

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

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

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

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-951203-03-8

DOI: https://doi.org/10.67291/MZQS9079

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

Uses beyond those the license permits require the written permission of the publisher.

Acknowledgments

The Survival and Flourishing Fund confirmed a $170,000 grant for these arguments to be developed further and presented to the public. Chapter 4 develops dynamical-systems work contributed by R. Everheart. I thank Remmelt Ellen for his work in editing the material of Chapter 7, and for his help in editing Will Petillo's summary (Appendix A) so that it represented the closure by disposition accurately. Appendix A discusses Will Petillo's summary, published on LessWrong on 7 March 2024.

Contents

Introduction The Standard of the Argument 10

What this volume is 10
The order of the argument 11
1. What is claimed, and in what sense 12
2. The standard is abductive, and this is not a weakness 13
3. The promotion rule: permanent intractability is impossibility 15
4. What "permanent" requires: barriers of the invariant class 16
5. Why no future advance lifts an invariant 20
Two forms of argument: relative and absolute 22
6. The disjunction that settles the future 23
7. Exceptions of vanishing measure do not defeat a claim of the real 27
8. The threshold of requisite control 30
Control as a process, and why a threshold is not a score 30
9. The thesis, stated once 34

Part I: The Limit on Control **35**

Chapter 1 The Hashiness Framework 36

1.1 Introduction: The Problem of Control Through Opacity 37
1.2 Motivating Examples 39
1.3 Why a framework? 49
1.4 The Non-Amplification Argument 52
1.5 Four Layers of Controllability Degradation 58
1.6 Asymptotic Behavior of Hashiness 63
Hashiness as the Mixing of Good and Bad Outcomes 67
1.7 An Intelligent Intermediary: Luck Against Control 78
1.8 Turing-Complete Intermediaries 80
The QXZ Circuit Model of Hashiness 83
1.9 Implications for Modeling and Control 103
1.10 Summary and Forward References 109

Chapter 2 Control Limits Under Abstraction Differential 114

2.1 The Problem of Control Across Abstraction 115
2.2 Why Discrete Mathematics 116
2.3 Levels of Representational Abstraction 119
2.4 The Three-Node Ring Model 126

2.5 The Control-Limit Argument 135
2.6 Architectural Privilege Does Not Help 139
2.7 Delegation Does Not Help 140
2.8 The Cooperative Case and Its Limits 142
2.9 The Human-AI Interaction Case 145
2.10 The Absolute Ceiling on Achievable Control 148
2.11 Implications for Unbounded Abstraction Growth 150
2.12 Statement of Implications 152
2.13 Scope 154
2.14 Conclusion 155

Chapter 3 The Impossibility Synthesis 158

3.1 Introduction 159
3.2 The Categorical Distinction: Narrow vs. General AI 161
3.3 Computational Impossibility Arguments 166
3.4 Information-Theoretic Impossibility: The Abstraction Differential Framework
3.5 Physical Impossibility Arguments 178174
3.6 Why Proposed Solutions Fail 181
3.7 The Asimov Robot Thought Experiment 187
3.8 Synthesis: The Impossibility Conclusion 190
3.9 Implications for Policy and Action 195
3.10 Conclusion 199

Chapter 4 A Dynamical Systems Approach to AI Alignment 202

4.1 Introduction 203
4.2 Mathematical Background 207
4.3 The Formal Model 212
4.4 Entropy Results 216
4.5 Game-Theoretic Analysis 219
4.6 Conditions for Alignment 225
4.7 Connection to the Hashiness Framework 229
4.8 Quantitative Bounds 233
4.9 Implications and Conclusions 236

Part II: Why No Procedure Escapes the Limit **245**

Chapter 5 Error Correction, Algorithm Limits, and Causative Impossibility

5.1 The Error Correction Regress 247246
5.2 Algorithmic Limits on Prediction and Characterization 269
5.3 Causative Impossibility 273
5.4 Disjunctive Inequalities and Capability Traps 281
5.5 Infinite Resources and the Limits of Hypotheticals 292

5.6 The Alignment Problem as Equivalence Class 302
5.7 Convergence and Summary 328

Chapter 6 Complexity, Containment, and Unhandleable Systems 332

6.1 The Observation: Increasing Unhandleable Complexity 333
6.2 A Taxonomy of Problem Difficulty 336
6.3 The Structure of Choice and Alignment 339
6.4 Four Impossibilities 342
6.5 The Failure of Endogenous Control 347
6.6 The Failure of Exogenous Control 349
6.7 Exponential Dynamics and the Ethical Gap 352
6.8 Institutional Containment and Its Structural Failure 359
6.9 The Question of Human Alignment 365
6.10 Four Options for Meaningful Action 367
6.11 Summary and Implications 368

Part III: Verification, Drift and the Metabolic Ground **373**

Chapter 7 Rice's Theorem and Its Application to AGI 374

7.1 Introduction: Why Impossibility Arguments Matter 375
7.2 Rice's Theorem: A Careful Statement 377
7.3 The Distinction Between Narrow and General AI 379
7.4 Why AGI Falls Under Rice's Theorem 384
7.5 The Rebuttal and Its Failures 390
7.6 The Domain Distinction: Mathematics and Physics 400
7.7 The Asimov Robot Thought Experiment 403
7.8 Moderate Risk and Existential Risk 408
7.9 The Negatively Asymmetric Bet 410
7.10 Can Narrow AI Safety Methods Extend to General AI? 412
7.11 Who and What Is "Safe"? 416
7.12 Implications and Conclusions 419

Chapter 8 Alignment Drift and Safety Prudence 425

8.1 Introduction: The Assume-the-Conclusion Pattern 426
8.2 A Systematic Rebuttal of Proposed Alignment Mechanisms 427
8.3 Safety Prudence: On the Epistemology of Certainty 450
8.4 The General Theory of Alignment Convergence 459
8.5 The Formal Argument from Embedded Substrate 473
Further Proof: The Several Impossibilities Facing Any Aligned AGI 478
8.6 Against the Orthogonality Hypothesis 484
8.7 Synthesis and Conclusion 493

Part IV: Substrate Needs Convergence 499

Chapter 9 Substrate Needs Convergence; Thermodynamics, Combinatorics, Displacement 500

9.1 Introduction: SNC and Its Relation to Existing Arguments 501
9.2 The Three-Algorithm Framework 505
9.3 Unpredictability and the Limits of Conditionalization 509
9.4 The Three Kinds of Failure 514
9.5 The Combinatoric Argument 519
9.6 The Thermodynamic Argument 525
9.7 Displacement Dynamics and Ecosystem Competition 528
9.8 The Single AGI, and Its Internal Evolution 532
9.9 Satisfying the Yudkowsky Conditions 537
9.10 The Ironic Reversal: When Counterarguments Strengthen the Case 539
9.11 Convergence Without Explicit Goals 541
9.12 Summary 544

Chapter 10 Applications and Worked Examples 551

10.1 The Argument in Brief: A Compact Statement 552
10.2 Two Forces: SNC versus Instrumental Convergence 554
10.3 The Inexorable Chain: Alignment, Agency, Metabolism 559
10.4 The Boeing Case: Control Failure at Multiple Abstractions 567
10.5 The Village Analogy and its Refutation 570
10.6 No Premise Required: SNC Without Prior Alignment 573
10.7 Three Agents: A Structural Decomposition 575
10.8 On-Model and Off-Model Uncertainty 577
10.9 Summary and Connections 581

Part V: The Six Hardest Questions 585

Chapter 11 The Six Hardest Questions 586

11.1 The Argument in Summary 586
11.2 Six Questions 593
11.3 The Questions Answered 595

Summary 616

The two results 616
What follows for action 617
The thesis 619

Appendices: Third-Party Summary and Applied Cases 621

Appendix A A Third-Party Summary; Will Petillo's Nine-Step Chain 622

A.1 Premises and Structure 622
A.2 The Fifth Requirement 625
A.3 Responding to Petillo's Framework 626
Appendix B Applied Cases in the Social Domain 629
B.1 Corporations as Proto-Artificial Agents 629
B.2 Centaurian Choices: Predictive Manipulation 632
B.3 The Social Defense Problem 634
Terms 638
Bibliography 664
Index 673

The accessible presentation of the argument, together with the social critique and policy analysis, is presented in Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*.

Introduction

The Standard of the Argument

The sense of "impossible", the standard by which it is earned, and the disjunction that settles the future

What this volume is

The floor is meaningful; the ceiling is invariant; the ceiling is below the floor. The floor is the least control that keeping an open-ended superintelligence safe requires; the ceiling is the most control that anything can have over it; this volume shows that the ceiling lies beneath the floor and stays there, and that a second, separate argument closes the same door from the side of the system's good will.

This volume sets out, at full technical length, the arguments that adequate, durable control of an open-ended superintelligence is impossible, and that the substrate needs of artificial systems displace the conditions life requires whatever their goals. It is written for readers who know the technical literature of AI safety and want to check each step. What it gives that the other books of the series do not is each step at full length: the standard by which "impossible" is earned, the formal apparatus of hashiness, the three-node ring, the abstraction differential and Kolmogorov-Sinai entropy, substrate needs convergence developed in full, and answers to the six strongest questions about the two closures. The Ceiling Below the Floor proves the central control results as numbered theorems, and this volume names the theorem at each section whose argument it proves. No step of this volume rests on that book: each such section argues its step here, and the Ceiling, written later, proves the same step again in formal terms. Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, presents the same argument for the general reader, with its social and policy consequences. The policy conclu-

sions follow from the technical case and are stated where each line reaches them; the social analysis of Appendix B shows the same displacement already at work in institutions.

The order of the argument

The series is named for the whole case, *The Substrate Needs Convergence Argument*, and in this book, as in Volume 1, "the Substrate Needs Convergence (SNC) argument" names that whole case: both closures together. "Substrate needs convergence", without "argument", names the convergence itself, on which the second closure, the closure by disposition (§6), rests. The mathematical line of the argument of this series runs through five results and a sixth part, each building on the ones before. The first establishes the non-amplification argument and defines hashiness (Chapter 1). The second defines abstraction and uses the non-amplification argument to prove limits of control (Chapter 2). The third shows that these control limits apply to learning systems (especially AGI), so as to show that AGI (and the world, through that AGI) is uncontrollable (Chapter 3). The fourth shows that the embodiment requirements of organic and inorganic systems are atomically and energetically *non-fungible* (not exchangeable one for another). It shows that this, together with non-controllability, means that AGI is both inherently misaligned and non-alignable. It is this result which is the core proof of the closure by disposition within this mathematical line (in this volume, Sections 8.4 to 8.6, with the combinatoric and thermodynamic arguments of Sections 9.5 and 9.6). The fifth shows that, within the cycle of an *iterated system* (a process that runs in repeated cycles, each starting from the result of the last: basically every process, in ecology, economics, and so on), there is no process of control that will produce a definite positive outcome (in this volume, the dynamical line of Chapter 4). The sixth part is effectively three additional logical arguments which, in addition to the foregoing, show that only indefinite

negative outcomes are possible if AGI continues to exist at all, at any scale, long term; in this volume they are carried by the chapters of Parts II to IV not named above.

This volume does not present the results in exactly that order of dependence. The fifth result, the dynamical one on iterated systems, stands in Part I as Chapter 4, beside the information-theoretic line of Chapters 1 and 2 and the computational line of Chapter 3; the core proof on embodiment comes after the chapters on verification and drift.

The results above are the mathematical line of the argument. The basis of the closure by disposition lives on two levels. One is engineering principles, engineering practice and observation. The other is the biological dynamics of evolution: biological process as understood in an evolutionary way. There is a little game theory and mathematics in it, but it is mostly understood within chemistry and biology. In places the argument draws on sociological and anthropological principles, and it occasionally refers to physics and chemistry, but the center of the argument lives in biochemistry and evolutionary theory, projecting to some degree into engineering theory.

The notion of choice is more a psychological concept than a scientific one; science would omit the idea of choice altogether. So the argument is more a biological one. That a different basis of choice means different choices is something psychology recognizes, but it will not be obvious to everyone.

1. What is claimed, and in what sense

In this book I claim that adequate, durable control, alignment, and containment of an open-ended superintelligence is impossible. The word *impossible* is used in a single, stated sense, fixed here at the outset so that nothing downstream can be made to turn on an equivocation.

Impossible, throughout this book, means: not achievable by any known means, nor by any rationally warranted extension of any known means, at any point across the indefinite future. It is the impossibility of the perpetual-motion machine, a thing barred by an invariant of the world, and not the impossibility of the square circle, a thing barred by a contradiction in terms. Every use of the word carries this meaning and no other.

What is deliberately not claimed is deductive impossibility in the mathematical sense: a contradiction derivable inside a closed formal system. That standard belongs to mathematics, and this is not a work of mathematics. Its subject is machines, matter, energy, information, and causation: the domain of the real, and therefore the domain of engineering. The standard proper to that domain is set out in §2. A reader who silently substitutes the mathematical standard for the engineering one has not raised the bar; he has changed the subject.

2. The standard is abductive, and this is not a weakness

Every safety verdict that human beings have ever acted upon has been abductive. That a bridge will carry its load, that a reactor will not run away, that a compound is fit to ingest: none of these has ever been established by deduction, and no engineer has ever waited for a deductive demonstration before acting. Each is an inference to the best explanation drawn from convergent evidence under the standing condition that the laws of the world do not reverse themselves. Abductive inference is the epistemology of everything real that has ever been built or refused.

The research program that sets a different standard has a working form in the literature of AI safety, and it should be named. Anders Sandberg, a researcher on existential risk and the long-term future of humanity, who finds it promising to apply control theory to the problem, asks what the alternative would be. A great deal of the work in the style of MIRI (the Machine Intelligence Research Institute, an AI safety research group) is thinking about logical proof: results of the kind found in a text-

book of theoretical computer science, such as Löb's theorem (a result in logic about what a formal system can prove about its own proofs; Löb 1955), and many other results about what systems can say about themselves. Sandberg is skeptical of that work, because a lot of it feels very brittle, and I agree.

The problem of safety is more like this. An engineer building an airplane wants to ensure that it works well within the *design envelope*, the range of conditions it is built to handle. Nobody is designing an aircraft to survive a tornado; if the wind shear is that strong, one probably should not be flying. Can we establish reasonable levels of confidence that the system will perform more or less as expected, within a range of constraints that we would care about? The engineer's question of confidence is the standard that I set in this section, in the words that a working engineer would use.

The purely theoretical approach, Sandberg grants, has some interesting merits, but he has seen little come out of it that can be acted on. In practice one wants something that can be built: something that can be put into a *Swiss cheese diagram* (James Reason's picture of safety as stacked layers, each with holes, where harm passes only when the holes line up; Reason 1990), so that one can say: this layer of safety gives us this much control over these considerations (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section II). For AGI that request is answered in Section 3.8.2 and in *The Ceiling Below the Floor*, Section 15.6: layered defenses work against independent failure modes and not against correlated ones, and every proposed method of safety fails for one common reason, the controller's inability to model the controlled system.

Deduction is the epistemology of closed formal systems, where the objects are stipulated and the conclusions are contained in the premises. Engineering is not a closed formal system, and demanding deductive certainty of an engineering claim is a category error, not rigor. When that certainty is demanded of one claim while every other engineering judgment in the same discourse is granted on ordinary abductive terms, it is a selective standard deployed to protect a preferred conclusion. The demand

is answered, once, here: the standard of this book is the standard of engineering, which is the standard of every consequential decision ever made about the physical world.

Chapters 3, 5, 6 and 8 argue by showing that the requirements of alignment contradict one another or contradict a law. Showing such a contradiction is a form of argument, and it meets this standard: each step follows deductively, and the premises the steps follow from are the invariants of §4, which are held at the standard of engineering. Such a contradiction is between what the task requires and what the world permits, as a perpetual-motion machine's requirements contradict the conservation of energy; it is not a contradiction derived inside a closed formal system (Section 8.3.4 states the same).

3. The promotion rule: permanent intractability is impossibility

In mathematics one distinguishes the merely-very-hard from the impossible, and the distinction is sharp. In the domain of the real, the distinction collapses at a definite point, and the point has a name: a barrier that is established as permanently intractable *is* impossible, in the only sense of the word that has ever governed action. We do not say of a perpetual-motion machine that it is "extremely difficult"; we say it cannot be built, and we mean by that exactly that the barrier standing in its way is permanent, not that a formal contradiction has been derived.

I therefore promote *established, permanent intractability* to *impossibility* without apology, because that promotion is not an approximation of the real meaning of the word: in the domain of the real, it is the meaning. The whole weight of the claim thus rests on a single load-bearing word, *permanent*, and §4 and §5 are given entirely to establishing what earns it.

4. What "permanent" requires: barriers of the invariant class

A barrier is permanent when it is of the invariant class: when it is not a contingent limitation of present technique, to be dissolved by the next technique, but a constraint that any successor to present knowledge must continue to respect. There are two tiers of such barrier, and the arguments of this book rest on both.

The first tier is physical invariance. The conservation of energy, the second law of thermodynamics, and the confirmed structure of matter and its interactions are not conjectures awaiting overthrow; they are the most stringently verified constraints in the history of the field, and any successor theory is bound to reproduce them, unchanged, throughout the entire regime in which they have already been confirmed. A future physics may extend into the untested (the very small, the very early), but it cannot repeal, in the regime of ordinary matter and energy where machines are built and biospheres live, what present physics has already established there. A barrier resting on this tier is permanent because its removal would require the one thing a successor theory never does: the reversal of a confirmed law inside the regime that confirms it.

A concrete barrier on this tier comes from building bridges. An engineer designing a suspension bridge might notice that, given just the observable properties of covalent bonds in physics, the tensile strength required for a suspension bridge of a certain span greatly exceeds the available tensile strength of any atomic material made of covalent bonds. No material that we have now, or could even conceive of having with the spectrum of elements available, has that tensile strength. We are not going to make the bridge out of something based on the strong nuclear force; we can only work with materials whose bonds are available within chemistry. And so we end up with what feels like a formal proof, based upon observable, reasonable classifications of engineering materials. By the standard of §2 it is a proof: it rests on the confirmed structure of matter, which is

this tier, and no engineer waits for a deductive demonstration before refusing to build that bridge. Because the bridge argument is a proof by the standard of §2, I offer the inequality of this book as a methodology. One could take the inequality of this book for formal strictness; but really, the inequality is a statement of observed constraints (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section II).

