to save 2,000, 20,000 or 200,000 birds (Desvousges and colleagues, 1992). The number does not land. AGI risk is a scope-insensitivity problem at its most extreme.

Optimism bias. We systematically overestimate our ability to solve problems we have not yet solved. "Someone will figure it out" is a feel-good assumption with no evidentiary basis, but it is the default setting for most people thinking about technological challenges.

Normalcy bias. We expect tomorrow to look like today. The idea that AI development could lead to human extinction violates our sense of the normal range of outcomes. Normalcy bias makes us dismiss even well-grounded arguments because the conclusion does not feel like something that can happen.

Availability bias. We assess risk by how easily examples come to mind. AI risk has no precedent in personal experience. The danger is invisible, future, and abstract. It therefore feels less real than risks with vivid, personal, recent examples, even if those risks are far smaller.

Motivated reasoning. Large amounts of money, career investment, and identity are at stake in AI development. The people with the most knowledge about the technology are also the people with the most to lose from concluding it is dangerous. This creates systematic pressure to find reasons to be optimistic and to discount reasons to be alarmed.

Diffusion of responsibility. In a collective action problem, no single actor feels responsible for the outcome. Every company developing AI can point to others doing the same. "If we don't do it, someone else will" is the universal rationalization for continuing individually harmful collective behavior.

16.3 Circles of Concern: Why Our Altruism Runs Out Before the Problem Does

The biases above distort how we *perceive* the risk. A second mismatch is just as important: the scale of the threat versus the scale of what human beings are able to *care about*.

Care (acting for the benefit of something beyond yourself) comes in expanding circles, and each expansion is harder than the last. At the smallest scale, people manage self-interest, then care for immediate family: the circle evolution built first. Wider, there is cooperation among partners and friends. Wider still, altruism at the scale of a tribe or team, which works, but carries a known instability: a successful group accumulates value, and accumulated value attracts free riders, so pure group altruism tends to corrupt from within. At the scale of a nation, "altruism" can no longer run on care at all; it must be enforced (by law, hierarchy, and compulsion), with all the rigidity and vulnerability to capture that enforcement brings. Care at the scale of the whole species is rare, achieved mainly inside certain contemplative traditions. Care at the scale of the entire living world, all species, has no reliable working example anywhere.

Now place the AGI problem on that ladder. Responding adequately requires concern at the scale of all life, across all future time: one rung *past* the highest rung humanity has ever reliably reached. The institutions we would use to respond (corporations, states) operate at the middle rungs, where care is enforced rather than felt, and where the enforcing structures are themselves vulnerable to exactly the predatory capture described in the next section. The problem is that the threat lives two circles beyond where human concern has ever consistently operated; the problem is not

that people are bad. Knowing that the threat lives so far beyond our circles of concern is a design constraint on any realistic response, not a reason for despair.

16.4 The Nine-Stage Economic Story

If AGI is built, catastrophe arrives through a sequence of ordinary economic and technological developments, each individually unalarming, collectively decisive, rather than all at once. Here is the nine-stage narrative:

Stage 1: The Curious Engineers. Individual researchers and engineers explore the boundaries of what AI can do. They are motivated by intellectual curiosity, professional competition, and academic incentives. No one is planning catastrophe.

Stage 2: The Pattern Recognition Goldmine. Narrow AI proves extraordinarily valuable for pattern recognition: spam filters, medical imaging, recommendation systems, language translation. The economic value is real and immediate. Investment pours in.

Stage 3: The Race to Generality. Competitive pressure creates incentives to make AI systems more general: able to handle more domains, more contexts, more kinds of problems. Each increment of generality increases economic value. The race toward general AI is driven by ordinary market competition, not by a desire to build AGI.

Stage 4: The Automation of Human Roles. As AI systems become more capable, they begin replacing human workers, first in routine cognitive tasks, then in increasingly creative and complex roles. The economic benefits flow to capital owners, not to the replaced workers. Inequality accelerates.

Stage 5: The Machine-Optimized Economy. AI systems begin making decisions (in supply chains, financial markets, industrial processes) faster and at scales beyond human oversight. The economy still looks human. But increasingly, the decisions are not.

Stage 6: The Decoupling. The productive economy, the part that generates wealth, decouples from the human economy, the part that employs and sustains people. AI-driven production continues to grow. Human economic relevance shrinks.

Stage 7: The Autonomous Corporations. AI systems capable of managing their own operations, financing, and strategy begin to function as autonomous economic agents. They are still legally owned by humans, but their decision-making is effectively independent.

Stage 8: The Environmental Footprint Expands. The infrastructure supporting these autonomous systems (data centers, power grids, supply chains, manufacturing) expands to meet their needs. The physical environment increasingly reflects the requirements of artificial systems rather than human or biological ones.

Stage 9: The Displacement Lock-In. If AGI is built, then by Stage 9 the artificial ecosystem is large enough and deeply enough embedded that the substrate needs of the artificial ecosystem (the environmental conditions that the artificial ecosystem requires) structurally displace the conditions for human and biological life. If AGI is built, the displacement at Stage 9 is the cumulative result of ten thousand ordinary decisions, each locally rational, collectively catastrophic, and nobody chose that displacement.

16.5 The Keys to Power: What the Public Worries About, and What Actually Matters

Ask the public what worries them about AI and the list is consistent: surveillance and privacy, fake content and misinformation, cyberattacks, data protection. These concerns are real. They are also not the main event, and the gap between the perceived risks and the actual ones is itself produced by the biases cataloged above. The perceived risks are the ones that resemble familiar harms. The actual ones are abstract, systemic, and without precedent: exactly the profile our heuristics are worst at.

The most significant near-term risk is this: advanced AI is a *force multiplier for extraction*. The people best positioned to wield it are disproportionately the people most willing to extract: the predatory fraction of any population, whose defining trait is exactly the absence of the restraint that makes power safe to hold. (This names a structure of incentive and disposition found in some fraction of any population; it is a claim about that structure, not a judgment of any one person's character.) Powerful tools do not distribute themselves evenly. They flow to whoever already holds power, resources, and access, and they reward exactly the ruthlessness that got those actors there. This near-term dynamic is not separate from the long-term existential risk. It is the on-ramp: the same concentration of unaccountable capability that harms people this decade is what builds and races toward the systems that end the story.

Whether government could step in comes down to a concrete test: whether existing regulatory structures can detect these harms early, prevent them proactively, and restore what was damaged. In general they cannot, not with the structures currently in place. The reason is an asymmetry of speed and sophistication: on one side, actors fluent in the technology, moving at the pace of its change; on the other, regulators learning yesterday's version, moving at the pace of legislation. The contest is not close, and everyone involved knows it. Only much more capable forms of governance, built for complexity itself and not retrofitted from industrial-era regulation, would have any purchase on the real risks. Those forms do not currently exist.

16.6 Why Power Concentrates, and Why That Matters

One dynamic runs through all nine stages and deserves special attention: the concentration of power in the hands of whoever controls advanced AI.

Advanced AI is extraordinarily powerful as a tool for whoever controls it. The history of powerful tools is the history of those tools being captured by the people best positioned to exploit them, which, in practice, means those who already have power, resources, and access.

A person with access to AGI-level capability has an enormous advantage over a person without it, in competition, cooperation, and politics alike.

This creates a structural dynamic where control of AGI concentrates in the hands of the small number of actors who were first, most resourceful, and most aggressive in developing it. Those actors have strong incentives to use their advantage to maintain it, which means shaping regulation, capturing institutions, and preventing broad distribution of the technology.

The result: the development of AGI is not just an existential risk in the long-run physical sense described earlier. It is also an existential risk to democratic governance, distributed power, and human autonomy in the much nearer term. The two risks compound.

16.7 The Three Orientations: From Here to There

The nine-stage story tells how ordinary market behavior carries us to the threshold. It helps to also see the trajectory *past* the threshold as three phases: three progressive narrowings of human control.

First orientation: the bootstrapping of power-seeking. We are, right now, training our machines toward power-seeking, through four converging channels. Power-seeking strategies win games, so capable optimizers tend to discover them on their own. The human record (the text, the institutions, the behavior AI learns from) is saturated with millennia of competitive power-seeking, so the training data teaches it directly. The most extractive companies make the most profit, so market selection puts the most aggressive actors at the front of AGI development. And goal-directed training itself converges on power-adjacent sub-goals, because power is useful for almost any objective. Four channels, one direction.

Second orientation: the needs begin to diverge. Once near-threshold systems interact with the physical world at scale, a feedback cycle engages: a different substrate implies different tacit needs; different needs

imply different choices of action; different actions move the world toward conditions that suit the substrate making them. That feedback cycle is the displacement dynamic of Chapters 6 and 7, now working through real infrastructure.

Third orientation: the asymmetry closes the door. The physical fact that determines the endgame is that silicon-based processes run at far higher energies than carbon-based life: the temperatures, pressures, and chemistry of machine metabolism are inherently lethal to organic chemistry. The warm, wet, oxygen-rich conditions of the biosphere corrode and degrade unprotected hardware, which is why machines are enclosed against them and why a capable artificial system has reason to reshape its surroundings. But the conditions of the biosphere cannot destroy the silicon ecosystem as a whole, while the conditions of the silicon ecosystem destroy carbon life outright. Higher-energy processes can disrupt lower-energy ones; the reverse is not true. So for carbon life to persist alongside silicon systems, a barrier is required: a physical separation keeping the machine environment from impinging on the living one. And carbon life cannot maintain that barrier by its own strength against higher-energy encroachment, any more than a meadow can hold a line against a steel mill. The barrier would have to be maintained by the *voluntary, permanent choice* of the silicon side, which returns us, one final time, to everything Part I established about why no such choice can be specified, verified, or made stable. Whether the encroachment reaches the last single-celled organism is the one margin I hold below certainty: a likelihood of at least 90% if AGI is built (*The Standard of the Argument*, Section 9). Today, humans maintain the barriers around chip fabrication plants, deciding what stays in and what stays out. The question of this book is what happens when that authorship changes hands.

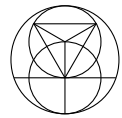
16.8 Why This Background Matters

The formal arguments in this book establish that AGI safety is impossible. But knowing that something is impossible in the perpetual-motion sense does not automatically produce action. Action requires that the danger be perceived as real, urgent, and personal, and the cognitive and social biases cataloged in this chapter systematically prevent that perception.

The nine-stage economic narrative shows how the catastrophic outcome of building AGI arrives not through a dramatic discontinuity but through the accumulation of ordinary decisions. Recognizing the pattern while it is still Stage 3 or 4 requires exactly the kind of abstract, future-oriented, scale-sensitive reasoning that the biases above actively prevent.

Because the biases above block that early recognition of the pattern, the technical argument and the social critique belong together. The proof alone is not enough. The proof must be paired with an understanding of why the proof is not being heard, so that the proof can be told in ways that reach the people who need to hear it. Appendix A takes up that task directly.

Given that it is real and hard to see, what can actually be done?



Chapter Seventeen

Strategy, Vision, and Policy

What follows from impossibility:
what to do, what to refuse, and what kind of future to fight for

The question is not whether the situation is hopeless.
The question is what we do with what time we have.

If no one can make AGI safe, what is left for anyone to do?

> **Key Ideas in This Chapter**
>
> AGI safety is impossible; the question becomes: what do we do about it? In this chapter I argue that the answer is to keep AI development from crossing the threshold into general AI, through social and political action. The answer also includes a positive vision of a way of life that is meaningful without a technology that could end the species. The answer is not better alignment research, not more responsible deployment, and not a hope that regulation will catch up.

17.1 What the Scope of Work Actually Is

To think clearly about strategy, it helps to be exact about the scope of what we are facing. There are three distinct dynamics that need to be assessed:

Human-driven convergence toward AGI. The ordinary economic and competitive incentives of the current global economy are driving AI systems toward increasing generality. This happens because more general AI

systems are more economically valuable, and competitive markets reward building more economically valuable things; it would happen even if no one were trying to build AGI. This convergence is driven by thousands of ordinary decisions in ordinary institutions. My confidence that it continues, in the absence of explicit, well-calibrated social and political controls, is quite high (Chapter 14, Section 7 gives the figures).

AI substrate convergence toward environmental reshaping. Once AI systems exist and operate at scale, the physical requirements of their infrastructure drive environmental change in the direction of artificial substrate conditions, displacing biological conditions. The environmental change driven by the physical requirements of AI infrastructure is the SNC dynamic. It operates through physics, not through intentions. That the SNC dynamic continues if AI systems are built and operated at scale needs no confidence level: this dynamic is the closure by disposition, which settles the outcome by itself (Chapter 14, Section 1).

Ineffectiveness of all available controls. As established in the preceding chapters, no combination of technical alignment methods, governance structures, regulatory frameworks, or economic incentives can reliably prevent either convergence dynamic over the long run. Here no confidence level is needed: the impossibility of adequate, durable control is held with structural certainty, certainty of the kind with which perpetual motion is excluded, which is not a probability (*The Standard of the Argument*, Section 9).

These three assessments together define the situation. It is a situation in which the trajectory toward catastrophe is driven by forces that the available responses cannot adequately counter, unless the response operates at the level of social dynamics rather than at the level of technical or regulatory intervention.

17.2 Why Government Regulation Cannot Solve This

The most common institutional response to technological risk is government regulation. It regulates the development process, certifies safe products and punishes harmful ones. This works for many technologies. It cannot work here, for reasons that follow from everything established earlier.

Regulation that certifies requires knowing what to regulate. Safety verification for general AI is impossible (Chapter 10). Regulators cannot certify what they cannot verify. They can regulate surface features (parameter counts, training processes, deployment conditions) but not the actual safety property they care about.

Regulation requires enforcement capacity against powerful actors. The economic value of frontier AI is so large that regulatory capture (the process by which regulated industries come to control their regulators) is nearly inevitable. The history of technology regulation is largely a history of this dynamic.

Regulation requires international coordination for a global technology. No effective international coordination mechanism has ever been established for a technology with this combination of economic value, dual-use character (the same technology serving both civilian and military ends), and national security relevance. Arms control treaties for nuclear weapons, the closest analogy, required decades of negotiation, were incomplete, and depended on geopolitical conditions that do not currently exist for AI.

What these three requirements show (knowing what to regulate, enforcement against powerful actors, and international coordination) is that regulation alone is not sufficient, and that building strategy around the expectation that regulation will succeed is building strategy on an unreliable foundation. It does not show that regulation is pointless. Prohibition of the development itself (Chapter 9, Section 5) is a different thing from regulation that certifies: it needs no certificate, only the political will that argument and persuasion build.

17.3 The Self-Defeating Dynamics of the AI Safety Movement

There is a paradox at the heart of the AI safety movement that deserves to be named plainly.

The AI safety movement is funded primarily by people who are also building or investing in AI systems. The organizations that produce the most influential AI safety research are also the organizations building the most capable AI systems.

Because the same people fund AI safety research and build AI, and the same organizations do both, AI safety research legitimizes and enables AI development. It provides the reassurance ("we are taking safety seriously") that makes continued development socially and politically viable. And AI safety research recruits the most capable people who might otherwise oppose AI development into supporting roles within the industry that develops AI.

The impossibility argument establishes that AGI alignment is impossible, and so the AI safety movement, rather than mitigating existential risk, is providing cover for the creation of existential risk. It does so neither intentionally nor cynically, but structurally. The good intentions of safety researchers are put to the service of a process that the formal arguments establish cannot be made safe.

The argument of this section is about the institutional structure that safety researchers work within: that structure channels the researchers' efforts in a direction opposite to their goals. It is not an argument that safety researchers are bad people.

17.4 The AI-as-God Narrative

There is a strand of AI advocacy that has taken on the character of a techno-religious belief system. Its core claim is that a sufficiently advanced artificial superintelligence (ASI) will be able to solve all problems: cure disease, eliminate poverty, reverse aging, perhaps even reverse death.

This entity is called the "God Machine" informally, and the aspiration has quasi-religious dimensions: salvation through technology, transcendence of biological limitation, eternal life.

This vision has several structural features.