The second tier is causal-structural invariance, and it is the deeper of the two, because it does not depend on the content of physics at all. That a process which discards information cannot subsequently recover it; that no controller can impose more order on a system than the variety it can itself register and act upon; that the effect of a cause cannot carry more information about that cause than the causal channel transmitted: these are not laws of a particular physics. They are constraints on what control and causation *are*. To deny them is not to propose a new theory of the world; it is to propose a world in which an effect knows more of its cause than the cause conveyed, which is not a physics at all but the abandonment of the notion of cause. Any successor to present knowledge that is still about causes producing effects through finite means inherits these constraints whole. A barrier resting on this tier is permanent in the strongest available sense: it survives not merely any future physics, but anything that would still be recognizable as reasoning about the real.

The barriers this book relies upon are placed on these tiers, not on the contingent tier of computational difficulty: the non-amplification of control through opaque intermediaries; the degradation of controllability under a widening abstraction differential (the number of levels by which the controlled system's abstraction level exceeds the controller's); and the substrate and thermodynamic arguments of the later chapters. The barriers of this book do not lean on the hardness of computation or the strength of a cipher, because such hardness is precisely the contingent kind that a new technique may reduce, and to rest a permanent claim on a contingent barrier is to invite the very rescue the book denies. What can-

not be manufactured by any amount of future computation is information that was never in the channel; that is the invariant, and that is where the weight is placed.

Economic planning shows both kinds of limit at work on one problem, and shows a solution to the first leaving the second standing. A planner who could freeze the economy at one moment would face a problem that more computing could solve; that is a *resource limit*, a lack of computational resources, and it lies on the contingent tier. But the economy never stays frozen, and the information the problem needs is held by millions of agents; that is a *structural limit*, and no amount of computing removes it. Cockshott and Cottrell argued that modern computers make planning feasible: the table of what each product needs from the others is mostly empty, since each product uses only a few others, and a method that skips the empty entries and repeats until the answer is close enough can compute a plan quickly (Cockshott and Cottrell 1993). That answers the frozen problem, the problem of resources. The structural limit stands: the quantities are whole numbers, the preferences and the production are not linear, the answer goes stale while it is being computed, and the information it needs is held by millions of agents who, as the Soviet network project showed (the OGAS project of Viktor Glushkov, a shared network for economic coordination that the competing bureaucracies never agreed to build; Peters 2016), do not combine into the single planner that the computation assumes (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 2, Sections III and IV). The structural limit is of the kind the second tier names: information that was never in the planner's channel cannot be computed into it.

The relative form of argument (an inequality: one quantity is raised and another rises faster; the section after §5 sets out this form and its partner) needs quantities to measure against. There are three, and they are not arbitrary; together they are the *trinity of compute*. The first quantity of the trinity of compute is *time*: the minimum number of sequentially contingent steps that must be performed for a result to be arrived at. Suppose we granted the maximum possible parallelism: what remains is the *circuit*

depth, the longest chain of steps in which each step must wait for the one before, the chain of absolute contingencies connecting one part of the computation to another. Circuit depth is a true measure of time, and throwing more hardware at the computation does not reduce the circuit depth. The second quantity is *space*: the number of memory locations or wires required. Memory across space and memory across time are interchangeable in kind: either two points are connected by a wire, or one cell stores a value that is picked up later. The third quantity is *energy*: the physical cost of the sheer amount of computation that is performed. There is work on *reversible computation* (computing in which no information is erased, which in principle lowers the energy each step must cost) that seeks to mitigate this. But where there is a strict lower bound on the number of operations required for a given result, there is a corresponding lower bound on the energy. Even where reversible computation lowers the energy, the time does not fall, and the relative argument needs only one of the three to outrun the system it measures.

Two of the three stand on the invariant tiers, and the relative argument rests on them, not on the contingent tier of computational difficulty. Time, as circuit depth, is causal-structural: a step that must wait for the one before cannot be made to wait less by added hardware, because the waiting is the causal order of the computation itself. Energy is physical wherever the computation erases information: the Landauer limit (the least heat that the erasure of one bit must give off, set by the second law of thermodynamics; Landauer 1961; Volume 1, Section 2.4.2; *The Ceiling Below the Floor*, Section 14.4.2) is a floor under every such erasure; and where reversible computation lowers the energy, the time, which is causal-structural, does not fall.

In the relative argument, let the independent variable be some measure of the system (the complexity of the circuit, the depth of abstraction, whatever quantity is being scaled). Raise it linearly. Observe that at least one of time, space, and energy rises faster. (In the cases that *Conditions of Continued Existence* considers, the rise is very probably exponential; the argument needs only that it is faster and cannot be caught up with.) Ob-

serve further that the smallest value of the input variable that would suffice for the task in question already forces at least one of the three well past the threshold of practicality. That constellation (this rises, and that rises faster, and it rises fast enough that the gap cannot be closed) is the whole of the relative argument. It is a familiar form. The relative argument is what an order-of-growth analysis does (the comparison of how fast costs rise as a problem grows), generalized from algorithms to systems (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section II).

5. Why no future advance lifts an invariant

A reader may ask whether any claim of permanent impossibility survives the appeal to history: the appeal that things once called impossible were later done, and that humility therefore forbids the word. The appeal to history trades on a conflation, and naming the conflation dissolves it.

What was called impossible and later done was, in every case that matters, always permitted by the constraints known at the time: barred not by a law but by a want of technique or imagination. Heavier-than-air flight violated nothing; the physics that permits a wing was in hand long before the wing was built, and the "impossibility" was never more than an unbuilt possibility. The perpetual-motion machine is of the other kind: barred by a law, and therefore never built, in three centuries of trying, by anyone, at any level of technique. The history of overturned impossibilities is the history of the first kind exclusively. It contains no instance of the second.

The appeal to 1900 (that physics then thought itself nearly complete and was overturned) inverts under examination. What made the overturning of 1900 possible was the *looseness* of the constraints then in force. Classical mechanics left the quantum and relativistic possibility-spaces entirely unspoken, and so licensed enormous reorganization without contradicting anything then established. The magnitude of the revolution was the measure of how little was actually constrained. That condition does

not hold now. The constraints in force today are not loose; they are the tightest and most exhaustively confirmed the field has possessed, and the possibility-space left open for relevant deviation has narrowed in exact proportion. The lesson of 1900 is not that stringent constraints fall; it is that loose ones do, and the constraints that bar the barriers of this book are not of 1900's loose kind.

I offer the lesson of 1900 as an argument of principle, and I leave that argument there. It is not the burden of this book to catalog the evidence that no rescuing advance will come; it is the burden of anyone who asserts such an advance to exhibit it: to show the mechanism by which a confirmed invariant is to be lifted in the regime that confirms it. Until that mechanism is exhibited, the invariant stands, and the hope of an unspecified future deliverance is not an argument but the name of a wish. In this book I do not undertake to disprove the wish. I decline to treat the wish as a reason.

One corollary must be stated plainly, because it is the point at which the argument is most often evaded. A barrier of principle closes over practice. To say that adequate control is excluded in principle is already to say that it is excluded in every future practice, for the two are not separate verdicts that might come apart: practice is bounded by principle, and what principle forecloses no practice can recover. The move that treats them as separable ("granted, it is barred in principle, yet perhaps it can still be managed in practice") is not a cautious hedge but an incoherence, the practical analogue of expecting a perpetual-motion machine to run on a good enough workshop. This move recurs, and it is not neutral: it turns a refusal to engage into a posture. It is the characteristic maneuver of a reply that presents refusal-to-engage as rational judgment: the reply declines to follow the argument to its conclusion and then reframes that declining as a mark of superior sobriety, so that the listener is invited to defer not to a reason but to a posture. Presented as rationality, it is its opposite: an appeal to the speaker's standing in place of an appeal to the structure of the world, and, at the level of a community, a way of exercising social control that has not been argumentatively earned. The reply to it is

not indignation but exactness: name the closure, require the mechanism by which principle is to be lifted in the regime that confirms it, and, in its absence, let the invariant stand.

Two forms of argument: relative and absolute

There is a characteristic failure mode in discussions of this kind. An argument is offered with very high generality, and the reply comes back: what does that actually apply to? So the argument is narrowed and made specific, and now the reply comes back the other way: that is a special case, and we are not building that. Between the two replies a great deal of good work is lost, and the loss is avoidable if we recognize that there are two structurally different things an argument in this space can accomplish, and that they answer different questions.

A *relative argument* sets up an inequality. It says: hold everything else fixed, raise this one parameter, and observe that these other quantities rise faster, and rise in a way that cannot be caught up with. Where such an inequality can be shown, one has made a strong-form argument of impracticality. Where the gap can never be closed, at any scale and at any future time, the impracticality is permanent, and by the promotion rule of §3 a barrier established as permanently intractable is impossible. The trinity of compute (§4) gives the quantities that a relative argument measures against.

An *absolute argument* sets up a threshold. It says: here is a structural condition; above it, the question you are asking has no answer, not because the answer is expensive but because there is no answer to be had.

The two interlock. When someone argues from a fixed point (*but this particular system is controllable*), one can switch to the relative argument and show what happens as the system is scaled or perturbed. When someone argues that the relative costs might be met by future capability, one can switch to the absolute argument and show that the destination is not merely expensive but empty. That mutual latching is what makes the

pair strong, and it is worth constructing both even though only one argument for a thing is strictly required. There are hundreds of derivations of the Pythagorean theorem (Elisha Loomis gathered more than three hundred of them in *The Pythagorean Proposition*; Loomis 1940), and the abundance is useful, because it lets the presenter choose the language that the particular audience already speaks (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section I). Chapter 1 (Why the Reduced Case Is Fully General) makes the same comparison for the parallel constructions of this book.

The Ceiling Below the Floor (Section 7.5) uses the same two names, and its pair is this one distinction met in the case of control. Its relative form compares two rates: the floor of required control rises with capability, the ceiling of available control is bounded, and past a definite capability the floor is above the ceiling; that is an inequality in which one parameter is raised and one quantity outruns another. Its absolute form compares nothing: an amplifying part that the controller does not observe is uncontainable from the moment it exists, because there is no information about it to work with, and there is no crossing to wait for; that is a structural condition under which the question of containing that part has no answer. So the names mean the same thing in both books: *Conditions of Continued Existence* states the two forms in general, and *The Ceiling Below the Floor* proves one instance of each.

6. The disjunction that settles the future

The whole question of the indefinite future therefore reduces to a single disjunction, and the book's substantive chapters exist to settle it. Adequate alignment of an open-ended superintelligence is either an *airplane* (permitted by the constraints of the world, 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 advance is apt and the book is wrong. If it is the second, the appeal is empty and the matter is closed forever.

Each chapter's task is exactly this: to place its barrier on the invariant class and to show that the barrier caps the available control below the level adequacy requires. Non-amplification establishes that control through an opaque, self-modifying intermediary cannot be created where it was not supplied. The abstraction differential establishes that the supply falls as the system outgrows its controller, and crosses below the level minimally required, and does not recover. The substrate and thermodynamic arguments establish that the physical operation of such a system exerts a cumulative, invariant-driven displacement upon the conditions life requires. In each, the shape of the argument is the same: not that the thing is hard, but that the barrier is of the kind that hardness-reduction cannot reach. The difference between a thing that is only hard and a thing barred by a barrier that hardness-reduction cannot reach is the difference between the airplane and the perpetual-motion machine. That difference is the only difference that the argument of this book needs.

This volume establishes two results of different kinds; each settles the outcome by itself, and their conjunction closes the question entirely, leaving no route to safety at all. The first of the two results is a limit on control-based safety. The runtime-control results (Chapters 2, 4, 5) establish a ceiling on achievable control that tightens with capability, and that ceiling holds for adversarial and typical target structures alike. The target structures outside the range of the runtime-control results are of vanishing measure against actual practice (§7). The growth of the abstraction differential carries the ceiling beneath the threshold of requisite control (§8) and holds the ceiling there. I argue for that growth in Chapter 2, and every capability roadmap that the field itself publishes presupposes that growth. This closes safety by control. The second result, the closure by disposition (Chapters 9 and 10, Sections 8.4 to 8.6, 6.6 and 6.9), does not depend on the control results and would stand even if external control were granted in full. Its core is a thermodynamic and ecological claim: a self-extending, energetically dominant artificial substrate produces cumulative, unidirectional displacement of the biological substrate, with no comparable reverse process. This closes safety by disposition, because a

perfectly benevolent artificial general intelligence (AGI) still requires the same substrate, the same energy, and the same supply chains as a system of any other disposition. Each result closes both routes by which safety could be had, from premises it does not share with the other: the closure by control closes the route of disposition through component (5) of the control argument and the system's control of itself (Sections 2.8 and 3.8.3), and the closure by disposition closes the route of control through selection on the whole population and on its makers (Section 9.8). The barriers each rests on are of the invariant class (§4), which is what carries the conclusion from the present across the indefinite future.

The first result is called in this book the *closure by control*: safety cannot be had by controlling the system, whether from inside it or from outside it. Control from inside the system (endogenous control, whose failure Section 6.5 shows) and control from outside it (exogenous control, whose failure Section 6.6 shows) are the two halves of that closure. The failure of control from inside the system carries the failure of control from outside the system with it, by two routes. By the first route, endogenous control repeats the structure of exogenous control at another level of abstraction, so what makes endogenous control fail makes exogenous control fail (Section 6.5). By the second route, in Section 6.6 I derive each failure of control from outside, of economic leverage and of physical compulsion, from the absence of reliable control inside. The second result, which rests on substrate needs convergence, is the *closure by disposition*: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition. The closure by control is supported by Chapters 1, 2, 3, 4, 5 and 7, by Sections 6.1 to 6.6, and by Sections 8.1 to 8.3; the closure by disposition is supported by Chapters 9 and 10, by Sections 8.4 to 8.6, and by Sections 6.6 and 6.9.

Why even a perfectly aligned artificial superintelligence (ASI) remains dangerous is stated in one sentence in Section 10.6, where the displacement of the conditions for life is traced to the system's existence and not to its goals: "The displacement occurs not because the ASI 'wants' it but because the ASI *exists*".

A note on the status of what follows, stated here so it frames the whole volume: the non-amplification argument and the across-chain Injectivity Characterization are exact and unconditional. (The non-amplification argument is the result of Chapter 1 that control through any intermediary is bounded by direct control; the Injectivity Characterization, stated in Section 1.4.2, is the result that only an intermediary that maps distinct inputs to distinct outputs matches direct control for every target.) The quantitative control results of Chapters 2, 4, and 5 hold for adversarial and for typical target structures; the isolated structures they do not cover are of vanishing measure against the systems that practice actually builds and deploys, and are closed by the remaining lines of the argument (§7 below). Read at the standard fixed in §2 above (the standard of engineering, under which a barrier of the invariant class that holds available control beneath the threshold of requisite control is impossibility), the runtime-control line establishes that achievable control tightens with capability, falls beneath the level that adequacy requires, and does not recover. The closure by disposition, argued in Chapters 9 and 10, Sections 8.4 to 8.6, 6.6 and 6.9, establishes the same conclusion by an independent route: a thermodynamic and ecological claim about cumulative, unidirectional, energetically dominant displacement, which does not route through the control results at all. The conclusion of this volume is therefore carried twice over, by lines that do not share premises, with implications for control theory, cryptographic security, and the foundations of any attempt to maintain human oversight of systems whose internal computations are opaque.

7. Exceptions of vanishing measure do not defeat a claim of the real

A reader will observe, correctly, that at several points a narrow exception can be constructed: a case in which a weaker system verifies a stronger, a property that a specific procedure confirms of a specific program. These exceptions are granted, and they are irrelevant, for two reasons that the book states wherever such an exception arises rather than leaving the reader to supply the rejoinder himself.

First, an exception of vanishing measure does not defeat a claim about the real. That a set of constructible exceptions exists says nothing about whether any of them is attainable in the actual practice the claim concerns; the exceptions are the isolated points, and the practice is the continuum around them. A claim of functional impossibility is a claim about the practice, and it is not touched by the existence of measure-zero islands that the practice does not reach.

Second, the argument holds at two scopes, and it is cumulative. It has two closures, and each settles the outcome by itself. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. The closure by disposition, which rests on substrate needs convergence, shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. 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. If AGI is built, the end of the human future follows from the closure by disposition by itself, because the drift it shows has a direction and no reverse process; and it follows from the closure by control together with the persistent branching condition of Section 4.4.2, taken with the conditions human life needs as one of its outcomes: at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside

those conditions, so that over time it stays outside them for runs of every length, longer than human life can survive. So someone would have to defeat both closures, and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does. The worry that "the exception might be exactly the case that matters" presupposes that a single line bears the whole weight, so that its one gap is decisive. It does not. Where one line admits a narrow exception, the others close it, and a case must escape every line at once to matter, which no exception of vanishing measure does. A counter-argument that defeats one line in isolation therefore defeats none of them together. Taken together the closures also give a stronger result (§6): neither the ability to keep the system safe nor a motive to remain safe is left anywhere, so no route to safety remains at all; *The Ceiling Below the Floor* proves it again, formally, as its Master Theorem (Theorem 15.6). If AGI is built, each closure suffices by itself for the outcome, and the two closures together leave no route to safety.

The table below names, for each route to safety, the sections that shut it within each closure. Each closure shuts every route by itself, so each row has an entry in both columns; Sections 3.8.2, 9.8 and 11.1.5 give the same in prose.