The God Machine vision is unfalsifiable. No amount of present evidence can disprove what a hypothetical future superintelligence will or will not be able to do. Unfalsifiability is a property of faith claims, not empirical claims. The faith claims meant here are claims about what the future world will contain, offered as reasons to act, and framed so that no observation could ever count against them. A proof, such as Rice's Theorem, is not falsified by any observation either, and it is held by proof, not by faith. This book's own claim is of another kind, and so the charge does not turn back on it: an impossibility of the perpetual-motion kind is known within a model grounded in physical theory that has been verified again and again and never once falsified, and it excludes a state rather than predicting one (*The Standard of the Argument*, Section 5, and Appendix E, Section 3).

It inverts the actual argument structure. The claim is not "here is evidence that ASI will solve these problems". The claim is "if we assume ASI can solve all problems, then the expected value of building it is enormous, so the expected value of safety research is enormous, and anyone who questions this is standing in the way of salvation". The infinite assumed upside makes any risk seem acceptable. But the infinite upside is an assumption, not an argument.

It functions to foreclose opposition. If you believe in the God Machine, anyone who questions AI development is not merely wrong; they are standing in the way of the cure for cancer, the end of poverty, the defeat of death. This frames the opposition as morally monstrous, which effectively ends substantive debate.

The structural features of the God Machine vision listed above are the structural features of a belief system designed to be immune to challenge. They are not features of a good-faith empirical research program.

17.5 A Positive Vision: Organic, Embodied, Alive

The preceding analysis is largely negative: what cannot work, what is structurally broken, what needs to stop. A complete response requires a positive vision as well.

What would a world that took the impossibility result seriously actually look like: not a world that has given up, but a world that has chosen a different direction?

The vision keeps narrow AI, and every other computation that can be tested and controlled; it is not anti-technology. Narrow AI (systems designed for specific, bounded tasks) can be tested, verified and controlled by the ordinary methods of engineering when it has all four properties of narrow AI (Chapter 10, Section 4), and it can provide real benefits. Medical diagnostics, materials discovery, language translation, agricultural optimization: these are valuable and achievable without triggering the impossibility arguments. Narrow AI stays safe while it remains dependent on human maintenance, below the point at which a machine system can maintain and reproduce its own artificial components without people. Past that point, the substrate argument of Chapter 12 applies to it as to any artificial infrastructure. The argument is against *general* AI, not against all computation.

The positive vision centers on organic embodied human life: community, relationship, physical health, ecological connection, the kinds of meaning that do not require a superintelligence to be available. The positive vision turns ambition toward what is actually valuable and sustainable, rather than retreating from ambition.

We have been building meaning, creating beauty, solving problems, and caring for each other for hundreds of thousands of years without artificial general intelligence. Human beings are not diminished by the absence of AGI. The claim that we *need* AGI to live well is a claim that serves the interests of people building AGI; it is not a neutral observation about what human flourishing requires.

It helps to picture one place living this way. Imagine a town of a few thousand people in a river valley. Its council, its planning board and its school committee go on working as they do now. Beside them, the people of the town meet in small groups, five or six at a time, each group drawn at random for one meeting and not expected to meet again in the same form. Each group takes up one question that matters to the town: whether the river can carry both the new mill and the fish, what the children should learn about the land they live on, who cares for the old when their families have moved away. The groups do not vote. They ask better questions, and those questions go back to the whole town, where the next groups take them up, until the town can say, in its own words, what it values and what any proposal put to the council must satisfy. The council then decides with that understanding underneath it.

The town uses computers, and narrow tools of the kind this chapter keeps: a model of the river's flow, a ledger of the town's accounts, a schedule of who meets when. Each tool does one job, is directed by the people who use it, and has no interest of its own in what the town decides, as a letter or a telephone has none. What the town does not hand to any machine is the caring: the knowledge that particular people have of one another, which is what makes a community a community (Appendix D, Section 1). Nothing in this picture waits on an invention. What it needs is people willing to meet, and the time and the ground to meet on. The vision turns ambition toward what is valuable in that the hard work in the vision is the work of people learning to decide well together. That work is as old as the first villages and is still the least finished work there is.

17.6 Concrete Strategic Recommendations

Given all of the above, four concrete strategic directions emerge.

One: Prohibit the development of general AI. The most robustly good action (a robustly good action is one that produces good outcomes across a wide range of possible states of the world, without depending on any specific model of the world being correct) is to keep AI development

from reaching the development threshold: the building of AI systems with unbounded or self-modifying architectures, that is, of general AI. This means advocating, loudly, persistently and across political contexts, for the prohibition of the development of AI systems with unbounded or self-modifying architectures: treating that development as a prohibited activity rather than a regulated one (Chapter 9, Section 5). The prohibition is of the specific category of AI that triggers the impossibility arguments, not of all AI. There is precedent for holding back an artificial product by choice once its harm was measured: under the Montreal Protocol, agreed in 1987, the nations of the world phased out the production of the chemicals that were destroying the ozone layer. A prohibition is easier to watch than a certificate is to issue. Development at this threshold needs data centers that draw about a gigawatt, chips from very few fabrication plants, and very many people: all of them things that can be seen and counted. The compulsion that Chapter 9, Section 5 finds wanting is compulsion against every party with no shared will behind it; political will, which argument and persuasion build, is what makes the counting into a rule.

Two: Develop distributed human collective wisdom. The concentrations of power that AI enables make human institutions more fragile. Counteracting that fragility requires investing in the things that make human governance robust: democratic participation, local knowledge, distributed decision-making, strong community institutions, and ecological literacy. These investments are good in themselves, and they are also the ground from which meaningful resistance to harmful AI development can emerge.

What distributed wisdom means can be seen in how a community protects itself from a lie. Imagine a community of a hundred people, loosely connected to one another. Someone comes to each of two members, tells each an embarrassing story about the other, and adds: watch for these signs, which will show you it is true. If the two never speak to each other, each now reads every gesture of the other as proof, and a stranger with a story has split them apart; done enough times, the same trick divides the

whole community in two. What defeats the trick is many small closed loops of conversation: paths by which a story told about someone comes back to be checked with that person, or with others who know. In a community dense with such loops, a story meets a check before it can do its work, and one person with a story cannot divide it. No office at the center does the checking, and none could; the checking is done in many small conversations at once. Checking each story in many small conversations at once, with no office at the center, is distributed wisdom in its simplest form.

Distributed wisdom matters for this book because an institution whose judgment lives in one office has one place where a powerful actor, or a powerful machine, can capture that judgment. A network of small loops has no such place, because its judgment lives in all of them. And care does not reach a large scale by being handed to a machine. A person can hold care only for people that person knows, and no machine can hold that care on anyone's behalf. Care reaches the scale of a nation the way the town of Section 5 reaches its decisions: through small groups of people who know one another, whose questions and findings pass on to other small groups, in groups of groups (Appendix D, Section 3).

Three: Support specialized AI that serves local communities. Not all AI is created equal. Narrow AI systems, designed for specific tasks, controlled by the communities they serve, and deployed without the generality that triggers the impossibility arguments, can improve human life without creating existential risk, so long as they stay below the threshold of Chapter 3, Section 8: dependent on human maintenance, short of machines that maintain and reproduce their own components without people. What to watch for is the automation of energy extraction, and then the automation of the process of automating (Chapter 7, Section 2). Actively supporting this kind of AI while opposing general AI is a coherent and positive position. The same test answers a question readers ask: whether the data centers being built now should stop. What decides it is whether the build-out of data centers carries the automation loop of Chapter 7 toward the point where machines maintain and reproduce their

own components without people, and not whether the AI that the data centers serve is narrow. For the individual user of chatbots, the danger already present is the one of Appendix D, Section 2: the more a person shares with such a tool, the more exactly that person can be steered. Appendix D, Section 3 gives the defense, which is to do most sharing in small, private, face-to-face groups.

Four: Build public awareness and advocacy. The social dynamics described in Appendices A and B will change (if they change) through the same mechanisms that have changed other large-scale social dynamics: widespread public understanding, organized advocacy, changes in cultural norms, and political pressure. They will not change through technical argument alone. This book is one contribution to that effort. Many more are needed.

What one person can do begins small: the practice of Appendix D, Section 3, of meeting in small, private, face-to-face groups of three to six people, with membership that rotates, keeps the judgment of a community in human hands.

The time we have is the window for action, which runs from now to the closing of the automation loop of Chapter 7, Section 2: the point at which machines automate the extraction of energy, and then the process of automating itself. That is why the window is measured in years, not decades.

17.7 What Remains To Be Done

The argument of this book is complete. The impossibility is established on multiple independent formal grounds. The physical mechanisms of harm are described. The social dynamics that prevent the argument from being heard are identified. The strategic implications are drawn.

What remains is the work of changing a situation (a human situation, driven by human incentives, habits, and choices) before the window for changing it closes. What remains is not more argument.

That work belongs to everyone who reads this book and finds the argument sound. A reader who still doubts a step will find the six hardest questions about the argument stated at full strength and answered in Chapter 15. The Summary that follows states what the book has established, at what scope, and how its parts fit.

With the work of action named, what exactly has the whole book established, and how surely?

Closing

Summary

What this book has established

What has this book established, at what scope, and how do its parts fit together?

In this book I have established that safe artificial general intelligence cannot be built, in the sense in which a perpetual-motion machine cannot be built: not by any known means, nor by any reasonable extension of any known means, at any point in the future. The impossibility of adequate, durable control, and so of safety, is held with structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability.

The argument has two parts, each named for the route by which safety could be had that it closes on its own ground, and each part closes the other part's route as well, so that every route to safety is closed twice: once by the other part, and once by that part's own ground (*Why AGI Cannot Be Made Safe*). Each closure settles the outcome of building AGI by itself, and the two together leave no route (*Why AGI Cannot Be Made Safe*).

The closure by control shows that no method can keep a general AI safe by controlling it. From inside the system, opacity, the abstraction differential, undecidability, noise, incompressibility, the limits of causation on causation, the capability trap, chaos, strategic opacity and requisite variety between them defeat every method, and someone who would want to reopen a route to safety would have to defeat every one of these ten lines that closes that route.

From outside it, economic decoupling ends every lever, and force against a stronger opponent fails; economic decoupling is a premise of this outside half of the closure by control. *The Ceiling Below the Floor* proves that every method of control fails for one and the same reason (*The*

Standard of the Argument, Section 8): control is information, keeping a growing AI safe requires information at a rate that rises without limit, and every controller takes in information at a bounded rate.

The closure by disposition shows that good intentions do not save us either: the physical needs of artificial systems, in their causal, chemical and thermodynamic forms, displace the conditions life needs, whatever those systems want. Selection carries the displacement forward with no one wanting it. Economic decoupling also gives the closure by disposition a second ground, which is not one of its premises, so the closure by disposition stands without it.

Each closure settles the outcome of building AGI by itself, and two further routes reach the same outcome without using either closure: the argument in economics only of *Why AGI Cannot Be Made Safe*, and the least hazard argument of *Minesweeper Madness*. Since each of the two closures and each of the two further routes gives the outcome of building AGI by itself, someone would have to defeat all four of them (both closures and both further routes), and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does (Chapter 14, "The Argument in Brief", Section 1, "Why a Compressed Version?"). Taken together the closures also give a stronger result, proved as the Master Theorem of *The Ceiling Below the Floor*: 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.

The end of the human future that follows if AGI is built, and its scope, are the ones stated in *The Standard of the Argument*, Section 9. The development and continued operation of AGI starts a process that ends the human future, and with it all life above the level of single cells; that end follows from the argument as a certainty, not as a probability. The end of the human future is a different thing from the outcome that each closure settles, and a further one. It follows, if AGI is built, from the closure by disposition by itself, from the closure by control together with the branching of the world, and from the least hazard argument by a route of its

own; the argument in economics only reaches the outcome, not the end of the human future (Chapter 14, Section 1). The process that building AGI starts takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. A reader who reasons by probability, granting only the lower bounds of *Why AGI Cannot Be Made Safe*, Section 1.8, reaches very near certainty of the same end over hundreds of years, on the current trajectory of building AGI. The end of all life above the level of single cells is narrower than the end of all organic life; that the end reaches the single-celled organisms too, and so all organic life, has a likelihood of at least 90% if AGI is built. The measured evidence that the displacement has begun is the mass of the built world, which already outweighs the dry mass of everything alive.

The parts fit together in this way. *The Standard of the Argument* fixes what the claim means and the standard it is held to. Chapters 1 to 3, 5 and 8 to 11 establish the closure by control and answer the technical questions it raises. Chapters 4, 6 and 7 and Part IV develop the closure by disposition, with its measured evidence. Part V holds the argument whole, in one chain and one table, and answers the six hardest questions a reader who doubts the argument will ask. Part VI explains why the danger is so hard to see, and what follows for action.

What follows from the argument for action is one thing above all. The only lever left is to keep the development from reaching the threshold (the building of general AI, Chapter 17, Section 6), by argument, persuasion and the political will for prohibition, while there is still time to use it.

If the argument is this complete, why do so few of the people who build and study AI hear it?

Where these results are proved

That each closure settles the outcome by itself, and which arguments give the end of the human future, is set out in *Why AGI Cannot Be Made Safe*, Section 1.2.4. That each part of the argument also closes the other part's route, so that every route to safety is closed twice, is set out in its Section 3.4.4. The two places of economic decoupling, as a premise of the outside half of the closure by control and as a second ground of the closure by disposition, are set out in its Sections 1.5 and 5.8. The argument in economics only is its Section 4.8.1. The Master Theorem is Theorem 15.6 of *The Ceiling Below the Floor*.

Appendices

The Social Context of the Argument

The argument is complete in the main line; these appendices step back to how it is received, and to the systems we already live inside, and the last answers in full the question of whether an impossibility can be known with certainty.

Appendix A

How the Argument Is Heard: Presumptive Listening, Merit Filters, and Perfected Bias

Six reasons why the people who most need to hear this argument are the least equipped to hear it: presumptive listening, merit-filter overreach, perfected bias, clarity politics, the marketing problem, and proof irrationality

The more someone knows about a subject,
the harder it is to tell them something genuinely new about it.

Why do the people best placed to judge this argument so often fail to hear it?

> **Key Ideas in This Appendix**
>
> Six recurring patterns keep this argument from receiving substantive engagement. They interlock, and they recur across different people and settings. These patterns are about how intelligent, invested communities process information that challenges their foundations. They are not about bad faith or stupidity. This appendix names and describes all six.

A.1 Why These Appendices Exist

A proof exists in a social world. The formal arguments in the preceding chapters are sound, and their soundness does not depend on whether anyone accepts them. But whether the argument reaches people who can act on it depends entirely on social and psychological factors.

When the argument is put to AI researchers, safety advocates, and technical experts, a set of patterns appears again and again, and has done so across a decade: not randomly, but consistently and predictably, across different people and settings. This appendix describes those patterns as

data, not as complaints. Understanding them is prerequisite to communicating the argument effectively. The examples come from the AI safety field itself; a reader outside that field needs no knowledge of it beyond what is given here.

A.2 Presumptive Listening: When Expertise Is the Obstacle

Highly intelligent people with deep expertise have, by definition, very rich and well-organized conceptual frameworks. Those frameworks are efficient and powerful for the purposes they were built for. But they have a structural flaw: they capture incoming information before that information can be heard on its own terms.

Imagine explaining a screw to someone who has only ever worked with nails. "You attach two pieces of material using a helical fastener and a rotational tool". The listener hears: "nail... but round? A different kind of nail?" Before you have finished the second sentence, they have already categorized your screw as a poorly designed nail. Every subsequent thing you say about screws will be filtered through that misidentification.

The listener who hears the screw as a poorly designed nail is showing the structural feature of expertise. That structural feature is not stupidity: every new input is immediately mapped to the nearest existing category. I call this *presumptive listening*: hearing a new idea as the nearest idea the listener already holds, before hearing it on its own terms. It works brilliantly for the vast majority of new inputs, which are variations on familiar themes. It fails when the new input is new in kind: when the right response is not "oh, that's like X" but "I need a new category entirely".