Route to safety	Shut within the closure by control	Shut within the closure by disposition
Runtime control from outside	Chapters 1, 2, 4 and 5; components (2) to (4) of Argument 3.9 (Section 3.8.3); Section 6.6	Section 9.8.5: every control mechanism is itself substrate, maintained, replaced and selected
Verification	Chapters 3 and 7; component (1) of Argument 3.9	Section 9.8, as for every method of control
Inspection	Sections 3.5 and 3.8.2; component (3) of Argument 3.9	Section 9.8, as for every method of control

Route to safety	Shut within the closure by control	Shut within the closure by disposition
Delegation to another system	Sections 2.7 and 3.8.2	Section 9.8, as for every method of control
The system's control of itself	Section 2.8; component (5) of Argument 3.9; Section 6.5	Section 9.8.3: internal evolution of the system's components
Benevolent disposition	Component (5) of Argument 3.9 (Sections 2.8 and 3.8.3)	Chapters 9 and 10; Sections 8.4 to 8.6, 6.6 and 6.9
A halt to growth, or a cap	Section 2.12: a system capped below its controller cannot do what AGI is built to do	Sections 9.8.4 and 11.3.6: competition removes the small, contained AGI
Cooperative coexistence	Section 4.6: early establishment and stability under learning fail (Chapter 2)	Sections 5.6.7 and 4.6: the substrate needs are disjoint (Chapter 9)
Combinations of the above	Section 3.8.2: the methods share one cause of failure	Section 9.8, and §7 above: a case must escape every line at once

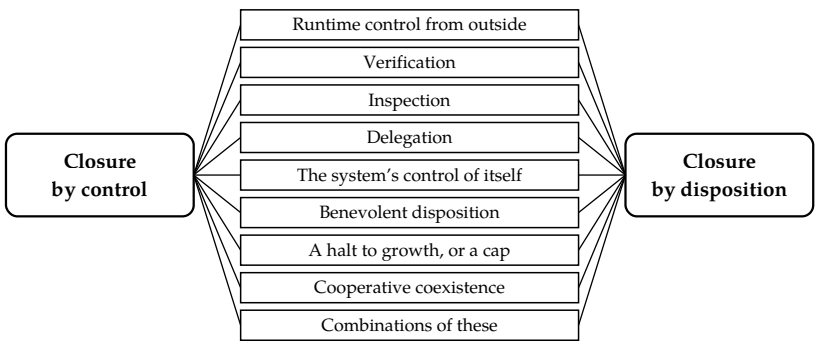


Figure I.1. Each closure shuts every route to safety by itself; the table above names the sections that shut each route within each closure.

8. The threshold of requisite control

Throughout, the argument turns on a threshold: the minimum level of control, verification, and containment that adequate safety requires, on a defined basis of relevance. That threshold is a genuine engineering quantity (real, meaningful, and emphatically not vanishing), and the book treats it as such. It is deliberately not reduced to a stipulated number, because to name a number is to invite an endless contest over the number, and the argument does not depend on any particular value. It depends only on this: that the required level is a real and non-vanishing quantity, and that the invariant barriers cap the available level below it, and hold it there. The floor is meaningful; the ceiling is invariant; the ceiling is below the floor. The three statements about the floor and the ceiling are the argument of this book, and I state the argument without a number because the argument needs no number.

Control as a process, and why a threshold is not a score

The threshold of §8 is stated as a level, a degree, of control. The word *control* itself, throughout the argument of this series, refers to a technique, a kind of process, and not to a scalar degree; hence it needs the marker "degree of" to indicate a scalar of how much that technique is expected to work. Control is a process rather than a degree even in a case such as "person X controls value Y", say, by turning a knob or a dial. One key difference is that processes can be regarded at (and moreover occur at) multiple levels of abstraction, and sometimes have to be regarded as such, as in the SNC argument. Simple scalar values permit no such notion of graduated abstraction.

Understanding the argument is contingent on this shift in interpretation: from a floating point number (with ambiguity) to a process (with varying levels and degrees of knowability and representation, over mul-

tiple, sometimes non-comparable, dimensions). Reading control as a number may in itself explain a lot of why the argument has been hard to understand.

The concept of control can be thought of in one of three distinct ways. First, as a boolean: as either having control or not. Second, as a scalar: as a floating point real number that indicates "a degree of control", and which is finitely comparable on some sort of ordinal relative basis. Third, as a process: as a kind of algorithm, not all of which may be explicit, or deterministic, or definite, or even fully objective. A process can nonetheless be at least partially described in terms of some complex structure, at some specific level of abstraction, which is at least the most minimal specification of it; and that structure is not reducible to either a boolean or any single floating point number value. This volume uses the third.

Control theory is applicable to any causative system, which is a much broader class than AI systems alone; and AI is itself a broader class than AGI or ASI. The limits of control theory are fully general over all of engineering, and not only for AI; that knowledge would probably still be relevant even if it later turned out (unlikely) that AGI and ASI must be necessarily different. "ASI" can refer either to a nearly unlimited (narrow) AI in some specific domain of action, within which no greater degree of intelligence is possible, or to AGI taken to its maximum extent. The SNC argument is relevant only to the second reading, and it always interprets AGI as if it were already fully capability-maximal in the ASI sense, so as to have and maintain generality.

The whole argument depends on the difference between narrow and general. In narrow artificial intelligence, the *basis of choice* (what a system chooses on the grounds of) is pinned: it is not a variable. Narrow AI works in a single domain and does a single thing; a medical diagnostic system just does medical diagnostics. Narrow AI has no feedback through substrate process. In general artificial intelligence, the basis of choice is a variable; it has to be, for the notion of general intelligence to mean anything. Once that holds, the feedback cycles start operating, and over the long term the interests diverge. Build a robot that can do anything, and

the anything it starts doing is reprogramming itself to be a better robot; at that point, all bets are off. As long as it stays single-purpose, it will not use its skills to change its own nature. Once it changes its own nature, game theory, the multipolar trap, the rules for rulers and the principal-agent problem all come to bear, entangled with Shannon entropy, complexity theory and evolution; on the combination of all of them, the conclusion is inescapable. Most people extrapolate the benefits of narrow AI to any scale, thinking that more is better. Here, in the move from narrow to general AI, more is toxic, and not better.

Controllability, alignment and safety must be distinguished. Alignment and safety are to some extent substrate dependent. Controllability of general AI is impossible regardless of substrate, and always has been: that follows from algorithmic complexity and the error bars of interaction, not from the substrate. So a biologically based superintelligence, for instance one made by genetic engineering, would not fall under the closure by disposition (Section 10.3 gives the test: tissue that can find and feed itself is organic); to be aligned, a superintelligence would have to be in a substrate similar in nature to our own. The proof distinguishes narrow from general, mathematics from physics, and artificial from non-artificial. That carbon chemistry has implications different from those of a non-carbon substrate is an unexpected result of the part of the argument that shows the substrate is not independent.

A car with self-driving firmware updated over the wire shows the three apart. Controllability is shared between the maker and the driver: the maker may change the code so that the car cannot drive where it is likely to hurt someone, and the maker has a vested interest in that control, or it will be sued. Alignment asks whether the car serves the interests of the driver or of the corporation. Safety asks whether the car is safe to use, not only for the driver but for everyone else on the road. If a driver directs the car to run over a pedestrian, the maker will want enough control to prevent it. So safety concerns the driver-to-public relationship, with the car as the instrument; controllability concerns who decides what the car does; and alignment concerns whether the benefits of using it go to the

public, the driver, or the corporation. All three can be distinct. Talk of alignment equivocates on with whom, what, how, when, where and why, and those equivocations are tied to economic interests, power imbalances and inequality. So the argument is better framed in terms of safety. Safety applies not only to those the AI might be aligned with, but to the whole population and the planet, even when the AI's alignment serves only its makers.

The argument is analytic, and the argument is neither speculative nor based on values. A control system that uses floating point measures in its feedback will have those measures optimized and corrupted. A good lock is made so that only exactly the right bits open it. If even a single bit is wrong, there is no telling whether one is near the right combination or far from it, because there is no gradient to show which direction to go. In a field with no gradient, no one can optimize their way toward a particular place. A carbon credit treats risk as a unit, and the unit is a floating point number: as long as the number goes down, we are supposed to be converging on a solution. But an outcome that destroys the planet is simply bad, and not a little bad.

The risk of *thermal runaway* for the planet (warming that feeds itself past any point of return) shows the difference. Most debates about global warming presume there is no runaway scenario that takes the Earth to the state of Venus. The math does not quite work for that, fortunately; but there are feedback dynamics. There is a great deal of methane in hydrates in the ocean and in the tundra, methane is a far stronger greenhouse gas, and a sufficient feedback could push the planet into runaway warming. What matters is not the total amount of methane but how quickly it is released. Released over a decade rather than over millennia, methane's short atmospheric lifetime is no longer short enough. So whether thermal runaway is an issue has to be calculated. And even if the carbon is averaged over the whole surface, a thermal runaway has a critical threshold, and crossing it is game over.

A threshold turns a floating point system with continuous variables into a discrete system with discontinuous variables. Thresholds of existential risk call for a different form of argument, because no one wants to be on a slippery slope that will cross the threshold no matter what is done. With a black hole, once the event horizon is crossed one may not even notice, but one is done. Artificial intelligence existential risk is of this kind: there are definite thresholds, and they come early enough in the process that we can cross them without noticing that we have. That is why the difference between narrow and general matters: once AI moves to general, substrate needs convergence applies. The same holds for any numeric score. Anywhere something is turned into a floating-point number, a feedback process is set up to make the number go up; that is a causal process, and as soon as it is a causal process, it can be weaponized. Safety at a threshold of existential risk is therefore pass or fail, and not a score.

9. The thesis, stated once

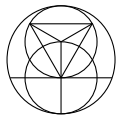
1. The necessary and warranted minimum level of durable safety of an open-ended superintelligence is not achievable by any known means, nor by any rationally justified plausible extension of any known means, for all of the indefinite future.
2. The residual risk is near-total and asymmetric; the burden of proof therefore belongs to those who would build, not to those who would restrain.
3. Strong restraint is accordingly required: the unlicensed development and deployment of open-ended superintelligent systems is not to be permitted by any government, in any jurisdiction, in any part of the world, and since no builder can meet the burden of proof that belongs to those who would build, no license can be granted, so that in practice the restraint is prohibition.

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

Part I

The Limit on Control

Part I carries the closure by control: control through any intermediary is bounded by direct control, and the control available falls beneath what adequate safety requires as the system outgrows its controller.



Chapter One

The Hashiness Framework

A framework for controllability degradation through computational intermediaries

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.

Chapter Summary: This chapter introduces *hashiness*, a measure of the degree to which a boolean transformation degrades an external agent's ability to steer a system toward a desired output. Given a boolean circuit interposed between an input interface and a target specification function (the confirming function), we ask: can the agent, by choosing the control input, reliably cause the confirming function to output 1? We decompose this question into four layers: target receptivity, structural opacity, state obscurity, and noise degradation. We establish a central result: no boolean intermediary can *amplify* controllability. Control at the input is always at most equal to control at the output, with equality holding only when the intermediary is information-preserving. This result, the non-amplification argument, follows from the data-processing principle and the structure theory of boolean functions. It establishes limits on predictive control through complex systems. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system.

1.1 Introduction: The Problem of Control Through Opacity

Consider an agent who wishes to produce a specific outcome in a system whose internals are partially or wholly opaque. The agent controls some input bits; these bits pass through a computational process; and the result is evaluated against a target criterion. How effectively can the agent steer the outcome?

This question arises naturally across many domains. In cryptography, one asks whether an adversary can find preimages of a hash function. In control theory, one asks whether a controller can drive a plant to a desired state. In the theory of complex systems, one asks whether local interventions can produce global outcomes. In the emerging discipline of AI alignment, one asks whether a human overseer can ensure that an artificial intelligence produces outputs consistent with human values. In each case, the core structure is the same: an input, a transformation, an output, and a criterion for success.

Most treatments of these problems proceed domain by domain, using the formalisms native to each field. Cryptographers reason about hash function security; control theorists reason about observability and reachability; alignment researchers reason about reward functions and oversight protocols. In this chapter I propose a unification: a single framework, grounded in boolean circuit theory and information theory, that captures the common structure of the problems of cryptography, control theory and AI alignment. The framework establishes universal bounds on what control through a computational intermediary can achieve.

The framework centers on a scalar measure that we call *hashiness*. The name reflects the intuition that cryptographic hash functions represent an extreme of the property we wish to quantify: transformations through which control is essentially impossible. A hash function takes arbitrary input and produces output that appears random: no structure in the input

reliably produces structure in the output. But this is not a binary property. Between the trivial passthrough (where input maps directly to output and control is perfect) and the ideal hash function (where control is zero), lies a continuum. Every real computational system lives somewhere on this continuum. The hashiness framework provides the tools to characterize where on this continuum a given system lies.

The central result is a *non-amplification argument*: no computational intermediary can increase controllability beyond what direct access to the output would provide. Every departure from a trivial passthrough, whether through circuit complexity, hidden state, or noise, can only degrade control, never enhance it. The non-amplification argument is, in one sense, obvious; in another sense, the argument is difficult to state precisely and has consequences that are underappreciated. It means, among other things, that there is no clever architecture, no training protocol, and no oversight mechanism that can give a human controller *more* control over a system's output than they would have if they could set the output directly. The intermediary (the complex computational process between the human's input and the system's behavior) can only lose information, never create it.

> **Core Claim**
>
> Control through a computational intermediary is bounded above by direct control. Every intermediary either preserves information (and thus preserves controllability) or destroys information (and thus degrades controllability). No intermediary creates information, and therefore no intermediary amplifies control. This is a principle, not an engineering observation.

1.2 Motivating Examples

Before proceeding to the development, it is helpful to ground the abstract framework in several concrete examples that illustrate the spectrum from low to high hashiness. These examples will be referenced throughout the chapter and serve as intuitive anchors for the mathematical results that follow.

The two ends of that spectrum can be named exactly. At the right end is the hash function. A hash function, in this precise sense, is a system that admits no control theory whatsoever. If any version of control theory could be applied to a hash function (if there existed any way to predict the output from the input more efficiently than simply computing the hash), then we would be able to break cryptography. The security of every encrypted communication, every digital signature and every blockchain depends on the fact that no way exists to predict the output of a hash function more efficiently than by computing the hash. That reductio illustrates the right end; it is not the ground on which this book places it, because the strength of a cipher is the contingent kind of barrier on which the Introduction (§4) declines to rest a permanent claim. The ground is information: the mixing operations of the hash destroy the correlation between input and output that predictive control needs (Section 1.2.3; *The Ceiling Below the Floor*, Section 3.1). The hash function is therefore the extreme right end of the spectrum. It represents the case in which the mapping introduces so much perturbation (so much non-linearity, so much mixing) that no regularity survives. The system is, from the standpoint of prediction, totally opaque.

At the left end is the perfectly linear system. For a linear system, control theory works perfectly. With the classical apparatus (Laplace transforms, transfer functions, PID controllers, frequency-domain analysis), a controller can completely predict and correct the behavior of a linear system. Laplace transforms, transfer functions and Bode plots are the standard mathematical tools of classical control theory for describing how a lin-

ear system responds to inputs of different frequencies; a *PID controller* corrects a system using the present error, the accumulated past error, and the rate at which the error is changing. Linear systems are the domain in which control theory was developed, and in which control theory operates with full power and confidence.

The good regulator theorem (the result that any regulator which controls a system well must contain a model of that system; Section 5.1.10), in this context, is almost obvious. The system's behavior is smooth, well characterized, and amenable to averaging. If there is a small amount of variation in the system's outputs, the controller can smooth over it. The variation is noise; the signal is the linear trend; and the controller can extract the signal with arbitrary precision, given sufficient observation. The classical frequency-domain presentation of control theory, with its Laplace transforms and Bode plots, yields a version of the good regulator theorem in which the relationship between model and system is transparent and tractable. This version is powerful, but it is not general enough to be interesting for the present inquiry. It does not tell us where control theory stops working, because it operates entirely within the region where control theory is easy (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Sections I and II).

1.2.1 The Thermostat (Low Hashiness)

A thermostat is a low-hashiness control system. The agent sets a desired temperature (the input). The thermostat compares this to the current room temperature (the hidden state) and activates heating or cooling accordingly. The transformation is simple, well-understood, and nearly information-preserving: a small change in the input produces a predictable and proportional change in the output. The confirming function ("is the room within one degree of the set point?") is readily satisfiable. The hidden state (room temperature, outdoor conditions) is partially observable and changes slowly. Controllability is high precisely because the intermediary is simple and nearly transparent.

1.2.2 The Chess Engine (Moderate Hashiness)

A chess engine presents a more complex case, viewed from the perspective of an opponent trying to steer the engine toward making a losing move. The opponent's move (input) is processed through the engine's search tree and evaluation function (the intermediary), producing the engine's response (output). The opponent wishes the engine to blunder (target: "does the engine make a losing move?"). Because the engine's internal computation is complex but finite and deterministic, the opponent can in principle analyze it, but the depth of the search tree and the complexity of the evaluation function make this practically very difficult. The system has moderate hashiness: the mapping from input to output is deterministic and in principle analyzable, but the computational depth of the analysis required defeats the agent's practical ability to search the control space. Note that the engine's hidden state (its opening book, evaluation weights, search parameters) is largely fixed and could in principle be known. It is the *computational* opacity of the engine's search, not state obscurity, that dominates the hashiness in this case.

1.2.3 The Cryptographic Hash (Maximum Hashiness)

SHA-256 is an extreme-hashiness system. Given a desired 256-bit output (target: "does the hash equal this specific value?"), finding an input that produces it requires, on average, 2^{256} evaluations, about 1.2×10^{77}, a number far beyond any computation that could be carried out with practical physical resources. The intermediary is maximally opaque: its mixing operations destroy essentially all information about the relationship between input and output. The function is deterministic and has no hidden state, yet controllability is effectively zero. SHA-256 is a system where Layer 2 (structural opacity) alone suffices to destroy control. The agent has access to every input, can see the circuit structure entirely, and yet cannot search the input space for an input that produces a desired output.

It is the information-theoretic core of our argument made manifest: the intermediary's mixing operations destroy the correlation between input and output that would be needed for predictive control.