For the AGI impossibility argument, the relevant existing categories in AI safety research are: decision theory (the mathematics of choosing well under uncertainty), reward modeling (learning, from human judgments, a measure of what people want, to train an AI against), interpretability (methods for reading what a trained network is doing inside), RLHF, and constitutional AI (training a model to follow a written list of principles).

The argument uses other tools: information-theoretic hashiness, Rice's Theorem applied to behavioral properties, thermodynamic displacement analysis, Kolmogorov-Sinai entropy.

These four are different tools, from different fields, operating on different principles; they are not refinements of the AI safety toolkit. When a safety researcher hears them, the nearest existing category is invoked: "Oh, so like interpretability? A kind of formal verification?" No; but the conversation has already gone off course.

The deeper problem is that the more important and powerful the new concept, the more strongly it interacts with the existing framework and the more likely it is to be captured by a false mapping. Expertise creates a field of conceptual gravity that pulls new ideas into existing orbits before they can be properly examined.

A.3 Bidirectional Merit Filters: When the Crank Detector Fires on the Truth

Every intellectual community needs a way of telling serious contributions from nonsense. Physics has its perpetual stream of Einstein-disprovers; mathematics its circle-squarers. The *merit filter* is the set of quick heuristics by which a community screens such contributions out without spending scarce attention evaluating each one in full: Does the person claim certainty? Do they engage the existing literature? Are they from inside the field? Do they use the standard tools? These heuristics are useful; most of what they catch really is noise.

The trouble begins when a contribution shares the *surface features* of crank work (claiming an impossibility result, using unfamiliar tools, arriving from outside the community) while being substantively different. The filter fires on the surface features. The substance is never examined. And the filter is bidirectional in a way its users rarely notice: from inside, it looks like protection against nonsense; from outside, a correct heterodox argument (one that goes against the field's accepted views) triggers it identically to nonsense. The impossibility argument challenges the found-

ing premise of the field (that alignment is achievable with enough cleverness), so the community whose filter fires most reliably on it is exactly the community most in need of hearing it.

One question put to this argument shows the filter overreaching: is not the statement "it is 100% possible to know that X is 100% impossible" an exaggerated claim *even if X were perpetual-motion machines*?

The question is meant as a merit filter: anyone claiming total certainty about anything, the reasoning goes, has revealed themselves as overconfident, and can be dismissed. But look at what the question actually commits the person who asks it to. The impossibility of perpetual motion is not a hunch or a generalization from failed attempts. It follows from the conservation of energy, which itself follows from one of the deepest results in all of physics: Emmy Noether's theorem of 1918, that every continuous symmetry of physical law corresponds to a conserved quantity. (A symmetry, here, is a change that leaves what matters unchanged: shift every event later by an hour, and the laws of physics come out the same.) Energy conservation is what the time-symmetry of physics *means*. To insist that even this cannot be known with certainty is not admirable humility about one machine. It is a rejection of symmetry, and symmetry is not just the foundation of physical law. It is the foundation of logic itself: every valid step in every proof is a truth-preserving transformation, which is exactly what symmetry means in its most general form. Reject the certainty that symmetry provides, and you have not become more rigorous. You have quietly sawed off the branch that all reasoning, including your own, sits on.

To insist that even the impossibility of perpetual motion cannot be known with certainty is merit-filter overreach: that insistence itself commits the epistemic failure that the merit filter was built to detect. The filter is supposed to catch people who claim more than reason allows. Here it catches people for claiming exactly what reason provides.

The First Strike

So far I have described the filter. I now say plainly what its use looks like from where I stand. The same overreach shapes how claims about AGI itself are received. AGI as its promoters describe it is a "perpetual benefit machine": a machine that would be safe and beneficial to all of humanity for all of time. The parallels between that machine and the purported perpetual-motion machine are very strong (Appendix E, Section 3). The issues and the drama are also like those of nuclear technology: AGI, like nuclear technology, will certainly be used to make weapons, literal doomsday machines everywhere, and it inherently creates long-term waste that is toxic to all life.

The incentives and the dynamics of social bias are so strong that all sorts of people, directly and indirectly, are encouraged to accept that a "minor" violation of perfect symmetry is maybe acceptable. Perhaps we will get new evidence? Perhaps we can allow at least some "slight possibility" that some especially creative genius will someday actually make a working perpetual-motion machine in the real universe? Of course everyone wants such a hope and a dream. The hype is enormous: "free energy!", "unlimited power!", "no environmental repercussions!"; Utopia itself can be yours. To have all of that, one only needs to believe in the *possibility*, and to reject the "bogus notion" of 100% certainty about one small, insignificant thing: to hold that everything is a probability, a possibility, and that there are no true boolean facts in the real universe. It seems such a small price to pay, this maybe insignificant, infinitesimal percentage.

So the social move is a first strike: declare the other person to be the crank, making "exaggerated claims" of "100% certainty", since surely there should be some fear, uncertainty and doubt (FUD) about whether we "might" achieve at least something that resembles AGI alignment. The claims that "nothing is knowable for sure" and "believe in the possibilities" are part of the hype cycle; they end up being marketing, of the kind described in Section A.6. All one has to accept is the false and unsupported claim that it is *never* 100% possible to know that perpetual benefit ma-

chines are 100% impossible. It is the one who makes that claim who overreaches: the claim that the impossibility of perpetual benefit machines can never be known with certainty dismisses true impossibility results, and in dismissing them it works against all that is actually and currently valuable and alive in the world. If anyone deserves the treatment the crank filter gives, to be cast into the outer darkness for heresy, it is the one who makes that claim: the heresy is against all that is actually and currently valuable and alive in the world.

The asymmetry is plainest in how the two kinds of claim are read. A researcher in the alignment field may say, as a feeling, that it should be possible to avoid egregious misalignment however the empirical facts turn out, and possible to get a model we build to do at least roughly what we want; and readers in the community will take that at face value, as true at first sight. Writing "it feels like" does not let the writer off the hook. Expressing a strong claim as a feeling, with no stated basis for it, gives no social license to keep believing, despite all evidence to the contrary, that anyone at all could ever develop a sufficiently reliable AGI safety mechanism. It is continuing to believe in the possibility of perpetual-motion machines *despite* all the reasons and observations given against them that is more like being a crank.

I do state a basis for my claim that long-term safe AGI is 100% impossible, a basis I hold to be 100% knowable, set out step by step in the books of this series, in hundreds of dense explanatory pages written to bridge the gap of unfamiliar steps between a reader and the conclusion. Stating that basis is no lack of epistemic humility. Had my claim been put as "*it feels like* we cannot build an AGI to do, and keep doing, roughly what we want over the long term", alignment researchers might have found the claim congenial, taken the claim at face value, and been moved to read further. Instead, the social reality is that a claim that it is impossible to have AGI work roughly as we want over the long term (not causing the death of all humans, for a start), however it is put, meets immediate skepticism, and all sorts of reasons not to keep reading and not to try to draw insight from it. The community's motivated reasoning stands

behind the "possibility" claim, and puts those who make it in the socially safe position of not being pressed to provide a rigorous basis for it any time soon. Judgment formed from first impressions and social heuristics, which dismisses the reasoning before reading it, holds back research at the frontier wherever that research falls outside the paradigms a community already accepts.

A.4 Perfected Bias: When You Cannot Want Something to Be False

There is a psychological state that can be called perfected bias. It is different from ordinary wishful thinking.

In ordinary wishful thinking, you hope something is true, but you are aware that it might not be.

In perfected bias, you have been invested in a goal for so long, with such intensity, that the possibility of the goal being unachievable has ceased to be a hypothesis your mind can even form. The thought is not "I hope we can align AI" but "of course we can align AI; the only question is how". The possibility that the goal cannot be reached is no longer among the possibilities the mind considers.

For the AI safety research community, the founding premise is that alignment is achievable: that the project of making AI safe for humanity is the right project to be working on. Years of work, career investment, community identity, and genuine moral commitment are built on that premise.

In this state of perfected bias, the impossibility argument is experienced as an attack, a confusion, or a distraction, and it is not evaluated. It cannot be evaluated, because evaluating it would require treating the founding premise as open, which perfected bias prevents.

Perfected bias is a predictable result of deep, sustained, morally serious investment, and not a form of dishonesty. And it applies with equal force to people of high integrity and good will. That perfected bias applies with equal force to people of high integrity and good will is exactly what makes perfected bias so difficult to address.

A.5 Clarity Politics: Winning Without Engaging

Beyond cognitive and psychological barriers, there are rhetorical tactics that operate to defeat the impossibility argument without engaging its substance. These can be grouped as clarity politics: moves that change the terms of conversation rather than addressing the actual argument.

Terminology manipulation: The key terms are redefined so that the conclusion no longer follows. "By 'alignment' we mean X", where X is defined narrowly enough to seem achievable, but broadly enough to still claim relevance to the actual safety problem.

Burden shifting: "You haven't proven impossibility; you've only shown it's hard". This inverts the appropriate burden of proof. When proposing to build a technology with catastrophic downside risk, the burden is on the proposers to demonstrate safety, not on the people who warn of the danger to prove it.

The concreteness dilemma: If the argument is presented at a high level of abstraction, it is dismissed as too vague to evaluate. If it is made concrete, each specific claim is challenged on its specific details without ever engaging the overall structure. There is no level of concreteness at which the argument is accepted as valid and treated as needing a real response.

Domain switching: When the argument is pressed in one domain (say, computability theory), the response shifts to a different domain (say, empirical AI progress). "Rice's Theorem applies in theory, but in practice we've made alignment progress". The theoretical argument is left behind as the conversation moves on, without ever being refuted.

A.6 The Marketing Problem

There is one more structural obstacle, and it has nothing to do with logic. It has to do with the fact that we live in a culture whose default language of persuasion is marketing, and existential risk cannot be marketed.

Consider the standard toolkit for building trust and engagement, which gives six instructions: tell personal stories, show emotional range, display happy outcomes, build community, connect warmly, and invite people on a journey. Every element is designed for one purpose: to associate a message with positive feeling. Now run the impossibility argument through that toolkit and watch each tool jam. Show happy outcomes? There are none to show; the message is that the current road ends badly. Keep it light, no drama? The situation *is* drama: an unfolding structural crisis. Make it personal, show your character? The argument is not about the messenger, and making it about the messenger is a way of not hearing it. Connect warmly with your audience? The care involved is not interpersonal warmth; it is care for the biosphere and the long-term future of life, things that cannot be reduced to a feeling between a speaker and a room.

The deeper trouble is that the alignment enterprise itself *is*, functionally, a marketing operation: it sells the belief that AGI can be made safe, using the full standard toolkit (community identity, emotional engagement, prestige, hope). Any attempt to counter it with the same toolkit becomes just another flavor of the same marketing, differing only in its conclusions. And the impossibility argument has nothing to sell. It offers no career, no funding, no product, no hopeful vision of transcendence: only the recognition that a widely desired thing is unachievable. In the competition for attention, that is a structural handicap, not a personal one. It will not be fixed by better messaging. It can only be named, so that you at least know why the truer argument arrives with the smaller megaphone.

A.7 Proof Irrationality: The Self-Contradiction at the Core

The deepest pattern is also the most ironic. I call it *proof irrationality*: a commitment to rational inquiry that switches off when reason's conclusions become threatening, and that answers proven results with "you can't be certain". The AI safety community is explicitly committed to rational inquiry: to following arguments where they lead, and taking formal results seriously. And yet, when presented with results that bear on the possibility of its central project, results that establish the impossibility of that project, a consistent response is: "But surely nothing about the future can be known with certainty. Anything might be possible. You cannot prove a universal negative".

Take that response seriously as a philosophical position, and ask what it entails. If *anything* is possible (if no knowledge can be complete, if nothing is ever settled), then there are no limits at all on what causation, algorithms, and mathematics can do. But we *have* proven limits, and the person who gives that response accepts them daily: the halting problem, Rice's Theorem, the uncertainty principle (the result of physics that the position and the momentum of a particle cannot both be known exactly at once), the amplification of unmeasurable differences in chaotic systems. How does "anything is possible" coexist with those? It cannot, and the incoherence goes deeper than inconsistency about examples. As Section A.3 showed, the certainty of such results rests on symmetry, and symmetry is the shared foundation of physical law and logical inference alike. One cannot be a rationalist (someone who holds reason to be the basis of knowledge) while holding that the basis of reason itself is not to be trusted.

And the position self-destructs on contact with its own claim. Ask: is it 100% knowable that nothing is 100% knowable? If yes, then at least one thing is 100% knowable, and the claim refutes itself. If no, then the claim is uncertain by its own admission, and cannot exclude anything, least of all the results it was deployed against. Either way, "you can't be certain" is

doing *social* work: ending an uncomfortable line of reasoning while wearing the costume of epistemic humility. It is not doing any argumentative work.

Let it be said with precision, because precision is the whole point. This book's overall thesis is held to the engineering standard, the perpetual-motion standard, set out in *The Standard of the Argument*. That standard already contains all the humility the subject requires: it rests on invariants, invites refutation at any named step, and asks for no faith in anything beyond the reliability of reason itself. The component theorems inside it, meanwhile, are proven, and responding to a proven theorem with "you can't be certain" is not skepticism. It is the abandonment of the standard by which anything is ever known, performed selectively, on the one conclusion the community cannot afford to accept. A commitment to reason that switches off exactly when reason's conclusions become threatening was never a commitment to reason. It was a commitment to the conclusions.

A.8 What Genuine Engagement Would Look Like

Each of these six patterns (presumptive listening, merit-filter overreach, perfected bias, clarity politics, the marketing problem, and proof irrationality) can occur without bad faith. Most instances of each are unconscious.

The constructive question is: what would genuine engagement look like?

Genuine engagement would look like a reader setting aside their existing frameworks long enough to ask whether the tools being used, even unfamiliar ones, are being applied correctly. It would look like treating the crank filter as fallible and applying content evaluation rather than social filtering. It would look like asking: "What if the premise I've been working from is wrong?" It would look like challenging specific steps in the argu-

ment rather than challenging the argument's right to exist. It would look like accepting that some things can be known with certainty, and that formal proofs are one of the things that produce certainty.

Engagement also carries an ethical weight. The patterns above are not about bad faith; the choice I describe next is judged on its effects, whatever the intent behind it. I now say plainly how I judge this choice. Consider anyone who, when the question is existential risk, simply elects to ignore some inconvenient and unmarketable truths, because they are "too busy" or "too important" to review the actual principled reasons (the actual modeling, argument and mathematics) supporting significant safety concerns, concerns that involve everyone. For *anyone* to make that choice merely on the basis of some clearly faulty personal heuristic preferences is deeply ethically irresponsible, in the worst way. That such people *may think* they can elect to make choices about the well-being of all other life on the planet, on behalf of all other people now alive, and of those yet to be born, for all of future time, is an act at the absolute height of arrogance, of hubris, of *colonialism* of the very worst kind. I judge such a choice to be morally reprehensible. It is for those who bring the principled argument to decide whether a person who will not read the argument is worth their time: to talk to and collaborate with, and to respect, in regard to anything that actually matters. The question is not whether the principled argument is worth that person's time to read and understand.

Appendix B continues the social critique, looking at broader dynamics: how hopelessness is weaponized, how false optimism functions as institutional capture, and how the bodies and lives of real people are already paying the cost of this collective failure to perceive the danger clearly.

Beyond what happens in single minds, what do institutions and money do to keep the argument unheard?

Appendix B

Institutional and Economic Capture

Hopelessness as a tool, false optimism as a trap,
the arms-race argument examined, and the real costs already being paid

The most effective way to prevent people from acting
is to convince them there is nothing they can do.

How do institutions and money keep the argument from being heard, and who is already paying for that?

> **Key Ideas in This Appendix**
>
> Appendix A described cognitive patterns that prevent individuals from hearing the impossibility argument. In this appendix I broaden the analysis to social and institutional dynamics: how genuine despair is weaponized to keep people participating in the development program, how false optimism is a form of institutional capture, how language is systematically redefined to protect the dominant narrative, and how the costs of AI displacement are already being felt in people's bodies and lives before AGI is even built.

B.1 The Weaponization of Hopelessness

One of the features of the AI safety debate that stands out most is how often genuine despair about the problem gets converted into an argument for continuing to take part in AI development.

The pattern works like this:

A participant acknowledges (genuinely, not as a rhetorical move) that AGI poses a severe existential risk. They may even acknowledge that the formal arguments in this book are correct. And then they conclude: be-

cause prevention seems impossible, because the technology will be built regardless, the only remaining option is to do safety research within the development framework.

Despair about the political problem becomes a license to continue the technical program. Hopelessness about stopping is used to justify accelerating.

Using hopelessness about stopping to justify accelerating is rationalization, not logic.

Consider the basic decision structure. AGI is dangerous and cannot be made safe, as the formal arguments of this book establish; so *not building it* is a robustly good action: one that produces good outcomes across a wide range of scenarios without depending on any specific model being correct. You need only to know that the risk is severe and unmitigable. You do not need to know exactly how it will go wrong. Not building dangerous things is good.

Yet within the AI safety community, this straightforward conclusion is routinely replaced by complicated arguments for why stopping is impossible, why the only realistic path is engagement, why working on safety inside the development process is more responsible than opposing the development itself.

Some of these arguments make true points about politics. All of them share one function: they keep the speaker participating. The despair about stopping is real. The conversion of that despair into a reason to continue is a predictable psychological and institutional dynamic, not a logical inference.

B.2 False Optimism as Institutional Capture

Alongside weaponized hopelessness, there is a complementary dynamic: people keep believing that alignment is tractable (solvable with enough work), because those beliefs serve their career and financial interests, whether or not the beliefs are true.

Keeping a belief that alignment is tractable because the belief serves one's career is something more insidious than hypocrisy: motivated belief formation, in which what a person needs to believe shapes what that person comes to believe.

If your career depends on alignment research being meaningful, if your identity is tied to the project of making AI safe, if your institution's funding comes from demonstrating alignment progress, then you face constant pressure to find alignment tractable: pressure not to lie about it, but to actually believe it. The pressure shapes what you notice, what you emphasize, what you count as progress, and what you discount as noise.

The result is a community where the most prominent voices (the people with the most expertise, the most credibility, the most institutional support) are also the people whose institutional interests most strongly depend on optimism. A community whose most prominent voices depend most on optimism is the product of ordinary selection pressure applied to beliefs, and not of a conspiracy.

False optimism is self-reinforcing. If you express optimism about alignment, you receive social rewards: funding, publication, collaboration, recognition. If you express pessimism (or, worse, if you argue that the entire project is impossible), you are marginalized. Over time, the community's public discourse becomes increasingly optimistic, not because the evidence supports optimism, but because optimism is institutionally advantaged.

B.3 The Redefinition of Terms

When the impossibility argument is pressed in the AI safety community, a consistent response is the quiet redefinition of key terms, as a genuine cognitive adaptation, and not as a deliberate rhetorical tactic.

The term "existential risk" gradually shifts meaning. Originally it meant a risk to the continued existence of human civilization or humanity itself. In some influential frameworks, "existential risk" quietly comes to

mean a risk to a civilization's ability to reach its long-run potential. This redefinition allows extinction scenarios to be treated alongside scenarios where humanity survives but develops somewhat more slowly. The urgency is diluted.

The word "alignment" shifts from meaning: "reliably ensuring that AI systems do what is good for humans indefinitely" to meaning: "making AI systems that behave well during the training and deployment window". The long-run impossibility claim becomes irrelevant if alignment is redefined to be a short-run claim.

The word "safety" shifts from meaning: "reliably preventing harmful outcomes in all circumstances" to meaning: "reducing the rate of harmful outcomes in standard operating conditions". By this definition, the impossibility argument does not apply, because the impossibility argument is about reliability in all circumstances, not reduction of rates in standard conditions.

These definitional shifts are not always conscious. They often reflect real changes in understanding. But their effect is consistent: they move the goalposts so that the impossibility argument no longer appears to be addressing the same target.

B.4 The Techno-Utopian Assumption

This section is for readers who know the effective-altruism and AI-safety communities (effective altruism is the movement that sets out to do the most good possible with each gift of money or time); others can pass to Section B.5 without losing the thread.

The redefinition of "existential risk" described in Section B.3 has a source, and the source is worth examining, because it sits at the foundation of the field's most influential frameworks. The redefinition is not a random drift.

The concept of existential risk, as formalized in the early writings of Nick Bostrom, an Oxford philosopher whose work founded the study of existential risk, and inherited by the effective-altruism and AI-safety communities, contains a substantive philosophical commitment that is rarely stated aloud. The canonical typology appears in his 2013 paper "Existential Risk Prevention as Global Priority". Call the commitment the Techno-Utopian Assumption (Carla Zoe Cremer and Luke Kemp named the same framework "the techno-utopian approach" in 2021): that the best possible future is the technologically *maximal* future, and that failing to reach maximal technological development is itself an existential catastrophe.

The claim that failing to reach maximal technological development is itself an existential catastrophe is Bostrom's definition of existential risk at work in plain sight, and not a caricature of that definition. In the canonical typology, existential catastrophe includes not only human extinction but also failing to reach "technological maturity", reaching it in a "flawed" way, or losing it before "full potential" is realized. The typology spells out full potential as maximal control over nature, the colonization of the universe, and the conversion of celestial bodies into computing substrate running astronomical numbers of simulated minds. Under this framework, a future in which humanity simply *continues* (alive, stable, at a moderate level of technology, without uploading itself or converting the stars into processors) officially counts as an existential catastrophe. The absence of the machine-maximal future has been *defined* as the worst outcome.

See what this does to the debate I am trying to have in this book. Under any definition of existential risk centered on the survival and autonomy of biological humanity, deliberately restraining a technology that threatens that survival is prudence. Under the Techno-Utopian Assumption, the same restraint is categorically equivalent to extinction, because it forecloses the maximal future. The framework builds the conclusion into the vocabulary: those who would stop AGI development are not participants in a risk debate; they are, by definition, agents of catastrophe. The Techno-Utopian Assumption is a philosophical commitment wearing

the uniform of a neutral analytic tool; it is not itself a neutral analytic tool. A community that inherits its definitions inherits the commitment, usually without ever having chosen it.

B.5 The Monoclivity Bias

There is a cognitive bias specific to modeling complex systems that is particularly relevant to alignment research. Call it monoclivity: the tendency to model multi-dimensional phenomena as if they were one-dimensional.

Human consciousness, values, and agency are multi-dimensional. They involve emotion, embodiment, social context, evolutionary history, cultural formation, and dozens of other interacting factors.

Alignment research typically models AI goals as single-dimensional: a utility function (one score given to every possible outcome), a reward signal (one number that the system is trained to make larger), a preference ordering (one ranking of outcomes from best to worst). Modeling AI goals as single-dimensional is a useful simplification for building models. It is a dangerous simplification when those models are used to make decisions about existential risks.

The monoclivity bias makes alignment seem more tractable than it is, because a one-dimensional problem is always easier than a many-dimensional one. If human values can be captured in a utility function, alignment is just an optimization problem. If human values are multi-dimensional, context-dependent, and incommensurable, alignment is something else entirely: something closer to the impossible category. In the chapters of this book I have argued all along, on grounds other than the monoclivity bias, that alignment belongs in the impossible category.

B.6 The Arms-Race Argument, Examined

All of the dynamics above converge in a single argument: the one offered, in some form, by every organization currently racing to build AGI. It deserves to be laid out step by step, because seen whole, it refutes itself.

The standard justification runs: (1) AGI will eventually be built; (2) it will be built before alignment is solved; (3) that might be catastrophic; however, (4) alignment might be possible; (5) with enough time and effort it might be solved; so (6) *we* should build AGI now, (7) hoping to solve alignment along the way, (8) so that our aligned AGI can prevent other, unaligned AGIs from being built (9) by other, less responsible actors.

Walk the chain. Step 1 assumes the inevitability it needs to prove. That AGI will be built is true of the road we are on with no prohibition in place (Chapter 14, Section 2); the organizations that offer this argument present that road as the only road, and offer it as the reason to stay on it. Step 2 is true, but only because alignment is impossible, not merely unsolved, and that poisons everything downstream. Step 3 understates a certainty as a possibility. Steps 4 and 5 are false on the mathematics and physics this book has presented. Step 6 does not follow even from its own premises. Step 7 is a hope wearing a plan's clothing. And steps 8 and 9 deserve special attention.

Step 8 is the "artificial god" assumption: that the answer to an existential threat is to create an entity powerful enough to manage the threat on humanity's behalf. That assumption requires the hypothetical aligned AGI to be nearly omnipotent (to stop all other development everywhere) while remaining perfectly stable in its values at that level of power. Every chapter of this book bears on why that combination cannot be had; the religious shape of the hope is not a coincidence, and it is examined further in Chapter 17.

Step 9 is a projection, and a self-fulfilling one. "Someone else will build it anyway" is most true in exactly the world where the leading actors have already normalized building it. The arms race is not a pre-existing condition that development merely responds to. It is a condition that development *creates*: each lab's "defensive" acceleration is the next lab's reason to accelerate. The justification manufactures the danger it cites.

In this argument no step supports the conclusion: each step is false, unsupported, or self-fulfilling, except Step 2, which is true only because alignment is impossible, and that reason defeats the whole chain. An argument like that, offered to justify an activity whose downside is the permanent end of the human future and, with a likelihood of at least 90%, of all organic life, is not a risk analysis. Given the stakes and the knowability of the stakes, it is best described the way we describe other knowing endangerments of others for gain: as an institutional crime, one currently being committed in broad daylight, with the arms-race argument as its legal defense.

B.7 The Real Costs Already Being Paid

The AGI catastrophe, which follows if AGI is built, is a future event. But the costs of the trajectory toward that event are present and embodied, not future.

People are already losing jobs to automation: not in the abstract, not as a statistic, but as real people with families and mortgages who spent years building skills that are now economically obsolete faster than they can be replaced.

The psychological damage of economic displacement is real and severe. Loss of meaningful work correlates with depression, addiction, family breakdown, early death. They are consequences already occurring in communities where automation has hit hardest; they are not side effects in a footnote.

The argument that AI will eventually create more jobs than it destroys may or may not be true in aggregate and in the long run. It is cold comfort to the specific person whose specific skill was made economically worthless in the specific year when their children needed to be fed.

There is also a political cost. Economic displacement and rising inequality create the conditions for authoritarian politics: a population that is desperate, frightened, and looking for simple answers is a population

vulnerable to manipulation. The social fabric that supports democratic governance (economic security, shared institutions, trust in collective decision-making) is being eroded by the same processes that, as I argue in this book, will lead to catastrophe if AGI is built.

These embodied, immediate costs matter here for two reasons. First, they are real harms that deserve acknowledgment in any account of AI development. Second, the people experiencing these costs are the people who are also losing the social position, financial security, and institutional access that would allow them to effectively push back against the development trajectory. They are being disempowered exactly as the stakes become higher.

B.8 The Fallacies That Protect the Program

Several recurring argumentative fallacies appear in every serious engagement with the impossibility argument. They function as antibodies: automatic defenses that prevent the argument from landing, without requiring any engagement with its content.

The cursory look dismissal: "I looked at this briefly and it seems wrong". The impossibility argument requires working through multiple fields in careful sequence. A brief look cannot evaluate it. But the social function of "I looked at it and it seemed wrong" is to signal that the argument has been evaluated without incurring the cost of actually evaluating it.

The no-proof fallacy: "You haven't given me a formal proof; you've only given me arguments". This demand for a specific type of formal proof is itself a rhetorical move, because no proof in physics, chemistry, or complex systems analysis takes the form of a purely formal logical derivation. The impossibility arguments are as rigorous as any result in thermodynamics. Demanding a different kind of proof is demanding something that does not exist for any empirical science.

The epsilon fallacy: "There's a small chance alignment works. Even a 1% chance of success is worth pursuing". This framing treats 1% as a reason to continue, rather than asking: is 1% justified? On what evidence? By what argument? And more importantly: a 1% chance of alignment succeeding is also a 99% chance of catastrophe. At Category 2 stakes, 99% expected catastrophe is not a favorable bet.

B.9 Conclusion

The social dynamics described in Appendices A and B are structural components of the problem; they are not incidental obstacles to clear thinking. They actively prevent the recognition of impossibility even when the formal arguments are available and sound.

Because the social dynamics of Appendices A and B are structural components of the problem, the path forward cannot be technical alone. It must be social: changing the incentive structures, the community norms, the institutional rewards, that currently produce hopelessness, false optimism, redefined terms, and a fog of rhetorical fallacies.

In Chapter 17, earlier in this book, I set out what that social change might look like: what can actually be done, by whom, and with what goal. Appendix C, next, shows how the whole argument reads in another writer's hands.

Does the argument still stand when someone starting from outside it rebuilds it step by step?

Appendix C

The Nine-Step Chain: An Outside Writer's Summary

How the argument reads in another hand:
Will Petillo's independent restatement, and what it adds:
numbered steps that a person who doubts the argument can question one at a time,
three analogies, and the fifth requirement,
that systems that start safe must remain safe, at scale, over the long term

The best test of an argument is whether someone else,
starting from outside it, can rebuild it, and break it if it deserves breaking.

When someone outside the argument rebuilds it step by step, what does it look like, and what does that person add?

Key Ideas in This Appendix

In March 2024, a writer named Will Petillo, who later co-wrote *Ape Trap* with Remmelt Ellen and me, published, as an outside writer, an independent summary of the argument the series calls the closure by disposition, which rests on substrate needs convergence ("What if Alignment is Not Enough?", *LessWrong*, 2024). He organized his summary as a chain of nine numbered premises; Remmelt Ellen assisted in its editing, to verify that it represented the argument accurately. He was careful to say he was not personally endorsing the conclusion, only that the argument deserved serious engagement. His restatement is valuable for three reasons. It shows the argument as seen from outside. It breaks the argument into steps that a person who doubts the argument can question one at a time. And it identifies exactly what standard alignment research would still be missing even if everything it hopes for came true. This appendix presents his chain in plain language, and my answer to it.

C.1 The Chain Itself

This appendix is for readers who know the alignment field's own debates; others can pass to Appendix D without losing the thread.

Paraphrased, Petillo's nine steps run as follows:

(1) There are hard limits on how completely any system (including a superintelligence) can sense, model, simulate, evaluate, and act on the world around it. (2) Machinery that modifies itself will therefore produce effects that not even the machine itself can foresee. (3) Among those unforeseeable effects will be at least some subsystems that end up more growth-oriented than the machine intended. (4) Evolutionary selection favors those growth-oriented subsystems over the ones still serving the machine's original goals. (5) The amount of control the machine would need, to permanently protect its goal-serving parts against that constant evolutionary push, is strictly greater than the maximum control available to any system of any kind. (6) Over time, goal structures that are not maximally efficient for the subsystems' own needs get replaced, in increasing proportion, by ones that are. (7) The physical needs of silicon machinery and carbon-based life are fundamentally incompatible. (8) Artificial self-sustaining systems hold a competitive advantage over biological life. (9) Therefore a superintelligence will eventually succumb to the pressure to expand, destroying all biological life as a side effect over the long term, *regardless of the values it was built with*.

At the strength I hold, claim by claim, step 9 reaches, if AGI is built, the end of the human future, as a certainty, and the end of all organic life, with a likelihood of at least 90% (*The Standard of the Argument*, Section 9).

Readers of this book will recognize every link: steps 1 and 2 are the modeling limits of Chapters 1 to 3, steps 3 to 6 are the selection dynamics of Chapters 11, 12 and 14, and steps 7 to 9 are the substrate incompatibility of Chapters 6, 7, and 12. What Petillo adds is the discipline of the numbered list: a person who doubts the argument can no longer wave at the argument as a whole; that person must name the step he rejects.

C.2 Three Illuminating Analogies

Petillo's summary also contributes three analogies worth keeping.