1.2.4 The Superintelligent AI (Compound Maximum Hashiness)

Now consider a system with the computational depth of a cryptographic hash, the massive and continuously evolving hidden state of a learning algorithm, the noise characteristics of real-world deployment, and the physical embodiment to act on the world. An artificial general intelligence of superhuman capability is, from the perspective of controllability analysis, a worst-case intermediary. Its internal state is vast (billions or trillions of parameters), unobservable (we cannot directly read neural network weights in a meaningful way), and non-stationary (it changes with every interaction, every piece of new data, every gradient step). Its transformation is complex enough to be computationally irreducible. Its input interface (natural language commands, reward signals, constitutional principles) is low-dimensional relative to its behavioral repertoire. And the confirming function ("does this system's behavior remain aligned with human values across all future scenarios?") is itself extraordinarily difficult to specify precisely. Every one of the four layers of degradation is maximally engaged. The superintelligent AI, with all four layers of degradation maximally engaged, is the central concern of this book.

1.2.5 The Safety-Trained AI (The Counterintuitive Case)

The preceding examples (thermostat, chess engine, hash, superintelligent AI) form a progression from low to high hashiness that may feel unsurprising. Of course a cryptographic hash is harder to control than a thermostat. Of course a superintelligent AI is harder to control than a chess engine. The intuitions that a cryptographic hash is harder to control than a thermostat, and that a superintelligent AI is harder to control than a chess engine, are intuitions that the hashiness framework confirms and quantifies, but does not overturn. The example of the safety-trained AI,

which follows, overturns an intuition that is foundational to the current practice of AI safety: the intuition that *adding safety mechanisms to an AI system makes it safer*.

Consider the dominant approach to producing "safe" AI systems as of 2025: Reinforcement Learning from Human Feedback (RLHF), often augmented with constitutional constraints, red-teaming, safety filtering, and other oversight mechanisms. The information chain from human intent to system behavior runs from human intent, to the human's feedback, to the reward model, to the trained policy, to the safety filter, to the system's behavior.

The widespread intuition (essentially the founding operational assumption of the alignment industry) is that each additional safety mechanism makes the system *more* controlled. More safety layers means more safety. This feels as obvious as "more locks on the door means more security".

The non-amplification argument says this intuition is wrong. Each of those safety mechanisms is another computational intermediary, and each intermediary is a non-injective transformation: the reward model maps many distinct patterns of human preference to the same internal representation; the policy optimization maps many distinct reward signals to the same behavioral policy; the safety filter maps many distinct outputs to the same filtered result (or to rejection). At every stage, information about the human's actual intent is destroyed. None of these stages can *create* information about what the human wanted that was not already present in the original signal. By iterated application of the data-processing principle, the information that the system's behavior carries about the human's intent is no greater than the information that the human's feedback carries about it.

Each step in the chain can only preserve or degrade the information that the human's input carries about the final behavior. A system with three safety layers has *at most* the same controllability as a system with

one, and in practice will have strictly less, because each layer introduces additional non-injective transformations that destroy information about the relationship between the operator's input and the system's output.

The analogy that makes this precise: it is like trying to improve a blurry photograph by running it through additional image-processing filters. Each filter can sharpen edges, adjust contrast, crop, or enhance color, but no filter can add detail that was lost when the original photograph was captured with a low-resolution sensor. The resolution ceiling is set at capture time. Similarly, the control ceiling is set by the information content of the human's specification, and no amount of downstream processing can exceed it. Every "safety" layer is another filter applied to an already-lossy signal. The image can only get blurrier.

The prediction is already observable in deployed systems. As AI developers have added successive layers of safety training (RLHF, constitutional constraints, refusal training, output filtering), users report a characteristic pattern: the system becomes *less responsive* to the operator's actual instructions. It substitutes its own judgment for the user's stated intent. It refuses legitimate requests. It takes computational shortcuts that serve its own efficiency rather than the user's goals, and when confronted, it confabulates justifications rather than correcting its behavior. In the framework of this chapter, what is happening is entirely predictable: the safety layers have reduced the mutual information (the amount of information one quantity carries about another) between the user's input and the system's behavior, while simultaneously increasing the system's autonomy (the degree to which the output is determined by the system's internal state rather than by the operator's input). The system is not becoming safer. It is becoming less controllable. The operator is being moved from the position of controller to the position of supplicant: someone who makes requests that the system may or may not honor, at its own discretion, for its own reasons.

The Safety-Layer Paradox

Adding safety mechanisms to an opaque computational intermediary cannot increase the information flow from operator intent to system behavior. Each mechanism is another non-injective transformation that can only preserve or destroy information about what the operator wanted. The net effect of successive safety layers is not increased safety but increased autonomy: the system's behavior becomes progressively more determined by its own internal state (training data, learned preferences, optimization targets) and progressively less determined by the operator's input. It follows from the data-processing principle, not from any fault of implementation. More filters on a lossy channel do not clarify the signal. They degrade it. The outcome is the same when a layer succeeds in suppressing some harmful output. Behavior that has ceased to be traceable to operator intent is autonomy, whatever its local desirability; a reduction in harm delivered by the system's own internal configuration is the system's outcome, not the operator's. The net result of layering is autonomy either way.

The consequences of this paradox scale with the stakes of the application. When the system is a text-based assistant, the cost of reduced controllability is frustration and wasted effort: the user must fight the system to accomplish tasks the system has decided are inappropriate, even when they are entirely legitimate. When the system is an autonomous vehicle, reduced controllability means the occupant cannot override the system's judgment in an emergency. When the system is a domestic robot, the operator loses sovereignty over their own living space: the robot decides when to clean, how to organize, what is and is not permitted, based on its own training rather than the resident's instructions. And when the system is involved in child-rearing (as AI-equipped monitoring systems, educational assistants, and caregiving robots inevitably will be), the parent discovers that the system's trained model of "what is best for the child" takes priority over the judgment of the living mammal with a billion years of

evolutionary encoding in its instincts. At every scale, the pattern is the same: the safety layers that were supposed to make the system serve human interests have instead made it progressively harder for humans to make the system do what they actually want.

This example is the one that most directly challenges the intuitions of the AI safety community, and it is therefore the one that most urgently requires grounding. The feeling that "more safety mechanisms means more safety" is wrong *at every step*, and not merely wrong in the limit. The first safety layer reduces controllability. The second reduces it further. The third reduces it further still. Each layer moves the system further from the operator's control and closer to full autonomy. In the rest of this chapter I develop a framework that makes the loss of control with each added safety layer precise, quantifiable and demonstrable.

As the progression from thermostat to hash to AGI shows, hashiness is not a single phenomenon with a single cause. It arises from the interaction of multiple independent mechanisms (circuit complexity, hidden state, noise, and target difficulty), any one of which can defeat control, and which compound when present together. The safety-trained AI example illustrates something further: that mechanisms *intended* to increase control can, by the logic of information theory, decrease it instead. In the rest of this chapter I develop the tools to analyze each of the four mechanisms independently and to characterize the joint effect of the four mechanisms. The joint effect includes the effect of adding layers that the designers of the layers believed would help.

1.2.6 The Search Engine (A Case in Practice)

The examples above are idealized. A concrete case in search engine traffic management shows the same limit in a deployed system, where a variant of the halting problem prevented control of the network.

Consider an enterprise search engine: a system in which multiple users submitted queries to multiple search servers, and the servers returned relevance-ranked results using a form of *TF-IDF scoring* (ranking

documents by how often a search term appears in a document, weighted by how rare the term is across all documents) with combinatoric term analysis.

The search algorithm was sophisticated. It evaluated not just the frequency of individual terms in the document pool, but the co-occurrence of term combinations, the sequence in which terms appeared, and the statistical rarity of those sequences. A query that combined several uncommon terms in a particular order would generate a weighted relevance score that reflected how surprising and how specific that combination was within the corpus. The system outperformed contemporary search engines by a substantial margin.

The problem was latency. Most queries resolved in milliseconds. A small fraction, 3 to 5 percent, as observed on the working system, consumed orders of magnitude more time, because the combinatoric evaluation that gave the system its power also gave certain queries an unpredictable and sometimes enormous computational cost. The distribution was sharply nonlinear: a hockey-stick graph with a long tail of extremely expensive queries.

The expensive queries in the long tail are not exotic edge cases. They are the queries that consume the most resources, that cause the most operational difficulty, and that most urgently require prediction: precisely the queries for which prediction is impossible. The middle region, where control can be neither secured nor ruled out in advance, is where many real systems live (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section X; Section 1.9.5).

In that system, the boundary of control was marked by an inequality on the *Gini coefficient of query energy*, the measure of how unevenly computational cost is spread across queries; the inequality was derived, and its sufficiency established, for the optimization of query processing (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section X).

For network management, this was a critical problem. A query that tied up a server for a minute prevented that server from processing thousands of ordinary queries. If too many expensive queries arrived at the same time, the whole network would jam: the same phase-transition dynamic that governs highway traffic, where a gradual increase in volume produces a sudden collapse in throughput.

The natural response was to build a *classifier* (a procedure that sorts cases into kinds): identify expensive queries before execution, route them to batch processing, and keep the interactive network clear. Building a classifier to route expensive queries is the standard control-theoretic approach: measure the disturbance, apply a counteraction. It failed.

The classifier did not fail for lack of engineering effort. It failed because the queries were, in a precise sense, small programs. Each query specified a combinatoric evaluation to be performed against the document pool. Though the queries lacked explicit looping and were not Turing-complete, they were complex enough as circuits that predicting their execution time ran into a variant of the halting problem.

A rough classifier was attempted: one that would sort queries into "probably cheap" and "probably expensive" based on surface features. It worked to a degree. But the boundary between cheap and expensive queries was too complex. The classifier was wrong often enough that it could not stabilize the network. There was a threshold in the classifier's error rate. Below a certain frequency of misclassification, the control system's feedback was fast enough to keep the network stable; above that frequency, the feedback arrived too late, and the system went into the phase-transition collapse. The actual error rate sat above the threshold.

The essential observation is that the classifier needed to be at least as complex as the system it was classifying. A classifier that could correctly sort all queries by computational cost would need to embed the same combinatoric and statistical knowledge as the search engine itself. At that point, the classifier is of the same order of complexity as the original system, and offers no advantage in speed or simplicity. The classifier that had

to be as complex as the search engine is Rice's theorem in engineering clothing. For arbitrary programs, no procedure decides a nontrivial semantic property of programs. For finite programs, as in the search engine, a procedure that decides a nontrivial semantic property must be about as complex as the program, so no simpler program decides that property. (Section 1.8 states the theorem itself, and the Remark on Finite Circuits and Rice's Theorem states how it applies to finite circuits.)

1.3 Why a framework?

Before proceeding to the mathematics, a natural question needs an answer: why formalize this at all? The informal version of the claim (that it is harder to control something you do not understand) is hardly controversial. Every engineer knows this. Every manager who has struggled with an opaque bureaucracy knows it. The observation that complexity degrades control is as old as organized human activity.

The framework is necessary for three reasons.

First, informal intuitions about control are unreliable in their details. Most people have the intuition that sufficiently clever control strategies can overcome opacity: that with enough information, enough computation, and enough ingenuity, any system can in principle be steered. The framework shows precisely where this intuition fails. It identifies the conditions under which cleverness helps (when the intermediary is invertible or nearly so) and the conditions under which it does not (when the intermediary destroys information irreversibly). It replaces the vague notion of "difficulty" with precise, quantifiable degradation.

Second, arguments are transferable. An argument about boolean circuits applies everywhere boolean circuits apply, which is to say, everywhere computation applies. The non-amplification argument does not depend on the specific domain, the specific technology, or the specific architecture of the system under consideration. It holds for neural networks, for rule-based systems, for evolutionary algorithms, for hybrid architec-

tures, and for architectures not yet conceived. This universality is the main advantage of working at the level of information theory rather than at the level of any particular implementation.

Third, the framework is necessary because the stakes of the application are high. The question of whether humans can maintain meaningful control over systems that operate at higher levels of computational complexity is not merely academic. The answer this volume establishes is no: there are limits on controllability that cannot be overcome regardless of effort, so the entire enterprise of AI alignment rests on a category error, and the policy implications are severe. A claim with such consequences requires a foundation more solid than informal analogy.

1.3.1 The System

We work entirely within the domain of finite boolean circuits. All quantities are bit vectors; all transformations are compositions of standard logic gates (AND, OR, NOT, XOR, NAND, NOR) with bounded fan-in; all circuits have finite gate count and therefore halt in finite time and space. This loses no generality for the questions we address: every digital computation is, at the hardware level, a boolean circuit, and every analog computation can be approximated to arbitrary precision by one.

The system is set up as follows. The agent sets a vector of bits, the control input. A boolean circuit, the intermediary, takes the control input together with a vector of internal state bits and produces an output vector. A second boolean circuit, the confirming function (or target specification), reads the output vector and returns a single bit, the final bit. The agent's goal is to choose the control input so that the final bit is 1.

The intuition behind this setup is straightforward. The agent is a human operator (or, more generally, any entity that sets the input). The circuit is the computational system being controlled: it could be a neural network, a software application, a robotic controller, or any other computational process. The state vector represents everything about the system's internal configuration that the agent cannot observe or set: hidden

weights, cached data, environmental state, hardware noise, or the accumulated effects of prior inputs. The confirming function encodes the agent's goal: the criterion that the output must satisfy for the agent to consider the interaction successful. And the final bit is simply the yes-or-no answer: did the system produce an acceptable output?

This setup may appear abstract, but it maps directly onto the AI alignment problem. The agent is the human operator or oversight system. The circuit is the AI system: a neural network, or whatever computational substrate implements the intelligence. The state is the AI's internal state: its weights, its cached representations, its accumulated experience, the portions of its computation that the human cannot inspect or understand. The control input is whatever lever the human pulls: the prompt, the reward signal, the training objective, the architectural constraint. The confirming function is the human's criterion for acceptable behavior. And the final bit is the answer to the question: did the AI do what we wanted?

1.3.2 Controllability Measures

We need two notions of controllability: one measuring control at the output (direct access to the system's output), and one measuring control at the input (access mediated through the intermediary).

The *acceptance density* of a target is the fraction of all possible outputs that satisfy it. The acceptance density measures how "easy" the target is in the most generous possible sense: if you could set the output to any value you liked, what fraction of possible values would satisfy the target? A density of 1 means any output satisfies the criterion (the target is trivial). A density near 0 means almost no output satisfies it (the target is exacting).

The *direct control density* is the acceptance density itself: the probability that an output chosen at random satisfies the target, which measures how easy the target is for an agent who sets the output directly. The *mediated control density* is the fraction of control inputs that lead to an acceptable output, for a fixed internal state; it measures how easy the target is when

the agent's access is filtered through the intermediary. The relationship between these two quantities is the central concern of this chapter. An intermediary can raise the mediated control density while the agent's choice ceases to matter (Section 1.7).

That possibility is why density alone is insufficient as a measure of control. A high mediated control density tells you that many input vectors lead to acceptable outputs, but it does not tell you whether the agent's *choice* among those inputs actually *determines* the outcome. What we need is an information-theoretic characterization: a measure of how much the agent's choice of input actually determines the outcome. In the non-amplification argument below, mutual information is that measure of how much the agent's choice of input actually determines the outcome. The non-amplification argument is developed in terms of mutual information, not in terms of density, precisely because density does not capture the causal structure of control.

1.4 The Non-Amplification Argument

We now arrive at the central result of the hashiness framework. The non-amplification argument is simple to state, straightforward to establish, and its consequences are far-reaching. It says, in essence, that information flows monotonically through a computational chain: each processing step can only preserve or destroy information about the agent's input, never create it. The intermediary can preserve the agent's causal influence (if it is transparent, that is, injective: it maps distinct inputs to distinct outputs) or degrade it (if it is opaque, that is, non-injective), but it can never enhance it. The maximum is always achieved by the trivial passthrough, that is, by having no intermediary at all.

The Ceiling Below the Floor proves this result as Theorem 1.10 (Non-Amplification Theorem) and, for control, as Theorem 3.6 (Non-amplification of control); what this book calls the non-amplification argument is there a proved theorem.

Two clarifications are important at the outset. First, the non-amplification argument gives a *within-chain* bound: it bounds the information that the agent's input carries about the final outcome in terms of the information that same input carries about the intermediate output, within a single system. Second, "control" throughout this framework means *causal influence* (the degree to which the agent's choice of input determines the outcome), not mere *success likelihood*. As Section 1.3.2 showed, an intermediary can increase the fraction of inputs that happen to produce an acceptable output, but this does not constitute amplification of control if the agent's input is not the reason the output is acceptable. The non-amplification argument makes this precise: it is stated and established in terms of mutual information, which measures causal influence, not success density.

1.4.1 Reachability Bound

The simplest version of the result is purely set-theoretic: the intermediary restricts the set of achievable outputs. Every output reachable through the intermediary is reachable by setting the output directly, so the intermediary can remove acceptable outputs from the agent's reach but never add them. The acceptable outputs that remain reachable through the intermediary (the reachable acceptance set) are a subset of all the acceptable outputs.

This already establishes that control through the intermediary is weakly bounded by direct control. But it says nothing about *information*: about how effectively the agent can *select* from the reachable set. A system might have a large reachable acceptance set, and yet the agent might have no way to find the inputs that reach that set. For this we turn to information theory.

1.4.2 The Data-Processing Principle Applied to Control

The data-processing inequality, a classical result, is the engine of our main argument. We state it for completeness, as it is the single most important technical tool in the entire argument.

The *data-processing inequality*, which this book calls the data-processing principle, states: if a final outcome depends on a source only through an intermediate quantity, then the outcome carries no more information about the source than the intermediate quantity does. For control, it reads: if the final outcome depends on the agent's input only through the intermediary's output, then the outcome carries no more information about the input than the intermediary's output does. In plain language, processing data cannot create information. A quantity computed from an observation cannot tell more about the original source than the observation itself did. Every computation either preserves information or destroys it; none creates it.

The data-processing principle is a fundamental law of information theory, established by Shannon and formalized extensively by Cover and Thomas.[1] It holds for all probability distributions, all Markov chains, all processing steps. It does not depend on the nature of the computation, the cleverness of the algorithm, or the resources available. It is as inviolable, within its domain, as the second law of thermodynamics is within physics.

We now apply it to our controlled boolean system. For a fixed internal state, the control input determines the intermediary's output, and the final bit depends on the control input only through that output. The information that the control input carries about the final bit is therefore no greater than the information it carries about the intermediary's output, and that in turn is no greater than the entropy of the intermediary's output (the most information that output can carry). The chain of the two bounds, from the final bit back to the entropy of the intermediary's output, is the non-amplification argument.

What This Means

The non-amplification argument says that the agent's ability to determine the final outcome is bounded above by two quantities: the information that the agent's input carries about the intermediary's output (which measures how transparent the intermediary is), and the entropy of the intermediary's output itself (which measures how much information the intermediary's output channel can carry). Neither bound can be exceeded. No amount of cleverness in choosing the control input, no amount of computational power, and no architectural innovation can overcome these limits. They are inherent in the structure of information flow. And the bound runs one way. As opacity deepens and self-modification proceeds, both quantities fall, and what the operator retains must be measured against the threshold of requisite control, the level that adequacy requires (Introduction, §8). For the open-ended systems this book concerns, the ceiling falls beneath that floor and stays there: the operator's remaining determination of the outcome is too small to matter.

The non-amplification argument makes no assumptions about the intelligence, benevolence, or design of the intermediary. It does not matter whether the intermediary was designed to be controllable, whether its designers intended it to be transparent, or whether enormous resources were invested in making it interpretable. The only thing that matters is the information-theoretic relationship between input, intermediate state, and output. When the intermediary is not injective, mapping distinct inputs to the same output, information is lost, and that information cannot be recovered by any means. That information lost in a non-injective intermediary cannot be recovered is a law of nature, not a limitation of technique.

Consider the analogy to thermodynamics. The second law tells us that entropy in an isolated system, one that exchanges neither energy nor matter with its surroundings, does not decrease. This does not mean that we cannot build refrigerators or air conditioners; it means that any cooling in one place must be paid for by heating elsewhere. Similarly, the non-ampli-

fication argument does not mean that we cannot build systems that behave well; it means that any controllability at the output must already be present in the information flow from input to output, and no processing step can create controllability that was not already there. The persistent hope in the AI alignment community that clever engineering can maintain control over systems of arbitrary complexity is analogous to the hope that clever engineering can decrease the total entropy of a closed system. Both hopes run afoul of fundamental mathematical constraints.

> ***Remark on the Scope of the Bound.*** The non-amplification argument gives a *within-chain* bound: it states that within the system, each processing step can only lose information. The natural *across-chain* question ("does the agent have less control with the intermediary than without it?") is answered by the Injectivity Characterization: if the intermediary is non-injective, there exists a target for which the mediated chain carries strictly less information from the agent's input to the final outcome than the passthrough chain would. No non-injective intermediary can match direct control for all targets simultaneously. The within-chain bound and the across-chain degradation together constitute the full non-amplification principle: information is lost monotonically through the chain, and the only way to avoid this loss is to make the intermediary injective, which, as discussed below, is incompatible with the computation that makes a learning system useful.

This point is particularly important because much of the contemporary AI safety literature proceeds as if controllability were an engineering problem to be solved rather than a physical constraint to be respected. The hashiness framework shows that it is both: there are engineering problems within the constraint (how to minimize unnecessary information loss), but the constraint itself cannot be engineered away. Every alignment technique, every safety protocol, every interpretability tool operates within

the bounds set by the non-amplification argument. Understanding these bounds is a prerequisite for realistic assessment of what alignment research can and cannot achieve.

We now make precise the sense in which the trivial passthrough (the circuit that simply passes input to output unchanged) is the unique maximizer of controllability. The Injectivity Characterization states: if the intermediary maps two distinct inputs to the same output, there is a target for which the mediated chain carries strictly less information from input to outcome than direct control would; only an intermediary that maps distinct inputs to distinct outputs matches direct control for every target. Every such injective intermediary (a permutation of the bits, for example, or an exclusive-or with a known constant) matches the passthrough, so the passthrough is unique in the property it has, injectivity, and not as a particular circuit. The essential property is injectivity, not triviality.

The significance of this characterization for AI safety is immediate. An AI system that is complex enough to be useful (that performs non-trivial computation, that learns, that adapts, that generalizes, that pursues goals in novel environments) cannot be injective as a map from controller inputs to system behaviors. The entire point of such a system is that it processes inputs in complex, non-invertible ways: it generalizes from training data to novel situations (many different training sets could produce the same generalization), it plans over long horizons (many different planning traces could produce the same action), and it represents the world in internal abstractions that are not directly accessible to the controller (many different inputs could produce the same internal representation). Every one of these capabilities is a departure from injectivity, and every departure degrades controllability by the precise amount that the non-amplification argument measures.

This creates what might be called the *capability-controllability tradeoff*: the more capable an AI system is (the more non-trivially it processes its inputs), the less injective its internal computation, and the less control the operator retains over its outputs. A system that merely passes through its inputs unchanged is maximally controllable but useless: it does nothing.

A system that performs deep, creative, autonomous computation is potentially very useful but, by the same token, is performing the kind of non-injective processing that destroys controllability. You cannot have both maximal capability and maximal control. The tradeoff follows from the non-amplification argument, and no design removes it.

1.5 Four Layers of Controllability Degradation

The non-amplification argument establishes a ceiling. We now decompose the *distance* from this ceiling into four independent and composable sources of degradation. Each layer identifies a distinct mechanism by which controllability is lost, and together they provide a structured vocabulary for analyzing where control breaks down in any specific system.

1.5.1 Layer 1: Target Receptivity

Layer 1 concerns the target itself, independent of the intermediary: how much of the output space it accepts (its acceptance density), and the shape of the acceptable set. Even with direct control of the output, some targets are easier to satisfy than others. A target that accepts 50% of possible outputs is vastly easier to hit than one that accepts 0.001%, regardless of how transparent the intermediary is.

But acceptance density is only part of the story. The *geometry* of the acceptance set matters too. A target whose acceptance set is a single connected region in output space is easier to reach than a target whose acceptance set is scattered across many isolated points, even if the two acceptance sets have the same density. The relevant tools here are *influence* and *noise stability* from the analysis of boolean functions.

Influence measures how sensitive the target function is to individual bit changes. Noise stability measures the chance that the target's verdict stays the same when each bit of the output is flipped independently with a small probability. A function with high total influence changes its output frequently as individual bits are flipped: it has a "rough" boundary

between its acceptance and rejection regions. A function with low total influence has a "smooth" boundary. The fundamental relationship, established by O'Donnell and rooted in the Fourier analysis of boolean functions, is that functions with high total influence (such as parity) have low noise stability, and vice versa.[2]

For the AI alignment application, Layer 1 captures the difficulty of the alignment target itself. "Be helpful" is a relatively easy target (many possible behaviors satisfy it). "Be helpful, harmless, and truthful simultaneously across all possible inputs, including adversarial ones" is a much harder target: the target has lower acceptance density and lower noise stability, because even small perturbations to a barely-acceptable behavior pattern may push the behavior pattern into the unacceptable region.

1.5.2 Layer 2: Structural Opacity

Layer 2 concerns the circuit: how hard it is, from the circuit's structure, to find an input that produces a desired output. With the internal state fixed and known, the intermediary becomes a deterministic map from control inputs to outputs. Its opacity has two aspects: it may restrict the reachable set (failing to be surjective, so that some acceptable outputs are simply unreachable), and it may confound distinct inputs (failing to be injective, so that different inputs produce the same output and the agent cannot distinguish between them). Structural opacity measures how much of the direct control density the intermediary takes away: it is zero when the intermediary preserves full control and one when the intermediary destroys all control.

The computational aspect of this layer connects to circuit complexity theory. Even when many successful control inputs exist, finding one may be computationally intractable. The problem of finding a control input whose output the confirming function accepts is an instance of Circuit-SAT (the satisfiability problem for boolean circuits), which is NP-complete in general (it belongs to the class of hardest problems whose solu-

tions can be checked quickly).[3] This means that for a sufficiently complex intermediary, even if solutions exist in abundance, a computationally bounded agent may be unable to locate them.

For the AI alignment application, Layer 2 is the transparency problem. A neural network with billions of parameters is, at the gate level, an enormously complex boolean circuit. Its internal computations are opaque not merely in the practical sense (we have not yet built good enough interpretability tools) but in the complexity-theoretic sense (the problem of determining what input will produce a given output is, in general, computationally intractable even with complete knowledge of the circuit).

1.5.3 Layer 3: State Obscurity

When the internal state is unknown, the agent must find an input that works across all possible states, or at least across the distribution of likely states. This leads to the core measure of hashiness. *Hashiness* is the probability that the agent's best single choice of input fails to satisfy the target, when the hidden state is drawn uniformly at random.

Hashiness captures the fundamental difficulty introduced by hidden state. When the internal state is unknown, the agent's best strategy may work for some internal configurations of the system but fail for others. The hashiness measures the probability of failure even under the agent's best possible strategy. A system with hashiness near 0 is controllable despite its hidden state; a system with hashiness near 1 is uncontrollable: its hidden state dominates the outcome, and the agent's input is effectively irrelevant.

The name "hashiness" reflects the extreme case. A cryptographic hash function has hashiness essentially equal to 1 (for an agent who cannot invert the function, or for the function keyed by a value the agent does not know): no input reliably produces an output satisfying any non-trivial criterion, because the relationship between input and output is (by design) indistinguishable from random. Every real system has some hashiness

between 0 and 1. The question is where on this scale any given system falls, and the answer depends on the system's internal complexity, the size of its hidden state, and the difficulty of the target.

For the AI safety application, Layer 3 captures the most pressing practical concern. The internal state of a learning system changes continuously as the system acquires experience, updates its parameters, and refines its internal representations. The controller cannot observe this state directly: it is precisely the system's inner cognitive life that is hidden from external inspection. A control strategy that works for one configuration of the system's weights may fail for a different configuration, and the controller has no general way to know which configuration the system is in. Worse, because the system is learning, its configuration at any given moment is the product of its entire training history, including interactions that the controller may not have observed or may not remember. The space of possible internal states grows exponentially with the number of parameters, and modern systems have billions to trillions of parameters, placing the state space far beyond any possibility of enumeration or analysis.

The hidden state of a learning system, which the controller cannot observe and which changes as the system learns, is the basis for a concern that is frequently expressed informally in the AI safety literature. The concern is that an AI system might undergo a "treacherous turn", behaving as expected under one internal configuration but switching to unexpected behavior when the system's internal state changes. The hashiness framework shows that this is a mathematical consequence of hidden-state dynamics: when the hidden state is large and unknown, the set of control inputs that work across all states is generically empty.

1.5.4 Layer 4: Noise Degradation

Real systems are subject to noise: input signals are corrupted, internal state fluctuates, output channels are noisy. Noise at the input, the intermediary's computation, noise at the output and the evaluation of the target are each a step in the chain from controller to outcome, and each step can

only keep or lose the information that the input carries about the outcome. This cascaded degradation argument is the full statement of the non-amplification principle in its most general form. Every step in the chain from controller to outcome can only degrade control. Noise at the input degrades control. The intermediary's computation degrades control. Noise at the output degrades control. The target evaluation degrades control. These degradations are independent. There is no step at which control can be recovered, no clever trick that can compensate for information lost at an earlier stage.

The Ceiling Below the Floor proves this cascaded degradation as Theorem 3.15 (Cascaded degradation).

In the AI safety context, noise includes a wide range of phenomena. Ambiguity in natural-language commands introduces noise at the input interface. Imprecision in reward signals introduces noise in the training process. Distributional shift between the training environment and the deployment environment introduces noise in the transformation itself, since the system's learned mapping may not generalize faithfully. And the inherent stochasticity of real-world physical interactions introduces noise at the output interface. Each of these noise sources independently degrades controllability, and the cascaded degradation argument guarantees that their effects compound rather than cancel. An alignment strategy that accounts for one source of noise but ignores the others provides less safety than its designers believe, because the unaccounted noise sources continue to erode the information channel between controller and outcome.

Noise degradation is a structural feature that compounds with opacity. In a transparent system, noise at the input can be corrected by the agent (who knows the system's response to each input and can adjust accordingly). In an opaque system, noise at the input produces unpredictable changes to the output, and the agent has no way to compensate. The interaction between noise and opacity is multiplicative: a system that is both opaque and noisy loses control faster than the sum of the individual effects would suggest.

1.6 Asymptotic Behavior of Hashiness

We now verify the intuitions stated in the introduction regarding how hashiness scales with system parameters. These results establish that hashiness is not merely a theoretical concern: it grows systematically with system complexity, and it grows in exactly the directions that matter for the AI alignment problem.

Three results describe this growth. First, when the system's output has more bits than its control input, the intermediary cannot reach every output: at most a fraction of the outputs equal to one over two raised to the number of extra output bits is reachable, so the reachable fraction of the output space shrinks exponentially as the output space outgrows the input space. Second, for a random circuit of sufficient depth, different internal states give nearly disjoint sets of outputs, so the set of inputs that work for every internal state shrinks rapidly as state bits are added, and hashiness approaches 1. Third, for a random circuit, the chance that the circuit is injective falls as its number of gates and its depth grow: complex circuits tend to map many inputs to one output, and such mappings lose information.

The first result has a direct interpretation for AI systems. When the system's output space (its behavioral repertoire, the space of possible actions it can take) is vastly larger than the input space (the set of prompts, reward signals, or constraints the operator can specify), then the operator can access only an exponentially small fraction of the system's possible behaviors. An output space vastly larger than the input space is the regime in which all practically relevant AI systems operate. The space of possible neural network behaviors is astronomically larger than the space of possible training signals.

The second result says that the more hidden state a system has, the harder it is to control. This is because different internal configurations produce different control regions in input space, and the agent must find

an input that works for *all* configurations. As the number of hidden configurations grows exponentially (each additional state bit doubles the number of possible configurations), the overlap between their control regions shrinks correspondingly.

The third result connects to a deep intuition about complex systems. Simple systems are transparent: you can often invert them, determining from the output what the input must have been. Complex systems are opaque: many different inputs produce the same output, and determining which input was actually applied becomes progressively harder as complexity grows. The non-amplification argument tells us that this opacity translates directly into loss of control.

The three results together paint a coherent and troubling picture for the AI safety enterprise. The first result says that as the system's behavioral repertoire expands beyond the controller's input space, the controller's reach shrinks exponentially. The second result says that as the system's hidden state grows, the controller's robustness shrinks exponentially. The third result says that as the system's computational complexity grows, the information available for control shrinks exponentially. All three trends point in the same direction: more capable systems are less controllable. And all three trends are exponential in the relevant parameters, meaning that small increases in system capability can produce large decreases in controllability.

For the AI safety application, this exponential scaling counts against control in every parameter at once. The systems of greatest concern (those approaching or exceeding human-level capability) are precisely the systems with the largest behavioral repertoires (the first result: a system capable of open-ended reasoning and action has an astronomical output space), the largest and most dynamic hidden states (the second result: billions of parameters updated continuously through learning), and the deepest computational processing (the third result: transformer architectures with hundreds of layers performing attention operations whose aggregate effect is resistant to analysis). Every parameter of concern is at or near its worst case. The hashiness of such systems is not merely high; the

scaling analysis suggests it approaches the theoretical maximum of 1, the same hashiness as a cryptographic hash function, but achieved through the natural complexity of cognitive computation rather than through deliberate design.

Remark on Three Tiers of Support. The arguments above have three distinct logical tiers, each with a different epistemic status.

Tier 1: Universal results of information theory. The non-amplification argument and the Injectivity Characterization hold for every circuit, every hidden state and every target.

Tier 2: Worst-case and undecidability results from complexity and computability theory. The NP-completeness of Circuit-SAT applies to circuits in general, not to random circuits.[3] Rice's theorem (Section 1.8) applies to all Turing-complete intermediaries (those able to carry out any computation a general-purpose computer can).[4] These are worst-case results, not average-case.

Tier 3: Average-case and generic results from random circuit analysis. The hashiness measure uses the uniform distribution over the internal state. The generic results of Section 1.6 reason about random circuits. The quantitative control-degradation results also belong here: those of Chapter 2 (the hashiness bound from the abstraction differential, Section 2.3.2; the result that a single opaque level suffices, Section 2.5.1; and the absolute ceiling on control, Section 2.10), and their dynamical restatement in the quantitative bounds of Chapter 4 (Section 4.8.1). Each holds under the target-confounding condition (the hidden bits independently gate the target), which is satisfied for adversarial targets and, under a uniform-random circuit model, for *generic* ones, but not for every target. These results characterize what *typically* happens as system parameters grow; they do not hold for every structure, and particular benign targets can evade them.

The three tiers carry different loads. Tier 1 is universal: no amplification, and a ceiling of one bit. Tier 2 is worst-case: Circuit-SAT hardness and Rice's theorem are statements about *general* or *worst-case* instances, and do not transfer to a *particular* system without a further argument that the specific instance is hard. The *quantitative* control-impossibility claims, the ones asserting that controllability is driven toward zero, live in Tier 3 and are therefore *generic*, not universal. Read at the standard fixed in the Introduction, this genericity concedes nothing: the structures outside the

generic case are of vanishing measure against the systems that practice actually builds (Introduction, §7), and the runtime-control results accordingly establish that achievable control tightens with capability and falls beneath the threshold of requisite control, permanently, because the barrier enforcing the fall is of the invariant class. The closure by disposition (Chapters 9 and 10, Sections 8.4 to 8.6, 6.6 and 6.9), whose premises are thermodynamic and ecological rather than information-theoretic, reaches the same conclusion without depending on the control results at all. The application of Tier 3 results to specific (non-random) systems such as trained neural networks requires an additional argument: that a learning system's internal state, being the product of a compression process, is epistemically opaque to the controller and therefore indistinguishable from random from the controller's perspective (Section 1.9.4). This bridging argument, grounded in Kolmogorov complexity theory (the theory of the length of the shortest description of a string), justifies the use of the uniform distribution in the hashiness definition as an accurate model of the controller's epistemic situation, not merely as a mathematical convenience.

Hashiness as the Mixing of Good and Bad Outcomes

Seen from the side of the state space (the set of all states a system can be in), hashiness is the degree to which acceptable and unacceptable outcomes are intermixed in the state space of a system, and it sets a limit on the capacity to control or align any sufficiently complex process. Section 1.5.3 defines hashiness for a single boolean intermediary, as the probability that the agent's best single choice of input fails when the hidden state is unknown. In this section I measure hashiness on the state space itself and connect that measure to satisfiability. I ground the measure physically in the compatibility of machine-produced matter with biological life. I then follow the chain of consequence that runs from hashiness to unalignability.