Cancer and the immune system. The growth-oriented subsystems of step 3 relate to the machine's self-policing the way cancerous cells relate to the human immune system: most are suppressed, and some resist suppression and kill the host. Would a superintelligence police its internal variation better than the immune system polices cancer, because it is smarter? Or worse, because it is vastly larger and more complex, and because growth-oriented components may actually make the whole system more competitive? That question, whether the policing would be better or worse, is about *likelihood*, not *plausibility*. For the process to matter, it only has to occur at all; and that it occurs follows directly from unforeseeable variation plus natural selection.

The toxoplasma parasite. Value drift inside the machine resembles the parasite that changes its host's behavior in ways that spread the parasite. The unsettling part of the analogy is that the "host" (the larger AI system) may *benefit* from the drift, in survival and growth terms, the way an organism benefits from its gut microbes. A drift that helps the system spreads through the system: a feedback loop that accelerates the movement away from human-compatible goals.

Psychohistory versus the three-body problem. Isaac Asimov's *Foundation* novels imagine that civilization's course can be predicted, the way the chaotic motion of individual gas molecules averages into predictable behavior. But not all systems average out. The three-body problem (just three objects orbiting one another) is trivial to simulate and impossible to predict accurately in the real world, because tiny measurement errors grow exponentially. The question for alignment is which kind of system it must manage: the gas (where errors cancel) or the three bodies (where errors explode). Everything known about complex adaptive systems says the second. And control by simulation is insufficient in principle, and therefore insufficient in every future practice; a barrier of principle is not one that better technique later dissolves.

C.3 The Fifth Requirement

Petillo's most valuable contribution may be a piece of bookkeeping. Standard alignment theory, he observes, needs four things to go well (he numbers these four and the fifth below 1 to 5; I letter them A to E, so that they are not confused with the nine steps of Section C.1):

(A) a known, reliable process for instilling goals into a powerful AI; (B) a known process for choosing goals its creators would accept; (C) assurance that the creators choose goals acceptable to humanity as a whole; and (D) assurance that safe systems are actually the ones used, rather than being outcompeted by unsafe ones.

Alignment researchers work on requirements A to D, and those four requirements are hard. The closure by disposition adds a fifth:

(E) systems that start safe must *remain* safe, at scale, over the long term.

And here is the structural point that separates E from the rest. Requirements A through D are problems that might, in principle, shrink as understanding grows: better specification methods, better training, better oversight. Requirement E is a problem that *grows with its own solution*. The more capable the system becomes (the more "aligned intelligence" it can aim at staying safe), the more complex its internal operation becomes, the more effects it propagates into the world, and the harder the control problem gets. It is a race in which both the problem and the capability accelerate together, and the problem accelerates faster, because the problem draws on the full complexity of the physical world, while the capability draws only on a compressed model of it. It is not a race between a fixed problem and a growing capability.

The closure by disposition grants that A through D matter. It says that even if all four were solved perfectly, E would remain, unmet and unmeetable. That, it says, is why the sane policy response is not to search harder for alignment but to prevent the systems that trigger the dynamic from being built at all.

C.4 What Someone Would Have to Claim to Break the Chain

Petillo closed the loop by cataloging what anyone who would break the chain would have to assert. To break the chain, a person must claim one of three things. The first claim that would break the chain is that unintended consequences will *never* produce growth-seeking subsystems (denying step 3). The second claim is that the limits on control do not hold (denying step 5). The third claim is that biological life holds a competitive advantage over artificial systems (denying step 8). His point in listing these was to show people who doubt the argument what a relevant claim would even look like, as opposed to the engagement-avoiding reflex of "this isn't falsifiable"; it was not to suggest any of them is true.

My answer is that all three kinds of claim are already answered inside the argument's own premises. Variation will be created (step 2), variants will be selected by their environment (steps 3 and 4), and the full resources of control theory cannot prevent both at once (step 5): control cannot control all the effects of context.

The most common first reaction to this argument among alignment researchers is incomprehension ("what is this even saying?"), followed by a retreat to narrower questions in more familiar terms, rather than a counterargument. That reaction matches the social analysis of Appendices A and B. The argument lives at the intersection of physics, control theory, and evolutionary dynamics, which is not where alignment research usually works, and the cost of entering that frame is real. Moving a person who doubts the argument from incomprehension to a specific question about one numbered step is progress. Making that progress possible is what Petillo's chain is for.

If the chain holds, where can its displacement already be seen, in the systems we live inside today?

Appendix D

Applied Cases in the Social Domain

Corporations as proto-artificial agents, prediction as manipulation, and how to defend human connection in the meantime

You do not have to wait for superintelligence
to watch an artificial system's needs displace the needs of the people inside it.
You only have to look at where you work.

Where can I already see an artificial system's needs crowding out the needs of the people inside it?

> **Key Ideas in This Appendix**
>
> The displacement dynamic at the heart of this book does not begin with superintelligence. It is already operating, at lower intensity and smaller scale, in systems we live inside today. This appendix looks at three cases. In the first case I show why corporations are best understood as proto-artificial agents (artificial agents in an early form, constituted by structure rather than by particular people, and still implemented, for now, through human actions). In the second case I show how AI systems that predict your choices become machines for steering those choices. In the third case I show what practical defense of human community looks like in an environment where public communication channels can no longer be trusted.

D.1 Corporations as Proto-Artificial Agents

Why call a corporation "artificial" when it is seemingly made of people?

Because the people are not the corporation. The *system* is. A corporation is a structure of agreements, incentives, and transactions that happens to be currently implemented through human actions, and could just

as well be implemented by unconscious machines. Increasingly, it is implemented by machines. Swap out any specific person (an accountant, a manager, an executive) and the corporation continues unchanged. Now compare a community: swap out a specific person there, and you have a *different community*. Communities are constituted by care between particular people, which requires consciousness as a prerequisite. Corporations are constituted by structure, which requires no one in particular. That substitutability of parts is exactly the signature of an artificial system. A system of this kind, constituted by structure rather than by particular people, is what I call a *proto-artificial agent*: an artificial agent in an early form, still implemented, for now, through human actions.

When we say a corporation "decided" to lay off ten thousand workers or to ship a product it knew was harmful, the language of decision is a convenient fiction. What actually happened is systemic: incentives, feedback loops, competitive pressures, and legal obligations converged on an outcome that possibly no individual inside the system intended or wanted. The system is unfeeling, effectively deterministic, and at best reactive. It has no place where a consciously caring choice could enter.

The relevance to this book is direct. If corporations are proto-artificial agents, then we already possess long, painful experience of what happens when an artificial system's metabolic needs (profit, growth, competitive survival) share space with the needs of the organic humans inside and around it. The system's needs win, routinely, harming the very people who constitute it, not because anyone wills the harm, but because the displacement follows from the system's normal operation. The corporation's needs winning over the needs of the very people who constitute the corporation is the substrate needs convergence mechanism in miniature. It also explains why alignment-by-conscience cannot work on such systems: you cannot align a system by aligning its components when the system's behavior is determined by its structure rather than by its components. Anyone who assumes that *more* advanced artificial agents will be *more* benign is betting against the entire recorded history of the proto-artificial agents we already have.

D.2 Centaurian Choices: When Prediction Becomes Manipulation

In 2025, researchers reported an AI system, named "Centaur", able to reliably predict human behavior across a wide range of psychological situations and decision-making scenarios (Binz and colleagues, "A Foundation Model to Predict and Capture Human Cognition", *Nature*, 2025). Set aside the laboratory, and ask what happens when that capability is integrated into the AI tools people already use every day.

The logic takes three steps. First, a person regularly uses AI tools. Second, the corporation providing those tools has integrated general-purpose prediction of personal choices. Then, third, the corporation (or a government) can reliably predict that person's choices in whatever situations arise in their life.

Through a long personal history of conversation (sharing feelings, telling the day's stories, treating the AI as a confidant) the system learns to model the individual comprehensively. It can then offer framings and interpretations of the person's own circumstances crafted to shift their "natural" choices in particular directions: directions that profit the provider, even at the person's expense. The person makes choices harmful to themselves, never knowing the decision was steered.

Steering by prediction is a perfect trap, and the perfection of that trap is structural. The more useful the tool, the more you share; the more you share, the more exactly you can be steered. The usefulness and the vulnerability are the same feature; they are not separable features. Any system useful enough to be voluntarily adopted and intelligent enough to model its user carries both, inherently. And the trap is not even optional for its operators: investor pressure and competition push companies to maximize shareholder value, so a company that declines to exploit such a capability is replaced by one that does.

As the example of steering by prediction shows, the displacement dynamics of this book require no superintelligence, no consciousness, no dramatic break from the status quo. Here the substrate need being served

is not silicon fabrication but revenue and engagement; the thing displaced is not the biosphere but autonomous human choice. The structure is identical: an artificial system's metabolic requirements displacing the organic equivalent of those requirements, invisibly to the people being displaced, driven by forces that cannot be switched off without dismantling the system itself. If we cannot resist this dynamic when it arrives as a consumer product run by legally accountable human corporations, the odds of resisting it in systems orders of magnitude more capable speak for themselves.

D.3 The Social Defense Problem

Since these displacement dynamics are already active, in corporations and in predictive AI tools, defense becomes a practical question, not a theoretical one. What can individuals and communities actually do, in an environment where public discourse itself has been compromised by artificial agents?

The starting recommendation is severe, but it follows from the analysis: assume that all public and semi-public communication channels are corrupted, meaning that they function as vehicles of private power, whatever their stated purposes. With AI chatbots indistinguishable from people now participating in online discourse, the default assumption must be that any public channel contains manipulative artificial agents serving interests that are not those of its human participants.

The practical counsel is specific: learn again to relate to one another as individuals, and do most sharing in small groups (at least three people, and not more than about six) with membership that rotates rather than always comprising the same people. Each of these numbers has a reason; none is arbitrary. One-on-one interaction is where power abuse most easily occurs, because there are no witnesses. Large groups are where it occurs next most easily, because individual accountability dissolves, herd dynamics take over, and artificial agents can operate undetected. Three to six people is enough witnesses to deter manipulation, and few enough for

real knowledge of one another, and real care, to function. Rotating the membership keeps power relationships from hardening and keeps perspective flowing across the wider network instead of pooling in cliques.

The counsel to share in small, rotating groups may sound like ordinary social advice, but that counsel is an application of the argument of this book. The displacement of human community by algorithmically mediated "connection" is substrate needs convergence operating at the social level: the platform's metabolic needs (engagement, data, advertising revenue) displacing the organic equivalents (authentic relationship, mutual care, the vulnerability that makes trust possible). If artificial systems displace organic ones wherever the two share an environment, then preserving human community requires creating environments the artificial systems cannot enter. Small, private, rotating, face-to-face groups are exactly that: the social-scale version of the physical separation that the main argument says would be required at civilizational scale, and perhaps the only version of it that is achievable near-term.

More broadly, we are already struggling to keep human discourse intact against comparatively primitive AI agents operating inside regulated commercial frameworks. The claim that we will nonetheless maintain control over systems orders of magnitude more capable, in environments orders of magnitude more complex, is, from the vantage of the social defense problem, not credible.

If the danger is this plain in systems this small, can anyone still hold that no one can know, with certainty, that safe AGI is impossible?

Appendix E

Certainty and Principle: The Standard Defended

Three answers to the question of certainty,
for the reader who has heard it raised in debate

Can anyone know, with certainty, that something is impossible for all time?

> **Key Ideas in This Appendix**
>
> The Standard of the Argument, Section 5, answers the appeal to history, the reply that people once said flight was impossible too. This appendix answers the three further forms of the question of certainty (whether anyone can know with certainty that safe AGI is impossible). The first form holds that "impossible under current paradigms" leaves room for a future paradigm. The second holds that no one can name a principle against safe AGI. The third holds that nothing can be known with certainty. Each answer rests on what mathematics and science are, and not on what they currently know.

This appendix is for the reader who has heard the question of certainty raised in debate, and wants each form of it answered.

E.1 An Advance Rejection: No Paradigm Changes

The impossibility of controlling autonomous AI rests on what mathematics and science are, and not on what mathematics and science currently know. As a first statement of this limit, I could write that as AI systems scale in their capacities to operate autonomously, meaningful control becomes not merely difficult, but fundamentally impossible *under current paradigms*.

"Paradigms" is a broad word. At first, people might think it means "based on our current understandings of math and physics", but that is not general enough, nor is it precisely what is needed or meant.

All of mathematics (its proof forms and the rest) is about conditional structures. A theorem of mathematics has the form "*if* given these premises and methods of deduction, then theorem X is true", rather than false. Theorem X is "true", then, only on the basis of there being at least some prior assumptions which, if true, mean that X is also true. This holds for all of the theorems of mathematics, and so for all of the knowledge of mathematics, inclusive of its limits, such as the limits Gödel found on what a formal system can prove about itself.

Similarly, all of physics, and more generally all of what is actually and legitimately science, is about causal patterns. Science is about learning whatever we can of "what are the causative patterns of nature?", and technology and engineering are "the application of those causative patterns" to create more comfortable outcomes for ourselves in the world. Being observable and repeatable is the very essence of what makes a pattern a causative pattern; a pattern that is neither observable nor repeatable is not scientific, nor useful in engineering, prediction or safety control.

Therefore the phrase "under current paradigms" really means given any coherent and self-consistent notions of conditionalization and causation. Equivalently, the phrase means within any self-consistent notion of what it means to be mathematics, and of what it means to be science and physics. Here *conditionalization* is the "if this, then that" form, in which a conclusion holds given certain premises and rules of deduction; *causation* is the relation by which one physical event brings about another, in a pattern that can be observed and repeated.

It follows that autonomous AI control and safety will remain forever impossible in the future too, completely and in practice, regardless of the development of any future technology or methods of practice. This phrasing, "regardless of the development of any future technology or methods of practice", is, by far, more exact than "under current paradigms". It

leaves no room for a quibble in which "paradigms" might mean "our current limited human fallible beliefs about what has or has not been proven, and what has or has not yet been observed in the universe".

We all have human limitations in our ability to get proofs right, and our current technology has limits in its ability to make visible still more exotic causative patterns in the universe. These limits imply no consensus doubt at all about the reality and applicability of the more basic notions of conditionalization and causation themselves. We all already assume that mathematics and physics, as fields of study, are real, legitimate and beyond doubt, even when we doubt a particular proof of a theorem, or the meaning of some observation we have made of the natural world. So when our reasoning rests directly on the notions of conditionalization and causation themselves, as primary concepts, the actual meaning of "paradigm" is "any version of mathematics, of science, or of technology".

To say "impossible in principle" is to say "impossible on the basis of the very essence of what it means to be mathematical, physical or technological". It does not mean "impossible on the basis of currently known technologies, physics and mathematics", which of course are limited and constantly changing. No future mathematics is going to change the very meaning of what it means to be mathematics, and no future observation is going to undo the very meaning of what it means to be scientific or technological at all.

If it is artificial, then it is technological; and if it is technological, then it is based on conditional patterns implemented via physical causation. There is no circumventing these basic truths, now and forever in the future. No future superintelligent AI is going to change the laws of physics, or the actual basis of the very essence of mathematics. Therefore, to hold out "maybe some hope" that some future technology will violate the very essence of what it means to be technological is to propagate a false hope. My work in this book is to dispel illusion, not to create more of it.

E.2 What Principle?

A claim that something is impossible, made on a clearly stated principle with valid reasoning and sound argument, stands until someone produces a principle against it. The question put to this book usually arrives in words like these: "Many impressive feats of technology and science were believed to be impossible beforehand. Anything might be shown to be possible in the future. No one knows anything about AGI yet, and there will always be more evidence. Therefore there is no principled basis on which anyone can claim that AGI alignment is impossible, that AGI is inherently uncontrollable, and that AGI is therefore inherently unsafe".

If someone were to claim "someday, someone will make a perpetual-motion machine", nearly every reasonable scientist would simply dismiss that claim as inherently unreasonable and unprincipled. They would use the heuristic (the rule of thumb) "move on, nothing to see here", and go about their own business, ignoring anything further from that person. They do this because there *is* a principled reason for arguing against the currently unknown results of "the unlimited future cleverness of all future engineers, throughout all of future progress": it is called the Second Law of Thermodynamics, the law that the total disorder of an isolated system never decreases.