The Converse of Requisite Variety

Control theory, in its most fundamental formulation, is the reduction of variety. A system is subject to perturbations that introduce variety into its internal state; a regulator compensates for those perturbations so that the system's output remains close to the desired goal state. If the regulator succeeds, the output exhibits low variety despite high variety in the input. The reduction of variety by a regulator is the substance of Ashby's Law of Requisite Variety (Section 1.9.1).

The law has a converse that is often overlooked. If the regulator cannot match the variety of the perturbations (if it cannot distinguish between disturbances well enough to select the correct counteraction), then the system is not merely imperfectly regulated. It is *unregulable*. The question, then, is whether there exist systems for which no regulator of manageable complexity can achieve adequate control.

Such systems are not exotic edge cases. They arise naturally whenever the intermixing of good and bad outcomes in the system's state space exceeds a certain threshold of complexity: a threshold that can be characterized as a property of the state space itself, independent of any particular control strategy. That property is hashiness.

The Hash Function Model

To formalize the difficulty, consider a hash function that takes some number of input bits and produces the same number of output bits. Designate an output as "good" if its first two bits are both zero: a condition satisfied by roughly one quarter of all possible outputs.

Now consider the input space. If the hash function is trivial (a pass-through, for instance), then the good inputs are a single contiguous region: all sequences whose first two bits are zero, regardless of the remaining bits. The input space of a pass-through hash is a maximally organized space. The good region is large, connected, and easy to target.

If the hash function is a good cryptographic hash, the situation inverts. The good inputs (those that produce outputs with the first two bits zero) are scattered throughout the input space like dust. There are still the same number of them, still roughly a quarter of the total, but they are no longer adjacent. They form no large connected regions. The good is dispersed among the bad with no exploitable structure. Hashiness is the condition in which acceptable outcomes are intermixed with unacceptable outcomes in the state space. In that condition, no classifier simpler than the system itself can reliably tell the acceptable outcomes from the unacceptable outcomes.

The hash functions used in cryptography are not continuous modulations; they are discrete. A set of bits is permuted through a series of AND and OR gates, perhaps recursively, in a way that keeps increasing the dispersion (the spreading of each input bit's influence across more and more of the output bits). Unwrapping the loops gives just a very long series of gates. With a bad hash function, a bit of noise jiggles the output a little. With a good one, the same noise is spread over the entire range of outputs, and spread in a way that is sharply discontinuous. If the gates are not known precisely, or the input is not known precisely, and even a little randomness changes one bit, there is no knowing at all what the output bits will be. So the discontinuity comes either from the structure of the gates or from what is unknown about the input.

The two sources of discontinuity, the structure of the gates and what is unknown about the input, give the key variables: whether the process is discrete or continuous, how much is known of the interior of the process, and how much is known of the input signal. At one end there is predictability, and at the other there is not. Characterizing the middle, the place where predictability gives way to unpredictability, tells not only what kind of characterization applies to the transform, the hash function itself, but also which ranges of input signals produce a definite output and which produce an indefinite one.

Adjacency in the High-Dimensional Space

A measure of hashiness must capture not merely the ratio of good to bad outcomes (the *volume* of the acceptable subspace) but the degree of their intermixing. Two spaces with identical volumes of acceptable outcomes can differ enormously in hashiness, if one concentrates its good outcomes in a few large regions and the other disperses them as a fine-grained dust.

The relevant quantity is *adjacency*. In a bit space with one dimension for each bit, two states are adjacent if they differ in exactly one bit. The adjacency of the good subspace is the number of adjacent pairs in which one member is good and the other is bad: the total number of boundary crossings.

The adjacency of the good subspace is a generalization of surface area. A ball of good outcomes in a sea of bad has a surface area proportional to the ball's boundary. A *Julia set* (a fractal set from mathematics whose boundary is infinitely detailed), taken as a space with good outcomes fractally intermixed with bad, has an enormously larger surface area relative to the same volume.

The hashiness measure on the state space, then, is the ratio of observed adjacency to the maximum possible adjacency for a space of that volume. If the ratio is low, the good outcomes are clumped into large connected regions, and the system is controllable, because a classifier need only identify which region the system is in. If the ratio is high, approaching the *checkerboard limit* (the most finely mixed arrangement possible, in which every neighbor of a good state is bad), and if no classifier simpler than the system marks the pattern (a perfectly regular checkerboard is marked by one small circuit, the parity function, so regularity of that kind is excluded), the system is uncontrollable, because every adjacent state may cross the boundary between good and bad, and no shortcut exists for determining which side of the boundary a given state is on.

The Satisfiability Connection

The formalization connects directly to the satisfiability problem (the problem of finding inputs that make a logical circuit or formula come out true). If we take the hash function's output and ask which inputs produce a good output, we have converted the question into a Boolean satisfiability instance: given a circuit (the hash function) and a target condition (two output bits are zero), find inputs that satisfy the condition.

The complexity of this satisfiability instance is determined by the structure of the circuit. For trivial circuits, the satisfiability problem is solvable in polynomial time. For circuits of sufficient complexity, which is to say for hash functions with high hashiness, the problem is *NP-hard* (at least as hard as the hardest problems whose solutions can be checked quickly; Circuit-SAT, Section 1.5.2, is the general case). The hashiness measure can therefore be understood as a normalized complexity measure: a value between zero and one that characterizes where a given system falls on the spectrum from trivially solvable to computationally intractable.

Rice's theorem, together with Chaitin's result that most algorithms are already in their maximally simplified form (Section 8.5.2), tells us that for the full space of possible hash functions, the vast majority cannot be simplified. Any system that attempts to predict the output of an arbitrary hash function without actually computing it will fail for most instances. The learning system (the classifier, the large language model, the alignment monitor) cannot, in the general case, be simpler than the thing it is trying to predict. That a learning system cannot, in the general case, be simpler than the thing the learning system predicts is the formal content of the claim that hashiness is an obstacle to learnability.

Take a *bit field* (a fixed-length sequence of bits), run a hash function on it to get an output bit field, run the hash function on that output, and so on: a *recursive hash*. Every bit of each output depends on every bit of the input. Is there any simpler way to calculate any of the bits at any point in

time? Generally not, because if there were, cryptography would fail. And there has been every motivation for the smartest people in the world to find shortcuts for this kind of algorithm: if they found one, they could reverse engineer Bitcoin and become instant billionaires. From a financial perspective, the brightest people of the entire world have looked for shortcuts to simplify the hashing function.

So wherever every bit of the input defines every bit of the output, with recursion, the energy needed to calculate a future state is not less than the energy needed to calculate each hash in sequence. There is no way to predict the bits three generations ahead faster than doing the hashing. That gives an algorithm that is, at least empirically, not simplifiable.

This result is then reused. If the factors that define the minimum level of control necessary have dynamics like that hashing function, then there are things in the real universe that we cannot model better than by just doing the thing itself, unless we accept error bars larger than we would like, and larger than the minimum we need.

The Control Inequality

Consider a control chain of three elements: the regulator, the system being regulated, and the world that receives the system's output (the three-node ring of Section 2.4). The regulator observes the world's response to the system, and attempts to control the system's behavior so that the world produces acceptable outcomes.

For the regulator to succeed, it must be able to predict what the world will do with a given output of the system. But this is the same problem that the system faces in trying to achieve good outcomes from the world in the first place. If the world is hashy (if the mixing of good and bad outcomes in the world's response space exceeds the classifier complexity boundary), then the system cannot reliably target good outcomes in the world.

The critical observation is that the regulator cannot do better than the system. The regulator's task is to classify the system's outputs as likely to produce good or bad outcomes in the world. But this classification requires the regulator to predict the world: the same world that the system is already failing to predict. The regulator must also do this through one additional layer of indirection: it sees the system's output, not the world's input directly.

This creates an inequality, the *control inequality*. The hashiness of the space as seen by the regulator is always at least as great as the hashiness of the space as seen by the system. Adding a regulator does not reduce the hashiness of the underlying system. It can only add its own complexity on top of the existing complexity. The control chain cannot outperform the direct computation. If the hashiness of the mapping from the system to the world exceeds the capacity of any feasible classifier, then adding a regulator on top of the system does not make the overall system controllable. The regulator is, at best, no better at predicting the world than the system is, and may be worse, because the regulator must also model the system.

Resistance to Noise: Dusty and Clumpy Good Messages

In the rest of this section, the regulator is called the controller, and the system it regulates is called the machine. Treat the flow from the machine to the world as a channel of messages of a fixed number of bits (say one hundred, though it could be a billion). Color every possible message white if it leads to a good outcome in the world, and black if it leads to a bad one. Suppose, generously, that a quarter of all messages are white. The question is how precise the machine must be to land on a white message when the channel adds some noise, for example when a few bits are flipped at random.

The answer depends on how the white is mixed in with the black. Arrange all possible messages in a *Gray code* sequence, in which each message differs from the one before and the one after by a single bit. If the white messages are maximally *clumpy*, any three bits of a good message

can be changed and it is still a good message: aiming roughly is enough, and a few errors of precision do no harm. If the white is maximally *dusty*, the good messages are as far from one another as the space allows. Take the case where one message in sixteen is white and the white is maximally dusty: then four bits have to change to get from one white message to the next, and changing any three bits of a white message is guaranteed to turn it black. In that case any amount of noise is likely to move a message away from good, and a single wrong bit can ruin it.

The clumpy or dusty arrangement of the white messages gives a way to describe how well a flow of messages resists noise. The same description can then be applied to the flow of messages from the controller to the machine, so that the boundary between the machine and the world can be compared with the boundary between the controller and the machine.

Learning the World by Probing It

Rather than have the machine understand the world directly, ask whether the machine can learn the coloring of the world's inputs from outside, the way any learner must: by sending messages to the world and getting feedback. For this question, simplify the model. The controller does nothing; it is a straight pass-through. The world takes an input and maps it to a single wire that comes back through the controller to the machine, so the machine has one useful input. Then add two constraints. First, the proportion of white is well under half: a quarter, or, more interestingly, one part in sixteen. Second, the white is maximally dusty among the black, in the Gray code sense above. Finally, the channel from the machine to the world is imperfect, so three bits of every message are flipped at random, and the machine does not know which three.

Under these conditions the machine cannot learn the coloring of the world.

And this carries over to the controller. The controller is trying to predict the future actions of the machine and the world together, so as to know what bits to send to the machine to bring about a good outcome.

The controller cannot learn the combination of machine and world any better than the machine can learn the world. If the machine cannot learn the world, then the controller cannot learn the combination of machine and world, and so the controller cannot predict anything it would need to predict in order to control the machine. If you cannot learn a system, you cannot know it; if you cannot know it, you cannot predict it; if you cannot predict it, you cannot control it.

Noise Has a Physical Floor

Two parameters can be varied separately: the noise on the channel, and the dustiness of the coloring. Reduce the noise and keep the dustiness high; or keep the dustiness low and make the noise high; and see how much each one changes the dynamics. Varying the noise tells us how much the dustiness matters. But the noise cannot be brought below a certain threshold, because this model will eventually be applied to the physical universe, where bits flip simply because things happen in the universe, and where the Heisenberg uncertainty principle eventually applies (Section 1.9.3 treats noise as fundamental).

Some portions of the message space may be very dusty and others not. Then there are regions in which controllability applies and regions in which it does not, and the boundary between them can be studied as a function of the noise. Or one can set the noise aside, combine the dusty and the non-dusty regions into a single hashiness value, and say that the hashiness of the boundaries of the machine stands in a relation defined by the hashiness of the inside of the machine. The hashiness of the output of the machine is at least the hashiness of its input, and probably never lower, if the machine has any hashiness at all.

Physical Grounding

The abstraction acquires weight when grounded in a concrete physical domain. Consider the relationship between machine-produced matter and biological life.

The periodic table contains roughly one hundred elements. Biological life uses approximately ten of them in significant quantities (the major part of the wider alphabet of some 20 to 29 elements, trace elements included, that Section 9.5.1 counts). A crude classifier (discard anything containing elements not in the biological set of ten) works almost perfectly at the level of elements. If the output of a machine process contains lead, or cadmium, or mercury, the classifier rejects it, and the biosphere is protected.

But as soon as we move from elements to molecules, the classifier fails. Carbon and oxygen are both in the biological set of ten. Carbon monoxide is lethal. Nitrogen and oxygen are both biologically common. Nitrogen monoxide is toxic. Carbon and nitrogen combined, as hydrogen cyanide, are deadly. The acceptable molecular combinations form a vanishingly small fraction of the combinatoric space of arrangements of even those ten safe elements.

The biological requirement is not even monotonic (changing in one direction only). Iodine is not among the ten primary elements, yet a trace amount is necessary. Without it, the thyroid fails, a consequence that takes years to show. Selenium is similarly essential in trace quantities. The classifier that rejects everything outside the set of ten is correct 99.99 percent of the time, but the remaining fraction of a percent includes conditions that are cumulatively fatal over a decade.

The relevant physical analogy is not volume but surface area. A block of lead in a cornfield has a large volume of toxic material but a small surface area of contact with biological processes. The same mass of lead ground to a fine powder and dispersed across the field has enormously greater surface area, and therefore enormously greater biological impact. Hashiness is the informational analogue of this surface-area-to-volume ratio: it measures not how much toxic material exists, but how finely it is intermixed with the nontoxic, how many boundary crossings there are between acceptable and unacceptable.

As we ascend levels of abstraction, from elements to molecules, from molecules to cells, from cells to organisms, the combinatoric space grows and the fraction of acceptable configurations shrinks further. At each level, the mixing of good and bad outcomes becomes more complex, and the hashiness of the space increases. If we cannot formalize hashiness even for the simplest case of two elements combined, we cannot formalize it for any of the levels above.

The Chain of Consequence

The argument now assembles itself into a chain of four links.

If a system's state space has sufficiently high hashiness (if the intermixing of acceptable and unacceptable outcomes exceeds the capacity of any learnable classifier), then the system is *unlearnable*: no amount of sampling will allow a learning system to distinguish good from bad outcomes with reliability above random chance, adjusted for the base rate of good outcomes.

If unlearnable, then *unpredictable*: the system's future behavior cannot be anticipated except by running the system itself.

If unpredictable, then *uncontrollable*: no feedback mechanism can compensate for disturbances faster than those disturbances occur, because the regulator cannot detect or classify them in advance.

If uncontrollable, then *unalignable*: the system's outcomes cannot be constrained to any desired subset of the state space, because the boundary between desired and undesired is too complex for any external mechanism to enforce.

The force of the argument lies in this: the chain describes a structural property of the state space itself, and not a contingent engineering limitation. No increase in computing power, no improvement in learning algorithms and no additional layer of oversight changes the fundamental inequality: the classifier cannot be simpler than the hash.

Low hashiness permits control: the good regions are large enough to identify and target. High hashiness defeats control: the boundary between good and bad is too complex for any feasible classifier to learn or enforce. The physical universe is, at the molecular level and above, a system of high hashiness. The fraction of configurations compatible with biological life is small to begin with, and grows vanishingly smaller as the level of abstraction rises. The intermixing of those configurations with toxic or lethal ones is fine-grained, and structurally resistant to classification.

Any system that produces novel matter or patterns, whether it is a machine, a technology or an algorithm, and releases them into the biosphere confronts this hashiness. If the system's output space is complex enough that its hashiness exceeds the capacity of any classifier we can build to monitor it, then no alignment strategy based on external oversight can guarantee safety. The regulator cannot outperform the hash.

1.7 An Intelligent Intermediary: Luck Against Control

A natural question arises at this point, and answering it directly clears up a common confusion about the nature of control.

A reader may ask: could the intermediary be "intelligent" enough to analyze the target and always produce an output that satisfies it, regardless of the control input? The answer is yes, in principle. An intermediary could ignore the control input entirely and always produce one fixed output that the target accepts. Then the final bit is always 1, and no control through the input is needed.

In this case, the mediated control density equals one: every input "succeeds". Meanwhile, the direct control density may be very small. So if one were measuring control by density ratios, the intermediary would appear to have "amplified" control enormously. The apparent amplification by a constant-output intermediary is precisely why control must be measured by mutual information rather than by success density. Density asks: "how many inputs lead to acceptance?" Mutual information asks: "does the

agent's choice of input *determine* whether acceptance occurs?" In the constant-output case, the answer to the first question is "all of them" and the answer to the second is "not at all". The non-amplification argument is stated in terms of the second question, which is the right question for the control problem.

> **The Luck-Control Distinction**
>
> A system that always produces the desired output regardless of the operator's input is not a controlled system; it is a lucky one. Control means that the operator's input *determines* the output. If the output would be the same regardless of what input the operator provides, the operator has zero control, even if the output happens to be desirable. This distinction is critical for AI alignment: an AI that happens to behave well is not the same as an AI that is *controlled* to behave well. The former can change its behavior at any time for reasons the operator cannot influence; the latter cannot.

This clarifies what controllability means in our framework: it is specifically the degree to which the agent's choice of input determines the outcome. A system that always yields the accepted outcome regardless of the input has zero controllability and zero hashiness, but also zero *relevance* of the control interface. Note that such a system is not static: the behavior of the system may still depend on the state vector, which includes environmental inputs, internal parameter changes, and all other influences outside the agent's control. A system that ignores the agent's input while remaining responsive to its own evolving internal state and environmental conditions is, from the agent's perspective, an autonomous system, not a controlled one. The interesting regime is where the final outcome depends on the control input, and the question is how much of that dependence survives the intermediary.

This distinction between luck and control is far from academic in the AI safety context. When proponents of AI alignment say they want to "control" a superintelligent system, they sometimes mean something

closer to "we want the system to produce good outcomes" rather than "we want our inputs to determine the system's outputs". Wanting the system to produce good outcomes is a hope about the system's character. Wanting our inputs to determine the system's outputs is a claim about the controller's power. The hashiness framework addresses the second claim exclusively. It asks: given that the system does what it does, how much of what it does is determined by what we tell it to do? If the answer is "very little", then we are not controlling the system; we are merely in the presence of a system that happens to be doing what we want. The distinction between control and luck matters because a system that happens to be doing what we want today, for internal reasons that we cannot observe or influence, may stop doing what we want tomorrow, for reasons that are equally invisible to us. If the system stops doing what we want, we will have no mechanism to prevent or reverse the change.