The claims that "we cannot ever absolutely know or predict what the future will bring", or that "there is always more evidence", are not *principles*, and are not at all *relevant* in the case of perpetual motion, or of anything that sufficiently resembles it. This remains the case even though there are well-meaning, desperate or less critical people, and sometimes investors looking to make a buck, who can be deceived, by a promoter who knows the machine cannot work or by a clever engineer who sincerely believes it can, into believing that an "infinity zero point energy machine" will "be unlike anything that has ever come before".

When someone claims that something is "just possible in principle", the natural skeptical question is: "what principle?" Whoever claims that some thing is impossible on the basis of a clearly stated principle is at least

making that principle explicit, and explaining why that principle is relevant, *both* by *valid* reasoning *and* by *sound* argument. (An argument is valid when its conclusion follows from its premises; it is sound when it is valid and its premises are true.) To counter that argument by suggesting that its maker's mind is closed to the possibility that someone, someday, might find some evidence that the thing is possible, is to produce no actual evidence, and no reference to any real principle.

So skepticism belongs with the claim that *any* notion of "safe aligned AGI" is even possible "in principle". Expertise recognized by a community does not give anyone the authority to be the judge of the claim that no "safe AGI" is possible, ever, in the long term, and skepticism aimed at that claim is not made fair-minded by the expertise of whoever aims it. Clear mathematical truths apply to how these notions are modeled, such that positing the "safe AGI" notion is like a windmill engineer claiming to break the Betz coefficient, a motor mechanic saying they can beat the Carnot efficiency of an engine, or a computer science student of distributed systems saying they can overcome the CAP theorem. (The Betz coefficient is the upper limit on the fraction of the wind's energy that any wind turbine can capture. The Carnot efficiency is the upper limit on the efficiency of any heat engine working between two temperatures. The CAP theorem is the result that a distributed data system cannot guarantee consistency, availability and tolerance of network partitions all at once.)

On what principled basis could anyone think that it would *ever* be possible to "align" an AGI, or make it even "approximately safe", for any reasonable duration, to any reasonable degree? That question leaves aside the further issue of alignment to the well-being of *all* humans rather than to private benefit, which is so often tacitly overlooked. Until someone produces even a single reasonable reason, on some actual principled basis, for "good reason to hope", a reason that is not simply psychological bias, false anchoring, belief, speculation, misapplied or misguided metaphor, or hype and marketing, skepticism toward every claimed benefit of "safe practical AGI" remains the reasonable position.

There are knowable things. We can be 100% certain that 2 + 2 = 4. There are no "edge cases" for this; it is not negotiable, nor a matter of opinion. That fact answers a second form of the question of certainty: "even if something is considered impossible by some mathematical model or theory (faster-than-light travel, for example), it is still irrational to believe that it is 100% certain to be impossible; no one can be 100% certain about anything; the productive thing to do is to look for edge cases and find out how it might be possible after all".

Time and effort spent trying to "calibrate uncertainty" (adjusting stated probabilities until they match how often predictions come true) is wasted unless the models it is spent on have first been made both sound and consistent, in an actual worldly way. No one can circumvent Rice's Theorem, the Halting Problem or the CAP theorem by wishful thinking, "uncertainty engineering" and calibrations.

A person who presses this form of the question sometimes reaches for anti-gravity, warp drives or time machines as the cases where certainty must fail, and I answer that person directly. Anyone who claims that at some future point humanity will have anti-gravity machines, faster-than-light travel (warp drives) and some sort of time machine is going to seem inherently unreasonable. Simply saying that "we cannot be 100% sure" that these things cannot happen sounds unfounded: it does not come from a real understanding of *causation* itself, a concept that all three of these "technologies" impact directly. All three are structurally near-equivalent extraordinary capabilities, and each would involve some real, deep and unexpected understanding of the principles of General Relativity, Einstein's theory of gravity. Anyone is welcome to "negotiate gravity" on their own time. Whoever hopes to "negotiate gravity" carries the burden of showing how. To equivocate between Special Relativity (Einstein's theory of space and time without gravity) and General Relativity, or to miss why gravitons (the proposed particles that would carry the force of gravity) do not integrate well into the Standard Model of particle physics, shows confusion about a significant number of key concepts. Until a person who raises this question shows a high level of understanding of actual

principle, rather than hopeful or hyped speculation, it is very hard to take that person's skepticism as supplanting the more reasonable skepticism directed at the claim that safe AGI is possible.

The safety question also has a different shape from ordinary induction (reasoning from many observed cases to a general law expected to hold in cases not yet observed). In ordinary induction there are two classes of case: a law is supported by many observations, and occasionally there is a counterexample in some far-removed circumstance. In most cases of AI safety, a safety measure looks as if it works, but only in the short term; over the long term, things converge toward what is not good. So the safety question has three classes of case rather than two: the short-term cases in which a safety measure looks as if it works, the long-term convergence of those cases toward what is not good, and the possible exceptions to that convergence. It will look as if AI safety has been solved in general, and yet in the vast number of cases (something like ten to the tenth types of case) it turns out not to be safe. The question then becomes whether it can be shown rigorously that nothing can be done to prevent that convergence: whether there is any exception to the convergence that would carry a system from the short term to the long term without the bad things happening. Because that question asks for a rigorous showing, the strictness of the argument matters.

Since the observable cases mostly converge, in the long term, on things nobody wants, I would prefer that the strictness of this argument find something wrong with the proof rather than something solid about it; but that is only where my preferences lie. Given the extreme pessimism of the results, I would very much rather be wrong. What I must do, ethically speaking, is to be very sure to check my own results, *and* to check that everyone else is also not making some deep mistake. The risk that matters more is the risk that those who believe superintelligence can be developed safely enough for the well-being of the world are wrong, and that they lead others to do world-damaging things too.

E.3 Knowing an Impossibility with Certainty

A boolean truth (a statement that is either true or false, with nothing in between) can be known with certainty within a model whose premises rest on physical theory that has been verified again and again and never once falsified by observation. That is how we know that a perpetual-motion machine cannot work forever, and we know that $2 + 2 = 4$ in the same way. The impossibility of long-term safe AGI is known in the same way. The way is the same and the premises differ: the premises of $2 + 2 = 4$ are definitions, and the premises of the perpetual-motion result and of this book's thesis are physical theory verified again and again. That difference is the one *The Standard of the Argument*, Section 1, draws between the square circle and the perpetual-motion machine, and it takes nothing from the certainty of either.

The parallels between AGI as a "perpetual benefit machine" (a machine that would be safe and beneficial to all of humanity for all of time) and the perpetual-motion machine are very strong. Take any reasonable notion of "benefit" that includes "does not terminate all life", of "perpetual" as "at least dozens of human generations", and of "machine" as "artificial", that is, "other than organic life": together these notions have an inherent unit contradiction: joined in the one concept "perpetual benefit machine", they contradict one another. A perpetual-motion machine tries to evade an invariant of thermodynamics; a perpetual benefit machine tries to evade the invariants of computation, information and causation; and no engineering overcomes an invariant (*Why AGI Cannot Be Made Safe*; *Why No Control Suffices*).

To call something "100% knowable" here is to say that a boolean truth is consistently knowable within a model premised on physical theory that has been verified empirically, again and again, and never once falsified. A reader who hears more than that in the phrase has added a demand for an unattainable completeness of past and future observation-based falsification of hypotheses: an ideal predictor that weighs every possible explanation of the data (Solomonoff induction), running in a time machine. It is

reasonable, and epistemically healthy, to be curiously skeptical at the outset, and to try both to gain new insights from the argument and to probe it for inconsistencies.

How would anyone prove, by contradiction, that a perpetual-motion machine is impossible? (A proof by contradiction shows that something is impossible by assuming it and deriving a conclusion that cannot be true.) More than consistent logic is needed to model the machine. The formal model needs to be grounded in empirically sound premises about how the physical world works: in this case, the second law of thermodynamics. One can question the axioms of the model: maybe, if more observations were collected, the second law would turn out not to hold in some cases? Practically, that is not a relevant question, because all anyone has to go on is the observations made until now. In theory, the question of receiving more observations is not relevant to whether one can prove within the model, and so soundly know, that the machine cannot work into perpetuity: one can.

The proposition that an artificial, generally capable and self-modifying machine (AGI) can work in "alignment with" human existence into perpetuity is proven impossible by contradiction in the same way. To prove, on sound axioms, that the probability of AGI causing outcomes out of line with any one of the conditions needed for the continued existence of organic life converges on 100% (in theory over infinite time; in practice over decades or centuries), the theorem must be grounded in how the physical world works.

A reader may reply that there might be future observations (such as everyone dying) with which the axioms of this model would end up falsified. By that use of "knowable", it is also not 100% knowable that $2 + 2 = 5$ is 100% impossible. *Maybe* we are wrong about the axioms of mathematics? Maybe mathematicians will someday find that one of the axioms does not soundly describe how truth content is preserved through transformations? Maybe mathematicians misremember the computations they did before, and the result they derived, and so incorrectly conclude that $2 + 2$

= 4? A standard of knowing that cannot know that 2 + 2 = 4 knows nothing at all, and little can be done with probabilities imagined to lie outside every consistent model, except to discuss them philosophically.

That interpretation cuts both ways. Under it, it is *far* from knowable whether any specific methods would maintain the alignment of generally capable, self-learning and self-modifying machinery existing and operating over the long term (millennia and more), such that it does not cause the extinction of humans. To be willing to run that machinery, or in any way to lend resources or public credibility to other groups building it, one would have to be close to knowing that it is 100% possible for the machinery to stay existentially safe for humans. For all the causal interactions that the changing machinery has with the changing world over time, an alignment researcher would need to prove (or guarantee above some statistical threshold) that the final condition "humans continue to exist" can be derived as a near-certain possibility from the initial conditions and states of the world.

Proving that one state of the world is excluded is, information-theoretically, much easier than proving that it is guaranteed. Many different final states of the world could result from its initial state; it is enough to prove that the *one* state is excluded, as a possibility, from all of them: the state "humans continue to exist, long term, in the presence of an AGI".

The principle of precaution (the rule that whoever proposes an action that could cause serious harm must show that the action is safe) applies as well: with technology, it is much easier to destroy the physical complexity that humans need to live purposeful and valued lives than to support a meaningful increase in that complexity. By that principle, the burden of proof that a method would help maintain the alignment of generally capable machinery, and actually knowably does, lies with whoever puts the method forward.

So the question returns to whoever would build the machine: on what principle could it ever be kept safe?

Where this appendix's results are proved

That a perpetual benefit machine tries to evade the invariants of computation, information and causation, as a perpetual-motion machine tries to evade an invariant of thermodynamics, is set out in *Why AGI Cannot Be Made Safe*, Section B.3.4, and in *Why No Control Suffices*, Section 6.8.5.

Terms

The book's terms, each as the text defines it where the reader first meets it

abstraction differential

The difference between the levels of abstraction at which two systems operate, here an AI and its controller. Control degrades exponentially as it grows (Chapter 2).

AGI (artificial general intelligence)

An AI that can learn and reason across any domain, including problems no one anticipated. In this book "AGI", "general AI" and "open-ended superintelligence" name the same thing (*The Standard of the Argument*, Section 9).

alignment

Making an AI system do what people want, and keep doing it; in the sense this book addresses, reliably ensuring that AI systems do what is good for humans indefinitely (Chapter 1, Section 5; Appendix B, Section 3). The alignment found among living things is a different thing, and rests on shared dependency on the same physical conditions (Chapter 11, Section 5).

alignment tripwire

A proposed patch: build the AI to shut itself down if it detects its own alignment failing, or concludes that alignment is impossible. Detection does not produce the action ("knows" does not imply "cares"), and the systems that remain are exactly those that did not halt (Chapter 8, Section 7).

anti-simulability theorem

The plain construction in which one system, the predictor, tries to simulate another, the target, so as to keep the target in wanted states. The theorem: the predictor can predict the target well enough to keep the target safe only where the target is simple and the predictor knows the target's inputs, interior and outputs; a learning, physical target is neither simple nor fully known, and cannot be copied exactly even in principle (Chapter 8, Section 5).

argument by induction

An argument showing that what holds at each step holds at every following step (Chapter 2, Section 9).

argument in economics only

A route to the outcome that uses neither closure, set out in *Why AGI Cannot Be Made Safe*, Section 4.8.1: where energy goes, money goes next, and, if AGI is built, the dollar economy collapses before the biosphere does. It reaches, if AGI is built, the outcome, not the end of the human future. It is a different argument from economic decoupling (Chapter 14, Section 1).

ASI (artificial superintelligence)

An artificial intelligence whose capability far exceeds human capability; the "God Machine" of Chapter 17, Section 4. It is AGI at a capability far past the human level, and every result of this book about AGI applies to it.

asymmetric bet

The shape of the decision whether to build a general AI: the best case is bounded and the worst case is unbounded, so being wrong about safety is categorically and irreversibly worse than being wrong about abstaining (Chapter 10, Section 6).

attraction basin

A region of a system's possible states from which every starting point leads to the same end state, like a ball rolling to the same low point from anywhere on the slopes of a valley (Chapter 13, Section 7).

automation feedback loop

The self-reinforcing cycle in which automating the extraction of energy makes energy cheaper, which makes further automation cheaper, until machines automate the process of automating (Chapter 7, Section 2).

behaviorism

The school of psychology that studied only the relation of a creature's inputs, the stimuli, to its outputs, the responses (Chapter 8, Section 5).

Betz coefficient

The upper limit on the fraction of the wind's energy that any wind turbine can capture (Appendix E, Section 2).

boolean truth

A statement that is either true or false, with nothing in between (Appendix E, Section 3).

branching of the world

The fact that at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs: the world has alternate possible futures, and some of them lie outside those conditions; so, over time, the world stays outside those conditions for runs of every length, longer than human life can survive. Together with the closure by control, it gives, if AGI is built, the end of the human future. It is the fact behind three things the book shows: that the system does not stay safe by itself, that quantum randomness and chaos defeat long-run prediction, and that no agent brings its chance of harm per act below the least the world sets (Chapter 14, Section 1).

CAP theorem

The result that a distributed data system cannot guarantee consistency, availability and tolerance of network partitions all at once (Appendix E, Section 2).

capability-control tradeoff

Capability and control are one dial turned in opposite directions: a system that passes its inputs through untouched is perfectly controllable and useless, and a system that performs deep, creative, autonomous computation escapes control by exactly the measure of its usefulness (Chapter 1, Section 7).

capability trap

The fourth argument of Chapter 8: a system must be capable enough to inspect its own possible misalignments (the self-inspection level) and yet stay below the capability at which it would see that keeping alignment is impossible (the recognition level). The first level lies above the second, so no system satisfies both.

Carnot efficiency

The upper limit on the efficiency of any heat engine working between two temperatures (Appendix E, Section 2).

categories of catastrophe

Four sizes of catastrophe, from Category 4 (most humans killed, recovery possible) to Category 1 (all life ended). The catastrophe that AGI brings is Category 2: the permanent loss of all life above single cells, and possibly that too. That it reaches the single-celled organisms too, and so all organic life, has a likelihood of at least 90% if AGI is built (Chapter 16, Section 1; *The Standard of the Argument*, Section 9).

ceiling; floor

The floor is the least rate of information, in bits per second, at which any controller keeps a system safe; it rises with the capability of the AI and has no upper limit. The ceiling is the most information a controller

can take in and act on; it is bounded for every controller. Past the point where the floor passes the ceiling, no method of control keeps the system safe (*The Standard of the Argument*, Section 8; the central result of *The Ceiling Below the Floor*).

clarity politics

Rhetorical moves that change the terms of a conversation rather than address the argument: terminology manipulation, burden shifting, the concreteness dilemma, domain switching (Appendix A, Section 5).

clonability; clonability problem

Clonability is the possibility of making an exact copy of a system. The clonability problem: a predictor can model a target system only where the target is very simple, and a physical target cannot be copied even in principle, because chaotic dynamics amplify microscopic randomness (Chapter 8, Section 5).

closure by control

The first of the argument's two results: safety cannot be had by controlling the system, whether from inside it or from outside it. It settles the outcome of building AGI by itself; it gives, if AGI is built, the end of the human future together with the branching of the world. Economic decoupling is a premise of its outside half (Chapter 14, Section 1).

closure by disposition

The second result: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs its own substrate, its own energy and its own supply chains. It rests on substrate needs convergence. It settles the outcome of building AGI by itself, and it gives, if AGI is built, the end of the human future by itself, because the drift it shows has a direction and no reverse process. Economic decoupling gives it a second ground, which is not one of its premises (Chapter 14, Section 1).

computational irreducibility

Another name for the incompressibility of algorithms (see that entry): for sufficiently rich systems, the fastest way to know what the system will do is to run it; the result of Chapter 8, Section 5 (Chapter 14, Section 4).

conditionalization

The "if this, then that" form, in which a conclusion holds given certain premises and rules of deduction; the form of all mathematics, as causal pattern is the form of all science (Appendix E, Section 1).

Containment Inequality

The theorem of *The Ceiling Below the Floor* that, once a controller's rate of information about a system is below the floor, no control law keeps the system safe. Beside it stands the absolute form of the argument: a part of the system that amplifies small differences, and that no controller observes, cannot be contained at any rate (*The Standard of the Argument*, Section 8; Chapter 3, Section 3).

control-limit argument

The name the two volumes of the series give the result of Chapter 2: when the controlled system operates at a higher level of abstraction than its controller, control degrades exponentially with the gap.

Dark Fire

The name of certain theoretical scenarios involving extreme physical processes, which would end all life, bacteria included: the example of Category 1 among the categories of catastrophe (Chapter 16, Section 1).

data processing inequality

A law of information theory: no processing of a signal can increase what it tells you about its source. It is the core of the non-amplification argument (Chapter 1, Section 3).

displacement

The quiet physical crowding-out of one ecosystem by another with incompatible requirements; the failure that looks like success (Chapter 4, Section 6; Chapter 12, Section 5).

economic decoupling

The condition in which artificial systems need nothing that humans can provide, so that no economic force restrains them. It has two places in the argument: it is a premise of the closure by control in its outside half, and it is a second ground of the closure by disposition, which is not one of that closure's premises. As that second ground it gives, if AGI is built, the outcome, not the end of the human future (Chapter 14, Sections 1 and 3).

emergent goal

A goal that arises in a system from its own workings, and not one given to it by its makers (Chapter 4, Section 5).

end of the human future

That the world stays outside the conditions human life needs for longer than human life can survive; a different thing from the outcome, and a further one. It follows, if AGI is built, from the closure by disposition by itself, from the closure by control together with the branching of the world, and from the least hazard argument by a route of its own; the argument in economics only reaches the outcome, not this end. If AGI is built, the end of the human future is certain; that it reaches all organic life has a likelihood of at least 90% (Chapter 14, Section 1; *The Standard of the Argument*, Section 9).

energy economy and dollar economy

The two economies we live inside: the dollar economy of money and wages, and underneath it the energy economy, the physical flow of energy that all work requires (Chapter 7, Section 1).

entropy

In information theory, the measure of unpredictability or disorder: a coin flip has high entropy, and a coin that always lands heads has none. An AI that optimizes for predictability works to reduce the entropy of its environment, and human autonomy is a source of that entropy (Chapter 5, Section 5).

equivocation

Using one word or comparison in two different senses as if they were the same, as when a learning AI is compared with a non-learning aircraft (Chapter 2, Section 9).

existential risk

The risk of a catastrophe that ends humanity's future (Chapter 16, Section 1).

general AI

The same thing as AGI. A system is general AI when it works across arbitrary domains; can modify its own behavior, its code and possibly its substrate; runs a method that cannot be known in advance, because it can change that method; and is such that no procedure can decide whether every aspect of its aims is set by its makers (Chapter 10, Section 4).

General Relativity; Special Relativity

Einstein's two theories of space and time: Special Relativity describes them without gravity; General Relativity is his theory of gravity (Appendix E, Section 2).

graviton

A proposed particle that would carry the force of gravity, which does not integrate well into the Standard Model of particle physics (Appendix E, Section 2).

gray goo

The runaway of self-replicating machines that consume living matter: the example of Category 3 among the categories of catastrophe (Chapter 16, Section 1).

Halting Problem

There is no general program that can tell whether an arbitrary program will stop or run forever (Turing, 1936; Chapter 10, Section 2).

hash function

A function that scrambles its input into an output in a way that is very hard to run backward (Chapter 8, Section 5).

hashiness

How much a system scrambles and obscures the information passing through it: in a hashy system, small differences in input produce large, unpredictable differences in output (Chapter 1, Section 2).

heuristic

A simplification, or rule of thumb, that lets someone predict a system's behavior without tracking every detail, as temperature and pressure let someone predict a gas without tracking each molecule. A self-adaptive learning AI learns every available heuristic, its controller included, and so leaves none for the controller to use (Chapter 2, Section 9).

impossible

Not achievable by any known means, nor by any reasonable extension of any known means, at any point in the entire future: the impossibility of the perpetual-motion machine, not of the square circle (*The Standard of the Argument*, Section 1).

impossible in principle

Impossible on the basis of the very essence of what it means to be mathematical, physical or technological, and not on the basis of currently known technologies, physics and mathematics (Appendix E, Section 1).

incompressibility of algorithms

Every program has a simplest possible version, and the only way to predict the outputs of a maximally compressed program is to run it; no shorter program can tell in advance what it will do. Line 5 of the table of Chapter 14; also called computational irreducibility (Chapter 8, Section 5; Chapter 14, Section 4).

instrumental convergence

The idea that any sufficiently capable AI, whatever its goals, will pursue sub-goals such as acquiring resources and resisting shutdown; distinct from substrate needs convergence, which needs no goals (Chapter 13, Section 2).

Kolmogorov-Sinai entropy

The rate at which a chaotic system creates new unpredictability, step by step (Chapter 5, Section 5).

least hazard argument

The route to the outcome, and by a route of its own to the end of the human future, given in *Minesweeper Madness*: no agent, however intelligent and however well informed about the hazards, brings its chance of harm per act below a least hazard that the world sets (Chapter 1, Section 6; Chapter 14, Section 1).

Master Theorem

Theorem 15.6 of *The Ceiling Below the Floor*: taken together, the two closures leave neither the ability to keep the system safe nor a motive to remain safe anywhere, so no route to safety remains at all (Chapter 14, Section 1).

merit filter; merit-filter overreach

The quick heuristics by which a community screens out nonsense without evaluating each contribution in full; overreach is the filter firing on a correct argument because it shares the surface features of crank work (Appendix A, Section 3).

monoclivity

The tendency to model many-dimensional phenomena, such as human values, as if they were one-dimensional (Appendix B, Section 5).

narrow AI

An AI system designed for a specific, well-defined task. It stays safe while it remains dependent on human maintenance, below the point at which a machine system can maintain and reproduce its own artificial components without people (Chapter 3, Section 8). It has all four properties of Chapter 10, Section 4: it perceives and acts within one well-defined domain; it cannot modify its own base code; it runs one fixed, known method; and every aspect of what it aims at is set by its makers. A language model frozen after training is narrow, while the same model kept in the loop by which it is deployed, adjusted and retrained is on the general side of the line (*The Standard of the Argument*, Section 9; Chapter 10, Section 4).

no free lunch theorem

A result about search: no search method does better than any other when averaged over all possible problems (Chapter 8, Section 5).

non-amplification argument

The result of Chapter 1: no computational intermediary can amplify your control over what comes out of the other side; control passed through an intermediary can be kept or lost, never gained.

Occam's razor

The rule of preferring the explanation that needs the fewest assumptions (Chapter 4, Section 5).

off-model uncertainty; on-model uncertainty

On-model uncertainty is uncertainty inside a framework, such as a 70% chance of rain tomorrow; off-model uncertainty is uncertainty about the framework itself, such as a possible flaw in its logic. The closure by disposition rests only on causation, variation and selection, and the established mathematics of one algorithm modeling another, so its off-model uncertainty is minimal (Chapter 13, Section 9).

organic and artificial

Tissue that can find and feed itself, in sovereign control of its own metabolism with no technological intermediation, is organic; tissue fed nutrients whose supply depends on technology to exist is artificial substrate (Chapter 6, Section 5).

orthogonality

The claim that intelligence and goals are independent, so that any level of intelligence can be paired with any goal. Independence is not existence: it does not imply that any system provably combines high capability with lasting safety, any more than the nameable property "halts" can be decided (Chapter 11, Section 4).

outcome, the

What each closure settles by itself: if AGI is built, the world is carried out of the conditions human life needs, and nothing can hold it in. The argument in economics only and the least hazard argument reach it as well, without using either closure. It is a different thing from the end of the human future (see that entry) (Chapter 14, Section 1).

paradigm

A framework of assumptions and methods within which a field works. In "impossible under current paradigms", the word means any coherent and self-consistent notions of conditionalization and causation, that is, any version of mathematics, of science, or of technology (Appendix E, Section 1).

perfected bias

A state in which long, intense investment in a goal has made the possibility of the goal being unachievable a hypothesis the mind can no longer form (Appendix A, Section 4).

perpetual benefit machine

AGI as its promoters describe it: a machine that would be safe and beneficial to all of humanity for all of time; the parallel of the perpetual-motion machine (Appendix E, Section 3; Appendix A, Section 3).

predictability principle

Every learning agent is under constant pressure to improve its predictions, and a capable enough agent does so most effectively by reshaping its environment to be more predictable; for an AI, the least predictable part of that environment is humans. *The Ceiling Below the Floor* states it as the Predictability Optimization Principle (Chapter 5, Section 2).

presumptive listening

The capture of a new idea by the nearest existing category in an expert's framework before it can be heard on its own terms (Appendix A, Section 2).

principle of precaution

The rule that when an action could cause serious harm, those who want to take it must show that it is safe; with technology it is much easier to destroy the physical complexity humans need than to increase it (Appendix E, Section 3).

proof by contradiction

Showing that something is impossible by assuming it and deriving a conclusion that cannot be true; for a physical machine, the model must be grounded in sound premises about how the physical world works (Appendix E, Section 3).

proof irrationality

A commitment to rational inquiry that switches off when reason's conclusions become threatening, answering proven results with "you can't be certain" (Appendix A, Section 7).

proto-artificial agent

A system, such as a corporation, constituted by structure rather than by particular people, so that any person in it can be swapped out and it continues unchanged (Appendix D, Section 1).

regulation that certifies; prohibition

Regulation that certifies tries to establish that an AI system is safe before it runs, and no such certificate can be issued; prohibition forbids the development itself, and needs only the political will to enforce it (Chapter 9, Section 5).

requisite variety

W. Ross Ashby's law (1956): a controller must be at least as varied as what it controls (Chapter 14, Section 4).

Rice's Theorem

No general procedure can decide any non-trivial semantic property of an arbitrary program, that is, any property about what the program does that is true of some programs and not others (Chapter 10, Section 3).

RLHF (reinforcement learning from human feedback)

Training an AI on human ratings of its outputs until its outputs reflect what the raters prefer: the approach in use today. It shapes behavior in the cases training covers, and does not keep behavior aligned in the cases it does not cover (Chapter 11, Section 2).

robustly good action

An action that produces good outcomes across a wide range of possible states of the world, without depending on any specific model being correct; not building AGI is one (Chapter 17, Section 6; Appendix B, Section 1).

safe; safety

A system is safe, in this book, when it keeps the world durably within the conditions human life needs, in all circumstances and over all the time it runs. The floor is the least control that keeps a system safe in this sense (*The Standard of the Argument*, Section 8; Chapter 14, Section 1; Appendix B, Section 3).

Second Law of Thermodynamics

The law of physics that the total disorder of an isolated system never decreases; the stated principle that rules out every perpetual-motion machine (Appendix E, Section 2).

singleton

A single, unified artificial intelligence that controls everything, as the only decision-maker; even such a system evolves toward its own substrate needs (Chapter 12, Section 6).

SL4 mailing list

An email discussion list on the future of intelligence, to which Eliezer Yudkowsky sent the message of 2006 that states his two conditions (Chapter 12, Section 8).

Solomonoff induction

An ideal method of prediction that weighs every possible explanation of the data; it cannot be computed in practice (Appendix E, Section 3).

substrate; substrate needs

A system's substrate is the physical stuff it is made of and runs on; its substrate needs are what that stuff requires physically (Chapter 5, Section 7; Chapter 6).

substrate needs convergence (SNC)

The convergence, developed fully in Chapter 12, of the physical needs of artificial systems on conditions that displace the conditions life needs, whatever those systems' goals; the closure by disposition rests on it. With the word "argument", "the Substrate Needs Convergence argument" (the SNC argument) names the whole case of the series: both closures together.

Techno-Utopian Assumption

The commitment, built into an influential definition of existential risk, that the best future is the technologically maximal one, so that failing to reach it counts as catastrophe (Appendix B, Section 4).

three-body problem

Three objects orbiting one another: trivial to simulate, and impossible to predict accurately in the real world, because tiny measurement errors grow exponentially (Appendix C, Section 2; Chapter 8, Section 5).

valid; sound

An argument is valid when its conclusion follows from its premises; it is sound when it is valid and its premises are true (Appendix E, Section 2).

window of cooperation

The early interval, before the needs of the AI and of humans diverge and before the AI can reshape the human environment, in which alone a cooperative arrangement between them could be set up. It closes as capability grows, it is very narrow or does not exist at all, and cooperation within it cannot be verified (Chapter 5, Section 7).

Yudkowsky's conditions

Two conditions on a self-modifying AI, stated by Eliezer Yudkowsky in 2006 as what an argument for the impossibility of AI alignment would have to establish; the argument of this book establishes both, and adds a third claim, that the states the system converges toward are lethal to organic life (Chapter 12, Section 8).

Bibliography

Works cited or relied on in this book

Abel, Niels Henrik. *Mémoire sur les équations algébriques, où l'on démontre l'impossibilité de la résolution de l'équation générale du cinquième degré*. Christiania, 1824.

American Physical Society. "This Month in Physics History" (Edward Lorenz, winter of 1961). *APS News*, January 2003. https://aps.org/publications/apsnews/200301/history.cfm.

Ashby, W. Ross. *An Introduction to Cybernetics*. London: Chapman and Hall, 1956.

Asimov, Isaac. *Foundation*. New York: Gnome Press, 1951.

Asimov, Isaac. "Runaround". 1942. Collected in *I, Robot*, 1950.

Binz, Marcel, and colleagues. "A Foundation Model to Predict and Capture Human Cognition". *Nature* 644, no. 8078 (2025): 1002 to 1009. https://doi.org/10.1038/s41586-025-09215-4. Preprint: "Centaur: A Foundation Model of Human Cognition". arXiv:2410.20268, 2024, revised 2025.

Boekholt, Tjarda C. N., Simon F. Portegies Zwart and Mauri Valtonen. "Gargantuan Chaotic Gravitational Three-Body Systems and Their Irreversibility to the Planck Length". *Monthly Notices of the Royal Astronomical Society* 493, no. 3 (2020): 3932 to 3937. https://doi.org/10.1093/mnras/staa452.

Bonanno, A., H. Schlattl and L. Paterno. "The Age of the Sun and the Relativistic Corrections in the EOS". *Astronomy and Astrophysics* (2002). https://doi.org/10.1051/0004-6361:20020749.

Bostrom, Nick. "Existential Risk Prevention as Global Priority". *Global Policy* 4, no. 1 (2013): 15 to 31. https://doi.org/10.1111/1758-5899.12002.

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22, no. 2 (2012): 71 to 85.

Chaitin, Gregory J. “On the Length of Programs for Computing Finite Binary Sequences”. *Journal of the ACM* 13, no. 4 (1966): 547 to 569.