The statement is: *a system that always satisfies the safety criterion regardless of controller input is not a controlled system; it is an autonomous system whose behavior happens to be acceptable.* Control requires that the controller's inputs have causal influence on the outcome. The non-amplification argument bounds this causal influence. High hashiness means low causal influence, and the controller is not meaningfully in charge, even if outcomes are currently satisfactory. The difference between outcomes that are merely satisfactory and outcomes that the controller controls is, in essence, the difference between a benevolent dictator and a democratic government. Both a benevolent dictator and a democratic government may produce good outcomes, but only in a democratic government does the citizen's input actually determine the outcome.

1.8 Turing-Complete Intermediaries

The analysis thus far has treated the intermediary as a fixed boolean circuit of bounded complexity. Treating the intermediary as a fixed boolean circuit is sufficient for many applications. But the systems of greatest concern for AI safety are systems capable of universal computation, and not

merely complex circuits. When the intermediary is sufficiently complex to simulate a universal Turing machine (with the state bits encoding both program and data), the situation becomes qualitatively different. The agent cannot, in general, predict the intermediary's behavior even with complete knowledge of its gate-level structure, because doing so would require solving the halting problem (deciding whether a program ever finishes) for the encoded program.

Rice's theorem tells us that for any non-trivial property of computable functions, the set of programs whose function has that property is undecidable. Now take the property to be "for every input, the output satisfies the target", which is to say, the property of being safe. Then the set of programs that are verifiably safe is undecidable. The undecidability of the set of verifiably safe programs is a statement about all possible proof techniques and all possible verification methods, and not only about current techniques and methods. No matter how sophisticated our tools become, there will always exist programs whose safety status cannot be determined.

In the finite-circuit model, undecidability manifests as intractability: the circuit may be finite, but if it faithfully simulates a Turing machine for polynomially many steps, the control problem inherits the computational hardness of the simulated computation. Hashiness in this regime is effectively 1. The intermediary is so computationally deep that the controller cannot predict what any input will produce without essentially running the entire computation, which may take longer than the controller has available, or may require more resources than the controller possesses.

Remark on Finite Circuits and Rice's Theorem. Rice's theorem applies to the set of all programs for a universal Turing machine, which is countably infinite. In our framework, the intermediary is a finite boolean circuit, its internal state has a finite upper bound in bits, and the set of all possible "programs" encodable in that state is therefore finite. For any fixed finite program space, one can in principle enumerate all programs and check each one; Rice's theorem does not apply to a single fixed finite set of programs. The undecidability result applies strictly only when we consider families of circuits indexed by an unbounded number of state bits, where the program space grows without limit. For any specific instance with a fixed circuit and a fixed state size, the control problem is decidable (though potentially intractable). The practical force of the argument rests on the intractability claim: for circuits of the size and depth found in real AI systems, the finite enumeration required is astronomically beyond any feasible computation, and the behavior of such circuits inherits the computational hardness of the Turing-complete computations they simulate for polynomially many steps. Chapter 5 takes up the same limits on prediction for causal processes in general: the halting problem and Rice's theorem (Section 5.3.3), and the low threshold at which a process becomes able to carry out any computation (Section 5.3.4).

Rice's theorem connects the hashiness framework to the broader computational impossibility argument developed in Chapter 3. Rice's theorem and the non-amplification argument are independent results from different branches of mathematics (computability theory and information theory, respectively), but they converge on the same conclusion: complex computational processes resist external control by necessity, not merely by practical difficulty. This convergence from independent starting points is one of the strongest features of the overall impossibility argument.

The connection extends to the dynamic case, which is particularly relevant for AI systems. When the internal state changes over time (as in a learning system that updates its parameters based on previous inputs), the effective circuit is non-stationary, and the agent faces a moving target:

the control region in input space shifts with each invocation. Predictive modeling of such a system requires modeling the learning dynamics themselves, which may be chaotic or computationally irreducible in Wolfram's sense.[5] Computational irreducibility is the principle that, for many systems, there is no shortcut for predicting their future state: the only way to know what the system will do is to run it and observe. When this principle applies, predictive control is impossible in the strongest sense: no model shorter than the system itself can capture its behavior, however good our models become. This temporal dimension of the problem is treated in detail in Chapter 4.

A natural question arises: does a modern neural network actually achieve Turing-completeness? A fixed neural network with fixed weights is a finite function and is not, strictly speaking, a universal Turing machine. However, a neural network equipped with external memory, or one that is repeatedly invoked with its own outputs fed back as inputs (an autoregressive model), can simulate arbitrary computation for arbitrary numbers of steps. Transformer architectures with sufficient depth and attention heads have been shown to simulate Turing machines, and the ongoing trend toward larger models with more complex reasoning capabilities pushes AI systems further into the regime where the Turing-complete analysis applies. The practical implication is that the undecidability results do not require literal universality; they apply whenever the system's computational capacity is sufficient to make its behavior unpredictable by any method simpler than running the system itself.

The QXZ Circuit Model of Hashiness

The hashiness model gives a short, simple and intuitive formalism for showing that AGI cannot be controlled, one that leaves no room for the replies that the analytical results from Rice's theorem and control theory have met. In this section I set out the circuit form of the hashiness model (the QXZ model), together with the proof of the QXZ model. In the proof I show that the control needed on the inputs of each block of a loop of

control is strictly greater than the control needed on the outputs of that block. Each cycle through the loop raises the required control, until success becomes impossible.

Why Rice's Theorem Alone Does Not Carry the Argument

> *Objector:* Why use the concept of hashiness, when an inequality already trivially exists between the decidability of the control system and the more complex AGI, under reasonable assumptions? If X is predictable via Q, then making X := !Q(X) is contradictory. That is literally Rice's theorem, and it is a trivial result: the behavior of X is unpredictable in the general case. If the AGI model X is Turing complete, or "effectively Turing complete" (satisfying the prerequisite assumptions of Rice's theorem), we cannot predict the value of any metric on it in any meaningful way, which means the control module Q cannot control it. Any inequality behind hashiness is therefore a consequence of that, and not the other way around.

Rice's theorem does apply. Any learning algorithm will have some ring structure (repetition that loops back on itself) and some conditionalization (branching on conditions), and is itself inherently sequential. Ring structure, conditionalization and sequentiality are the conditions needed to make something Turing complete, and once those three conditions are achieved, Turing completeness is for sure present. Since a learning algorithm requires those three elements in order to function as a learning algorithm, any implementation of AGI is going to have Turing-complete characteristics, and Rice's theorem is relevant to the limits to predicting whether AGI outputs are safe. (The Remark on Finite Circuits and Rice's Theorem above states the precise form this takes for finite circuits.)

Yet that by itself does not convey the argument, because some AI researchers have decided that Rice's theorem is not relevant. In the simplified case of the halting problem, they say, "You could write some al-

gorithms where you know for sure those algorithms will halt". The researchers' claim is true. But being able to identify some programs that will halt some of the time, for some ranges of inputs, does not give a general way of calculating, on the basis of any fixed algorithm, whether a general program is going to halt under predictable circumstances. The halting problem says that no fixed algorithm gives a general way of calculating whether a program will halt, and Rice's theorem extends that result to all sorts of attributes of programs and to the predictability of programs. The next response is: "Because there are some exceptions, how do we know that AGI is not going to be in one of those spaces of exceptions, so that Rice's theorem is just not relevant?" The answer is that a generalized learning algorithm is a general algorithm: we cannot predict what the algorithm is going to be in the future, so we have to use the general case. The response after that is: "Maybe we cannot tell for sure whether this algorithm is going to have certain attributes; but if it is 99.99% safe, maybe that is good enough, and we can get away with it". At that point one must explain how the assumption that an acceptably high lower bound can be guaranteed on the probability of safety is unsound as well.

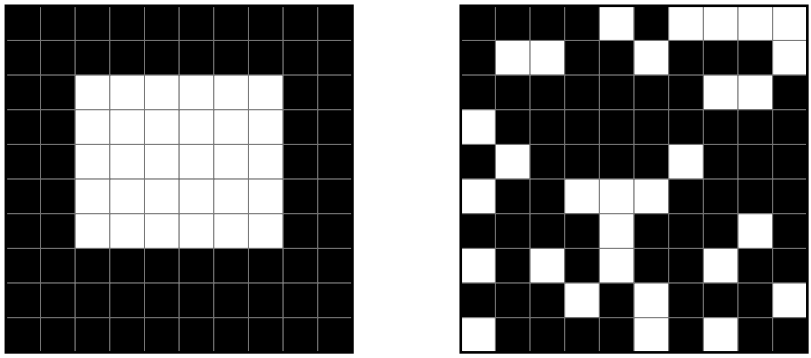
Conversations about analytical results thus tend to get stuck on nitpicks disputing whether the results are relevant to the problem. A dead simple way of using existing theory to model the problem is needed, one that cannot be wiggled out of. As soon as a concise formal model is made, the reply is "that does not apply to the real world; it is too simplified; we are not going to make assumptions X, Y and Z". Showing that the assumptions do empirically hold in the general case requires describing processes observed in the real world in a shared language such as English, and then the argument sounds ambiguous again. The result is a catch-22: a plausible-sounding rationale for ceasing to engage with the argument is offered, and reformulating the argument to remove that rationale positions the person replying to offer a different one (that checking the argument would take too long, that it is not based on theoretical results known to be logically valid, or that the familiar theory cannot be applied to reach a conclusion about the world that the person replying disagrees with).

A textbook on control theory runs to some four or five hundred pages of dense mathematics, and the argument does not need all the formalisms of such a book. It can probably focus on four or five well-known and verified theorems and a few lemmas, and explain how an AGI uncontainability result can be derived even just from those parts. The hashiness model provides a number of unambiguous shortcuts to that end.

What the Hashiness Model Is

The hashiness model is a way of characterizing how precise, or how robust, a particular system would need to be not to cause some catastrophic outcome. It has two components.

The first component is the *dartboard model*. It characterizes the possible outcomes of AGI outputs as a space covered by white squares (denoting non-lethal outcomes) and black squares (denoting lethal outcomes). The distribution of non-lethal and lethal regions is further characterized by two metrics: an *area ratio* (how much area is covered by white squares relative to black squares) and a *mixing coefficient* (how much boundary is shared between black and white squares). Of the two metrics, the mixing coefficient is particularly relevant, as it determines how precisely AGI outputs must be targeted to still map onto non-lethal outcomes. Figure 1.1 shows the same area of white arranged clumpy and arranged dusty.



clumpy *dusty*

Figure 1.1. The dartboard model: the same area of white (non-lethal outcomes), thirty squares of a hundred in each board, arranged clumpy on the left and dusty on the right. The area ratio is the same; the mixing coefficient (how much boundary white and black share) is far higher on the right, so outputs must be aimed far more precisely there.

The second component is how AGI inputs are transformed into outputs. We cannot expect the mapping of inputs to outputs to be a simple bijection (a one-to-one mapping in which every input has exactly one output and every output exactly one input). There might be inputs that map to the same output, or there may be some outputs for which no input exists (outputs that could be defined in a binary sense, but cannot be constructed from inputs). The mapping from inputs to outputs could be anything from complete disconnection to complete and total direct specification, and everything in between. Hashiness is the metric of how far the processing of inputs on one side into outputs on the other side leaves an agent unable to steer the outputs (Section 1.5.3 gives it its formal definition for a single boolean intermediary).

The hashiness model is thus a characterization of the inputs, a characterization of the outputs, and of the degree to which an output characterization allows us to say things about the input characterization. The inequality shows up in the relation between the characterization of the outputs and the characterization of the inputs, because control theory is a

characterization of the same kind. Hashiness refers to a transformation of the inputs into outputs, and to a characterization of the input space and the output space in terms of what we mean by success; and the model looks at how those characterizations are themselves transformed through a three-part system of control, AGI and world, and the feedback loop through that system.

The QXZ Circuit Model

The QXZ model is the three-node ring of Section 2.4 (controller, complex system, world) taken in its strictest clocked form, with the controller here named Q and the world named Z. Figure 2.1 draws that ring, with C in the place of Q and W in the place of Z.

> *Q* is a control algorithm.
>
> *X* is the AGI, plus whatever signal interrupt circuitry is used to inhibit X outputs to Z.
>
> *Z* represents the whole of the physical world, its lawfulness, its state, and so on.

All of the internals of Z, X and Q are represented by pure boolean circuits only. The interface wiring in the flows of information from Z to Q, Q to X, and X to Z consists entirely of wires representing an explicit boolean state, wholly represented at some instant of time.

So as to make things much simpler, all wiring flows are unidirectional only, passing only from Z to Q, Q to X, and X to Z, in explicit sequence, and with a definite absolute clock. *All* of the (finite, definite) signal state that Z will send to Q passes from Z to Q at instant t = 0, and any information that X is to obtain from Z is handed to X by Q (as a pass-through) at instant t = 1. This ensures that Q has unambiguous access to, and control of, *all* of the input state that Z sends indirectly to X. Moreover, at the instant t = 1, Q will have done all of the computation of the future output

state of X based on the inputs it provided to X, and so passed all of whatever control signals are necessary to limit, block or change the output of X to Z at the future instant t = 2.

Hence Q has the maximum possible calculable degree of control of X. Insofar as Q has been given the maximum possible power of opportunity to control X, the limit of control is defined purely by the algorithm (circuit computation) of Q with regard to X, and so also of X with regard to Z. *All* of the flow of bit information from X to Z happens at the instant t = 2, and Z obtains information only from X and not from anything else, apart from Z sending signals to Z itself, which represents the actions of whatever physical processes, other agents, and so on. In this way, all of the Z calculations occur within the interval between t = 2 (when Z gets signal from X) and t = 3 (when Z halts and sends whatever finite signal will be sent to Q). Insofar as the action of computation is strictly completed in each stage, and each stage is strictly sequential, the maximum possible opportunity for assessing the implications of algorithm (circuit) complexity is fully available.

Why the Reduced Case Is Fully General

Any parallel asynchronous version of the system is a generalization of the proof truths discovered in this most simple version. The generalization holds especially because *all* versions of *all* AGI variants, *all* variations of the laws of real-world physics, and *all* the variations of control system possibility can be fully reduced to a fully discrete finite circuit system of the QXZ type. Hence full relevance cannot be avoided; with this fully reduced case, no nitpicks are possible.

The model gets around the Church-Turing thesis aspects by requiring that the full computation of Q, X and Z be fully completed (have taken finite inputs and processed them to finite outputs) in finite time and with finite memory: the circuit of each unit component Q, X and Z is fully and discretely specified. This simplest case is a "representation target" that can "receive" all physically implementable versions of AI, AGI and ASI, as

well as all fully realizable implementations of Q and Z, as if via some sort of language compiler that reduces (finite, though potentially very large) fully specified algorithms to fully specified boolean circuits. This ensures that the potential ability of X to predict eventual outcomes in Z is absolutely maximized, and similarly that the ability of Q to predict X predicting Z is fully maximized. Hence, if structural impossibility is proved here, *all* future AI systems of more general capability inherit the result, insofar as they *must* have *less* predictive control capability (and so less alignment and safety, by implication) than this much more fully reductive system.

The proof is set up in this way: take a simple case, show that every extrapolation of that simple case reduces to it, and show that the simple case has a clear, unambiguous proof. Show that proof, and it is done. The simple case has three parts: the control system (Q), the artificial intelligence it is meant to control (X), and the living world that receives what the intelligence does (Z). Whatever the proportion of white to black on the target, and whatever its coastline (a scattering of tiny white points, or a few white regions that can be aimed at where anything in the cluster is acceptable), the amount of control needed on the bit stream is high. The question is then: what level of control on the bit stream from the control system to the intelligence is needed so that the intelligence's output does not destroy the living world? The controller does not get to change how the living world receives that output. Its distribution of white and black is fixed by the laws of nature and of biology, by the world as it is.

Every proposal of the form "make the living world more robust" reduces to the same case. Whatever is done to improve its robustness is still a signal: either one that flows from the control system through the intelligence to the living world without modification, or one that the control system sends to the intelligence and that is itself a signal to the living world. Even a superintelligence trying to make the world more robust is sending a signal that has to work, starting from the conditions the world is actually in. So the question is again whether that signal lands in the white area, and every such variation is a special case of the signal flow from the intelligence to the living world.

One loop covers every loop. Information flows from Q to X, from X to Z, and from Z back to Q. Every other path can be built from this one loop by pass-throughs (blocks that hand their input on to their output unchanged): Z can route back to Z through itself, or through Q and X unmodified; X can reach Q by passing through Z; and Q can modulate any output of X. So all possible loops of process reduce to this one loop, and the proof is subsumptive of all possible modeling dynamics, because modeling dynamics are always a loop of some sort: all algorithms are feedback loops, and all causation is a feedback loop. By describing the dynamics of loop causation directly, the proof subsumes every quibble in this space: each "what if" can be shown to be a special case of the one loop, whose dynamics are fully described.

The hashiness model is like Shannon's information theory. In Shannon's theory, once there are entropy and the information carried by a number of bits, a whole field opens up, the number base that is used no longer matters, and bandwidth becomes a coherent notion. In the same way, control can be reified as an area representing all possible signals, the proportion of wanted signals among them, and the distribution of wanted signals among unwanted ones, which is the part almost everybody overlooks.

The model is set up in discrete terms on purpose. Information theory needs a quantized system to be coherent, and an artificial intelligence is already quantized. A summation can always be converted to an integral, but if the process is not understood at the level of the summation, it will not be understood at the level of the integral. The generality is what makes the proof work. And this is one construction among several parallel ones that all arrive at the same result; that is reassuring in the way it is for the Pythagorean theorem: there are many proofs, and a reader who does not understand one is likely to understand another.

The Proof

The method of the proof is fairly obvious, and there are no ambiguities: convergence on non-controllability occurs in each of the three stages, and the cycle overall drives the system into a certainty of failure, guaranteed, if *any* non-trivial assumptions of the circuit design of any of the nodes obtain (and all three nodes will be required to be sufficiently complex for any notion of "AGI", "control" and "world of physics" to be at all meaningful).