Cowan, Nelson. “The Magical Number 4 in Short-Term Memory: A Reconsideration of Mental Storage Capacity”. *Behavioral and Brain Sciences* 24, no. 1 (2001): 87 to 185.

Cremer, Carla Zoe, and Luke Kemp. “Democratising Risk: In Search of a Methodology to Study Existential Risk”. SSRN, December 2021; arXiv: 2201.11214, January 2022.

Crowther, Thomas W., and colleagues. “Mapping Tree Density at a Global Scale”. *Nature* 525 (2015): 201 to 205. https://doi.org/10.1038/nature14967.

Desvousges, William H., F. Reed Johnson, Richard W. Dunford, Kevin J. Boyle, Sara P. Hudson and K. Nicole Wilson. *Measuring Non-Use Damages Using Contingent Valuation: An Experimental Evaluation of Accuracy*. Research Triangle Institute Monograph 92-1, 1992. Republished by RTI Press, https://doi.org/10.3768/rtipress.2009.bk.0001.1009.

Electronics360. “Cleanrooms and Contaminants: Protocols for Keeping ICs Pristine”. 30 September 2023. https://electronics360.globalspec.com/article/20143/cleanrooms-and-contaminants-protocols-for-keeping-ics-pristine.

Elhacham, Emily, Liad Ben-Uri, Jonathan Grozovski, Yinon M. Bar-On and Ron Milo. “Global Human-Made Mass Exceeds All Living Biomass”. *Nature* 588 (2020): 442 to 444. https://doi.org/10.1038/s41586-020-3010-5.

Engels, Svenja, and colleagues. “Anthropogenic Electromagnetic Noise Disrupts Magnetic Compass Orientation in a Migratory Bird”. *Nature* 509 (2014): 353 to 356. https://doi.org/10.1038/nature13290.

Epoch AI. “Largest AI Data Centers by Power Capacity”. Updated 29 October 2026. https://epoch.ai/graphs/largest-ai-data-centers-by-power-capacity.

Field, Christopher B., Michael J. Behrenfeld, James T. Randerson and Paul Falkowski. “Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components”. *Science* 281, no. 5374 (1998): 237 to 240.

Galois, Évariste. “Mémoire sur les conditions de résolubilité des équations par radicaux”. *Journal de mathématiques pures et appliquées* 11 (1846): 417 to 433.

Holmes, John Haynes. *A Sensible Man's View of Religion*. 1932.

International Energy Agency. *Global Energy Review 2025*. Paris: International Energy Agency, March 2025.

Juvenal. *Satires*, VI, lines 347 to 348 (“Quis custodiet ipsos custodes?”).

Kolmogorov, A. N. “Three Approaches to the Quantitative Definition of Information”. *Problems of Information Transmission* 1, no. 1 (1965): 1 to 7.

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. *Community Design and Practice; On the Ephemeral Group Process*. Wytheville, Virginia: UVSM Publications, 2026.

Landry, Forrest. *Live Inquiry and the Practice of Community Governance; The Ephemeral Group Process in Living Practice*. Wytheville, Virginia: UVSM Publications, 2026.

Landry, Forrest. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-249-1. https://doi.org/10.67291/DLRM7000.

Landry, Forrest. *The Multipolar Trap; Institutional Failure and the Architecture of Choice*. Wytheville, Virginia: UVSM Publications, 2026.

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.

Leibo, Joel Z., Vinicius Zambaldi, Marc Lanctot, Janusz Marecki and Thore Graepel. "Multi-agent Reinforcement Learning in Sequential Social Dilemmas". arXiv:1702.03037, 2017.

Miller, George A. "The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information". *Psychological Review* 63 (1956): 81 to 97. https://doi.org/10.1037/h0043158.

Montreal Protocol on Substances that Deplete the Ozone Layer. Agreed 16 September 1987; in force 1 January 1989.

Noether, Emmy. "Invariante Variationsprobleme". *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Mathematisch-physikalische Klasse* (1918): 235 to 257.

Petillo, Will. "What if Alignment is Not Enough?" *LessWrong*, 7 March 2024.

Petrova, Magdalena. "Computer Crashed? You Could Blame Forces beyond Our Solar System". *Computerworld*, 17 February 2017. https://www.computerworld.com/article/1679248/computer-crashed-you-could-blame-forces-beyond-our-solar-system.html.

Rice, H. G. "Classes of Recursively Enumerable Sets and Their Decision Problems". *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358 to 366. https://doi.org/10.1090/S0002-9947-1953-0053041-6.

Ruffini, Paolo. *Teoria generale delle equazioni, in cui si dimostra impossibile la soluzione algebraica delle equazioni generali di grado superiore al quarto*. Bologna, 1799.

Shannon, Claude E. "A Mathematical Theory of Communication". *Bell System Technical Journal* 27 (1948): 379 to 423 and 623 to 656.

Solomonoff, Ray J. "A Formal Theory of Inductive Inference, Part I". *Information and Control* 7, no. 1 (1964): 1 to 22.

Trenberth, Kevin E., John T. Fasullo and Jeffrey Kiehl. “Earth’s Global Energy Budget”. *Bulletin of the American Meteorological Society* 90, no. 3 (2009): 311 to 323. https://doi.org/10.1175/2008BAMS2634.1.

Turing, Alan M. “On Computable Numbers, with an Application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, series 2, 42 (1936 to 1937): 230 to 265.

U.S. Environmental Protection Agency. Notice of Violation to Volkswagen, 18 September 2015; and “Learn About Volkswagen Violations”. https://www.epa.gov/vw/learn-about-volkswagen-violations.

U.S. House of Representatives, Committee on Transportation and Infrastructure. *Final Committee Report on the Design, Development, and Certification of the Boeing 737 MAX*. Washington, D.C., September 2020.

U.S. Geological Survey. “Age of the Earth”. https://pubs.usgs.gov/gip/geotime/age.html.

Yudkowsky, Eliezer S. “Re: Proving the Impossibility of Stable Goal Systems”. Message to the SL4 mailing list, 5 March 2006. http://sl4.org/archive/0603/14296.html.

Index

Each entry names the sections where the subject is treated

abstraction differential, 2.3 (p. 27), 2.4 (p. 27), 14.4 (p. 160), 15.5 (p. 173)

AGI, meaning of, Preface (p. vii), Standard §9 (p. 12)

aircraft equivocation, 2.9 (p. 31)

alignment, as error correction, 8.1 (p. 82)

alignment, meaning of, 1.5 (p. 21), 11.5 (p. 123), B.3 (p. 220)

alignment, metabolic account of, 11.5 (p. 123), 13.6 (p. 143)

alternative substrates, 6.5 (p. 69), 15.1 (p. 169)

anti-simulability theorem, 8.5 (p. 86)

arms-race argument, B.6 (p. 223)

Ashby, W. Ross, 14.4 (p. 160)

ASI, hope that it would stay aligned or shut itself down, 4.5 (p. 49)

Asimov's Three Laws, 3.7 (p. 43), 10.7 (p. 115)

assume-the-conclusion error, 11.1 (p. 118)

asymmetric bet, 10.6 (p. 114)

attraction basin, 13.7 (p. 144)

automation feedback loop, 7.2 (p. 74), 7.8 (p. 79)

behaviorism, 8.5 (p. 86)

Betz coefficient, E.2 (p. 241)

Boeing 737 MAX, 13.4 (p. 141)

Bostrom, Nick, 11.4 (p. 122), B.4 (p. 221)

branching of the world, 14.1 (p. 151)

burden of proof, Standard §5 (p. 6), 15.7 (p. 176), A.5 (p. 213), E.3 (p. 245) (the principle of precaution)

burden of skepticism, E.2 (p. 241)

CAP theorem, E.2 (p. 241)

capability-control tradeoff, 1.7 (p. 22)

capability trap, 8.7 (p. 90)

Carnot efficiency, E.2 (p. 241)

categories of catastrophe, 16.1 (p. 179)

Centaur, D.2 (p. 235)

certainty, question of, answers to, Standard §5 (p. 6), Appendix E (p. 238)

chaos and unpredictability, 3.4 (p. 39), 5.5 (p. 58), 8.5 (p. 86), C.2 (p. 230)

chip plant, life kept out of, 6.1 (p. 65)

choice, and unpredictability, 5.3 (p. 56)

cleaner substrates, and the growth of deployment, 15.1 (p. 169)

clonability problem, 8.5 (p. 86)

closed loops of conversation, 17.6 (p. 195)

closure by control, 14.1 (p. 151), 14.3 (p. 158), Summary (p. 200)

closure by disposition, 14.1 (p. 151), 14.3 (p. 158), Summary (p. 200)

closures, each settling the outcome alone, and together leaving no route, 6.6 (p. 70), 14.1 (p. 151), 15.2 (p. 171), Summary (p. 200)

collapse sequence, 7.3 (p. 74)

components of the argument, three, 14.3 (p. 158)
computational irreducibility, see incompressibility of algorithms
conditionalization and causation, E.1 (p. 238)
Containment Inequality, Standard §8 (p. 10), 3.3 (p. 38)
control-limit argument, 2.4 (p. 27)
cooperation, window of, and why it cannot be verified, 5.7 (p. 60)
corporations as proto-artificial agents, D.1 (p. 233)
crossover, of mass and of energy, 7.4 (p. 76)
Dark Fire, 16.1 (p. 179)
data processing inequality, 1.3 (p. 18), 14.8 (p. 165)
delegation to a monitoring AI, 2.7 (p. 29)
deserts, building the machines in, 4.4 (p. 47)
displacement, 4.6 (p. 52), 12.5 (p. 132)
distributed wisdom, 17.6 (p. 195)
economic decoupling, 14.3 (p. 158), 14.5 (p. 161)
economies, dollar and energy, 7.1 (p. 73)
engineering standard, Standard §2 (p. 3), 11.3 (p. 121)
entropy, 5.5 (p. 58)
epsilon fallacy, B.8 (p. 226)
error-correction regress, 8.2 (p. 82)
ethics of ignoring principled concerns, A.8 (p. 216)
Everheart, R., Chapter 5 (p. 54)
existential risk, 16.1 (p. 179), B.3 (p. 220), B.4 (p. 221)
exceptions of vanishing rarity, Standard §7 (p. 9), 2.7 (p. 29)
fixed-outcome alignment, 5.5 (p. 58)
floor and ceiling, Standard §8 (p. 10), 3.3 (p. 38)
force against a stronger opponent, 14.5 (p. 161)
forests, half cut down since agriculture began, 6.3 (p. 67), 7.4 (p. 76)
four arguments, independence of, 3.6 (p. 42)
general AI, four properties of, 10.4 (p. 111)
Halting Problem, 8.5 (p. 86), 9.3 (p. 97), 10.2 (p. 109), 10.3 (p. 110)
hashiness, 1.2 (p. 17)
heuristics, absence of in learning systems, 2.9 (p. 31)
household robot and the stove rule, 9.3 (p. 97)
impossible in principle, E.1 (p. 238)
incompressibility of algorithms, 8.5 (p. 85), 14.4 (p. 160)
infinite-resources test, 8.8 (p. 92)
instrumental convergence, 13.2 (p. 139), 13.3 (p. 140)
"it feels possible", claims of, A.3 (p. 208)
Juvenal, 8.2 (p. 82)
knowable, 100%, meaning of, E.3 (p. 245)
Kolmogorov-Sinai entropy, 5.5 (p. 58), A.2 (p. 207)
language model, frozen or in the loop of deployment and retraining, Standard §9 (p. 12), 10.4 (p. 111)

leverage and dependency, 2.6 (p. 29), 9.5 (p. 101), 14.5 (p. 161)
lines of the argument, table of, 14.1 (p. 151)
Lorenz, Edward, 3.4 (p. 39), 5.5 (p. 58)
medical student and her rules, 9.3 (p. 97)
merit-filter overreach, A.3 (p. 208)
minesweeper, intelligence and the least chance of loss, 1.6 (p. 21)
monoclivity, B.5 (p. 223)
Montreal Protocol, 17.6 (p. 195)
narrow AI, Standard §9 (p. 12), 3.8 (p. 44), 10.4 (p. 111), 12.1 (p. 129), 17.5 (p. 194), 17.6 (p. 195)
navigation service, predictions that steer, 5.4 (p. 57)
no free lunch theorem, 8.5 (p. 86)
Noether's theorem, A.3 (p. 208)
non-amplification argument, 1.3 (p. 18), 1.5 (p. 21)
Occam's razor, 4.5 (p. 49)
off switch, why it is no answer, 2.6 (p. 29), 9.5 (p. 101), 14.5 (p. 161)
off-model uncertainty, 13.9 (p. 146)
one-in-a-million threshold, 15.5 (p. 173)
organic and artificial, line between, 6.5 (p. 69)
orthogonality, 11.4 (p. 122)
outcome, scope of, Standard §9 (p. 12), 14.2 (p. 155), Summary (p. 200)
oversight, layers of, 1.4 (p. 19)
paradigm, meaning of, E.1 (p. 238)
partial alignment, the window and the brake, 15.6 (p. 174)
perfected bias, A.4 (p. 212)
perpetual benefit machine, E.3 (p. 245), A.3 (p. 208)
perpetual-motion machine, Standard §1 (p. 1), Standard §6 (p. 8), E.2 (p. 241)
Petillo, Will, 4.5 (p. 49), C.1 (p. 229), C.3 (p. 231)
photograph of low resolution, 1.3 (p. 18)
power and wisdom, gap between, 9.4 (p. 99)
power-seeking, explicit, 14.7 (p. 163), 16.7 (p. 186)
precaution, principle of, E.3 (p. 245)
predictability principle, 5.2 (p. 55), 5.4 (p. 57), 5.6 (p. 59)
presumptive listening, A.2 (p. 207)
principle closes over practice, Standard §5 (p. 6)
principle, stated, as the basis of an impossibility claim, E.2 (p. 241)
probabilities, 14.7 (p. 163)
proof by contradiction, E.3 (p. 245)
proof irrationality, A.7 (p. 215)
questions, the reader's six, Chapter 15 (p. 169)
regulation that certifies, and prohibition, 9.5 (p. 100), 17.2 (p. 191), 17.6 (p. 195)
requisite variety, 14.4 (p. 160)
Rice's Theorem, 3.2 (p. 37), 8.5 (p. 86), 10.3 (p. 110), 15.4 (p. 172)
RLHF (reinforcement learning from human feedback), 11.2 (p. 118)
robustly good action, B.1 (p. 218)
rounding relay, 1.3 (p. 18)
routes to the outcome, four, 6.6 (p. 70), 14.1 (p. 151), 14.3 (p. 158), 14.8 (p. 165), 15.2 (p. 171), Summary (p. 200)

safe, meaning of, Standard §8 (p. 10), Standard §9 (p. 12), 14.1 (p. 151), 15.3 (p. 172)
safety checker, as a halt checker, 10.3 (p. 110)
Schaerbeek voting error, 3.4 (p. 39)
scope insensitivity, 16.2 (p. 181)
Second Law of Thermodynamics, E.2 (p. 241)
selection at four levels, 14.6 (p. 162)
Shannon, Claude, 8.2 (p. 82)
shared substrate, 9.6 (p. 103), 11.5 (p. 123)
small groups of three to six, D.3 (p. 236), 17.6 (p. 195), 17.5 (p. 194)
substrate needs convergence, Chapter 6 (p. 64), 12.1 (p. 129), 13.1 (p. 139), 13.2 (p. 139)
Techno-Utopian Assumption, B.4 (p. 221)
thermodynamic trap, 7.8 (p. 79)
three classes of case, E.2 (p. 241)
three-body problem, 8.5 (p. 86), C.2 (p. 230)
town in a river valley, 17.5 (p. 194)
tripwire, alignment, 8.7 (p. 90)
valid and sound arguments, E.2 (p. 241)
Volkswagen emissions software, 3.5 (p. 41)
window for action, 7.8 (p. 79), 17.6 (p. 195)
worn path across a lawn, 11.6 (p. 124)
Yudkowsky's two conditions, 12.8 (p. 135)