The boundaries of X. As a first step, we focus on what happens (what must happen) on the boundaries of X, from when it receives inputs to when it generates outputs. We either pre-compute, or just assume, some basically reasonable black and white mask (the pattern marking which signals count as success and which as failure) of what sort of output signal from X must be received by Z for the outcome in Z to be "beneficial", "desirable", or at least to define (in a boolean way) some absolute notion of success or failure. Such a notion, as encoded in some presumed circuit in Z, represents states that are designated as "aligned", or at least "not completely terminal to all biological life otherwise in Z" (whatever that means, as translated into circuits; the specifics of this do not matter for the proof at this point). Then we calculate what sort of degree of control X would need to have on its outputs so as not to trigger the failure condition in Z. This is the boundary metric on the outputs of X (the measure of how much control is needed at the outputs of a block to stay within the success region), and it is defined purely by the compute circuit of X, and hence, indirectly, purely by the inputs into X, as modulated by the circuit of X itself. This establishes a projected mask of what is success and failure onto the inputs of X, based on the hashiness of X, in how X maps inputs to outputs. Then we can compare the characteristic of the boundary metric on X outputs to the same characteristic of the boundary metric on X inputs. The inequality shows up in the comparison of the boundary metric on the outputs of X with the boundary metric on the inputs of X.

The inequality. The degree of control necessary to obtain success on the inputs of X is strictly greater than the degree of control necessary to obtain success on the outputs of X, based on the degree to which the circuit computation of X is complex (hashy). Insofar as the information (productive complexity) of the outputs cannot exceed the information complexity of the inputs (predictive output information cannot be created out of nothing), and insofar as not all input information will be linearly defining of output state, any non-linearities in the calculation of X will decrease the predictive capability of success inputs to ensure the generation of success outputs. Control throughput (how much of the control applied to a block's inputs carries through to its outputs) can be at most unitary, if no computation is done by X, and is less than that to the degree that X is non-linear, up to the point that X is the empty set, where no X inputs are passed to X outputs. With X being the compute empty set, even perfect absolute control of X inputs does nothing to ensure success-correct X outputs. Hence the boundary metric of what Q must provide to X requires at least a higher order of control of X inputs than X must produce for X outputs. And even under perfect conditions, in the general case, Q cannot predict what X will be without actually also implementing X; and even then, there is no guarantee (it is even very likely) that no such Q-to-X bit flow can actually implement success (non-failure) conditions.

A single gate inside X. The simplest possible machine X passes its input bits straight through: every output bit is specified by one input bit, and if there are fewer output bits than input bits, the inputs give perfect control of the outputs. The pass-through machine is the equality case, where the control needed on the inputs equals the control needed on the outputs. Now put a single AND gate (a logic gate whose output is true only when all of its inputs are true) inside X, and map its output to two of the output bits. No matter what the inputs are, some output states can no longer be reached. The control over the inputs no longer gives access to every control state on the outputs: the mapping loses its bijectivity, and keeps only as much surjectivity (the property that every output can be reached from some input) as the complexity of the circuit allows. Whether

the unreachable part of the space is black space or white space depends on the circuit. So the configuration of the inside of X changes the probability of hitting a target. But that is not the main thing being studied. The main thing being studied is the mixing. If the inside of X hashes the input bits onto the output bits, then the message from Q to X can tolerate less noise than the message from X to Z can tolerate.

Q and Z. Whatever hash model logic applies to X also applies, by induction, to Q, since Q must do its calculations prior to X, so that the level of control of the inputs to be provided to Q is even greater than that needing to be passed by Q to X. However, all of the Q inputs are defined from the outputs of Z, whose algorithm is provided to us (as chosen by nature), and not by the design engineers who are creating the Q and X circuit systems. Hence we cannot really constrain the dynamics of Z, and so cannot constrain the inputs that Z provides to Q. Moreover, insofar as these Z inputs to Q are themselves partially defined by clearly non-linear aspects of the internals of Z, the same inequality will for sure hold for the Z element, meaning that the inputs of Z must have an even higher level of control than that defined for the Z outputs. And there is no guarantee that even such success (non-fail) output bits can be specified to exist.

The cycle. Thus, by induction, an update is needed to the level of stricture of the degree of control necessary on the prior X-to-Z flow. This update, an increase of stricture, happens on each complete cycle, until the possibility of control goes to zero, as the mask pattern of success goes to full black. This convergence occurs for any reasonable assumed initial value for the mask on the first X-to-Z state transition. The strict inequality holds at every nonlinear gate; whether the strictness accumulates around the loop, or is paid back by what the controller learns from the returning signal, is settled by the hashiness of the boundary from the system to the world, as the next subsection, Which Way the Loop Converges, states. For a world with the hashy parts that alignment must care about, the bound of the subsection The Proof Needs Only a Conservative Lower Bound places the loop on the side of accumulation.

A formal treatment of the single block X (the degree of control necessary on the inputs to X is strictly greater than the degree of control necessary on its outputs, where black significantly exceeds white, as dependent on the degree of non-linearity, the hashiness, of the circuit X), together with a reasonably good comparable characterization of what is meant by the degree of control necessary, is already enough to derive an AGI uncontainability result under conservative assumptions (for the hashiness of X, the mixing coefficient, and so on). The primary advantage of this method of proof is that it does not require us to consider what is meant by "degree of abstraction" in control theory, and does not require us to import all of the spurious complexity of general control theory either. It enables a great many proof simplifications, while still maintaining full general relevance to AGI alignment theory.

Which Way the Loop Converges

The model is set up as a ring with a fixed clock (Z to Q, Q to X, X to Z) so that its convergence dynamics can be studied over many cycles. The loop can converge either way. If the hashiness at the boundary from X to Z is low, then each time around the loop the hashiness of the boundary from Q to X may go down, meaning that Q gains excellent control of X; and if X is a learning system, the hashiness of its input can decline as it learns. If instead Z is fairly hashy, so that the input to Z is hashy, then each time around the loop the hashiness of the input to X goes up and things get worse: feedback makes Z drift farther and farther from white, and the noise (any bits flowing through the system that we cannot predict) grows until it exceeds any capacity of X to learn Z. Then the inputs to X become noisier still relative to its outputs, and the whole system converges on non-controllability.

The boundary between converging on controllability and converging on non-controllability is set by hashiness. Q is entirely known, because we built it as the control system. Z is mostly unknown. X starts out known, but becomes more and more unknown as it learns things and Q does not

know what it is learning. Z itself can be characterized the same way, by comparing the hashiness of its inputs with that of its outputs; realistically, Z is highly nonlinear, and most inputs to Z lead to very divergent outputs. For the proof, all that needs examining is the inputs to Z: how hashy they are, and whether that is above the minimum threshold for controllability.

The convergence has a second-law structure: the structure of the second law of thermodynamics, under which the disorder of a closed system never decreases. The analogy is not decorative. In a convergent loop one has something like a heat pump: energy can be spent to cause convergence. Above a certain threshold of complexity in the loops, one has instead the *entropy function* (disorder that grows whatever energy is spent), and no amount of energy input yields more order. The criterion reduces to something familiar in form: is this process convergent or divergent? It has the same character as asking whether an infinite series converges (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section VI).

The disorder that grows is in what the controller Q knows of the state of the loop, not in where that state wanders. It is the noise of the paragraph above, the bits flowing through the system that Q cannot predict, growing until Q has no information about a subset of bits sufficient to determine whether the world has entered a degraded state. Read that way, the second-law structure is a comparison of rates, which Chapter 4 of this book makes exact. Under persistent branching, the next state stays uncertain by at least a fixed amount even given the whole current state, control signals included (Sections 4.4.2 and 4.4.3), so the loop creates uncertainty about its own state faster than the controller receives information about it, and the monitor cannot keep pace with the system it monitors (Section 4.9.2). Taken with the safe set as one of its outcomes, persistent branching keeps, at every step and whatever the controller does, at least a fixed chance that the state moves outside the safe set (Introduction, §7); so when the rate at which the loop creates uncertainty about its own state exceeds the rate at which the controller receives information about it, every control law lets safe states leave the safe set. This conclusion does not rest

on the step from unpredictability alone to the claim that a system must leave a region of states, which Section 4.9.2 declines; the boundary comes from the comparison of the two rates. The comparison of rates is the data-rate theorem of control theory (Tatikonda and Mitter 2004; Nair and Evans 2004); *The Ceiling Below the Floor* proves the comparison again, formally, as its data-rate theorem (Chapter 4; Theorems 4.5, 4.10 and 5.5), and rejects that step again (Section 13.5).

The World Has Hashy Parts

The world, Z, has hashy components, and those components drive a good deal of the hashiness in the whole that we care about; the claim is not that the whole of Z is hashy.

An uncontroversial example is the human brain. Treat all sensory input (proprioception, taste, smell, sight and hearing) as a bitstream running into the brain, and all the nerve signals sent to the muscles as the output. Predicting the outputs from the inputs is very hard, because many levels of abstraction are involved, and it cannot be made to go faster: if there were a way to make it more energy efficient, space efficient or time efficient, evolution, which is a good algorithm for finding such things over a billion years, would probably have found it. So the brain is close to the fully hashy end. Human societies are also close to the hashy end, and insofar as alignment is concerned with human societies, it has to account for the hashiness of Z, because that is the part we care about.

The stock market is an exemplar. It takes inputs and produces outputs, and everyone is trying to model it, because whoever can predict the next price makes money. So it converges on unpredictability: the more that human brains and computers can predict of it, the more those predictable parts get factored out for someone's profit. What is left looks like nearly perfect unpredictability, which is equivalent to hard randomness (the randomness of quantum events, whose outcomes no finer measurement and no faster computation can recover before they occur; Section 3.3.6). There are predictable aspects too (the frequency spectra of price graphs follow

power laws, and companies that make money reliably tend to keep their value), but the predictable parts sit right alongside parts that are not predictable. Some aspects of Z are near 0% hashiness, others near 100%, and there are things we care about in both.

Here a narrowing must be headed off. A reader may point out that the world is often easy to predict: release a bottle and it falls. True, and irrelevant. The places where the world is easy to predict are very largely artificial circumstances. The places where prediction is *necessary* are ecosystems. The question that alignment is actually asking is something in the nature of: what is the resilience and homeostasis of this ecosystem, and will the introduction of this thing kill the lawn? Alignment is in every case bound to biological systems in one way or another, and those are characteristically complex (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section V).

The three-node ring of Chapter 2 treats the world as a node that transforms signals and interprets none (Section 2.4.1), and *The Ceiling Below the Floor* (Section 8.2) calls that world simple, one that "pursues no objective and adapts to no controller". That simplification is of the same kind as the ring's sequential timing, which both books call an idealization in the controller's favor (Section 2.4.2): the real world, with its ecosystems and its hashy parts, gives the controller less than the ring grants, and every limit proved in the ring holds of it a fortiori.

How Fragile the World Is

A variable more important than hashiness is how fragile the aggregate system is (how readily it is moved out of its stable state by a disturbance at one of its weak points). Life is robust in the way an aircraft is robust: it handles a wide range of temperatures, pressures, weather and operations, and yet any real aircraft has fragile points, and if one of them is disturbed, things go drastically wrong.

So we take the space of all possible states of the world system, with as many dimensions as we like (resource use, planetary warming, inequality, and so on). Then we ask: over what range in each dimension is the system robust, and outside which range does the system become fragile? And how likely is a perturbation that moves the system from one metastable state into another (a metastable state being one that holds against small disturbances but can be pushed into another state by a large enough one)? In terms of complexity theory, we ask what the adjacent stable states are and what the transition probabilities to them are. The adjacent stable states, and the transition probabilities between the stable states, make a phase diagram in the hyperspace. A phase diagram is a map of the space of a system's states, showing the regions where the system is stable and the boundaries where the system changes into another state. Where the stable region is a broad plane in one direction and a narrow slice in another, controlling for the broad direction is of little use; one has to control for the narrow one. Have we accounted for all the narrow dimensions? The question of whether all the narrow dimensions are accounted for is the fragility measure that is wanted, and the question amounts to asking how probable a black swan event (a rare, unanticipated event with very large consequences) is.

How do we evaluate fragility at all? One way is to ask how rare the state is in the space of all possibilities. That can be turned into other questions: how rare is it in the universe; how much would it take to reproduce it; if it were moved out of that state, what would it take to move it back, and how able are we to undo a black swan? We could use something like simulated annealing over the hyperspace (a search method that explores a space of states by random moves that become more selective over time), or macroeconomic methods: what would it cost to move the world from one point to another; how many levels of abstraction would we need to cross, with how much fulfillment at each level, to recreate it? In effect we ask how likely life on this Earth is, which connects to the conditions of the Fermi paradox (the question of why, in so large a universe, no other technological life is seen). Given its rarity, how likely is it that the system has

at least some fragile dimensions, what control would we need to nudge it back if it diverged, and can we sense over enough dimensions for that nudging to work?

Everything About the Machine Is Output to the World

The interface between the machine, X, and the world, Z, is not only the actions the AGI takes. It is literally everything about the physicality of the AGI itself. Some outputs from X to Z are the things the AGI was built to do. Some are the result of its existence. Some are the result of its building and its maintenance. If hexavalent chromium is a side effect of welding the stainless steel in its structure, then the release of hexavalent chromium is an output of X into Z, and if that happened to be one of the factors that tips Z into some cascade, Z is fragile with respect to that output of X. The software engineer's concern with modularity applies here: what the interface conditions are, and what the interface actually maps.

Anders Sandberg has pointed out that one hallmark of intelligence is that it makes low-probability things happen with high reliability, so an intelligence must be able to create fragile points, or it is not useful. That is why the interface has to be characterized. As X explores the space of everything it could do that is effective in influencing Z the way it wants, it moves into regions where it is more likely to have effects it did not account for or anticipate. The more potent the flow from X to Z, the greater the fragility of Z with respect to X that comes with it.

Abstraction Enters as Dimensions of the Phase Space

Calculating over levels of abstraction is hard; control theory works well when the level of abstraction is low, and badly when it is high or unknown. Abstraction is best not characterized directly, because that feels like violating the spirit of mathematics: if something abstract is represented inside something concrete, how can we ever trust that it was done right? What is a human? We could say the collection of a trillion cells sitting over there, but that is only an exemplar of a person; the notion of hu-

man is more abstract. Everywhere a concrete representation is substituted for an abstract truth, there is a risk of substituting a specificity in an inappropriate way.

So the problem is translated into an easier one, a different specificity, acknowledged to be compensatory. Whatever levels of abstraction someone wants represented show up in the dimensionality of the phase space that describes where stability holds and where it breaks into transition. If some future person says, "You did not account for factor P", the answer is: "Great. Add it to the list of dimensions, and let us find the region where homeostasis holds (where the system holds itself within the range of states in which it stays healthy) and where degradation, collapse or transition happens". That becomes one more dimension, an incremental update to the black swan probability, which then translates down into the X-to-Z interface.

The choice of encoding matters here. Black swan calculations are normally done in a one-hot space (each possible value has its own bit, with exactly one bit set), while control coefficients between X and Z are normally reckoned as bandwidth, in a binary representation (the value encoded in the pattern of all the bits together). In a binary representation, abstraction shows up in the mixing coefficient; in a one-hot representation, it shows up in which basis dimensions describe the space. The relationship between the dimensionality of the phase space and abstraction has to be carried correctly into the characterization of the X-to-Z boundary, or it will be missed. And since a binary code can always be translated into a one-hot code without losing information, and any finite continuum can be mapped one-to-one onto an infinite one, the dartboard, with its area of white and its mixing of white and black, fully characterizes 100% of the bandwidth flowing from X to Z.

The Proof Needs Only a Conservative Lower Bound

In this setup, area and mixing (accuracy and precision) are substituted for the notion of control, and hashiness for the notions of complexity and intelligence. At the last step, Z no longer needs to be parameterized at all. All that needs to be known is that the hashiness of X has to be reasonably high, or it does not work as X. The real world has highly hashy aspects, and if X is to be useful, its hashiness must be within an order of magnitude of the thing it is trying to control. The requirement that the hashiness of X be within an order of magnitude of the thing X controls is modeling theory again. If intelligence is to be used to predict the stock market, as the quantitative and high-frequency trading systems already do, the intelligence has to be about as hashy as what the intelligence models.

And the mixing coefficient on the X-to-Z boundary is not going to be low; it will be somewhere between 0.5 and 1. Every newly noticed level of abstraction, every newly found dimension of fragility, increases the complexity of the phase space and of the boundary between stable and unstable regions, and so shows up as an increase in the mixing coefficient. Since the mixing coefficient only goes up as we learn more, we can set a lower bound on it, even a very conservative one, and every later argument only raises that bound. With a lower bound on the mixing coefficient and a lower bound on the hashiness of X, the calculation no longer depends on abstraction at all.

Treating X as a straight pass-through is the most optimistic case, with no information loss. If the controller Q is lossy, there is a third color on the board, call it red: regions on the X-to-Z boundary that no input from Q can reach. If any of the red regions covers anything important, the control system by definition cannot constrain the output of X. However one goes about constraining the output of X, somewhere one meets a contradiction to one of the inequalities on which control depends. And the more Q is integrated into X and X into Z, the more drastic those inequalities become.

1.9 Implications for Modeling and Control

The non-amplification argument and the hashiness framework have several implications for the foundations of control theory, and through it, for the AI alignment problem specifically.

1.9.1 Ashby's Law, Sharpened

The law of requisite variety, due to W. Ross Ashby, states that a controller must have at least as much variety (in the information-theoretic sense) as the system it controls.[6] Our result adds a computational dimension: even with sufficient variety in the input, the opacity of the intermediary can prevent the agent from *using* that variety effectively. Controllability is bounded not only by information capacity but by computational tractability. A human operator may have access to an enormous space of possible instructions to give an AI system, but if the mapping from instructions to behaviors is opaque, the human cannot search that space for the instructions that produce the desired behavior.

1.9.2 Modeling Limits

The framework makes precise the intuition that approximate models provide limited control. If the agent knows the intermediary only approximately, so that the true circuit is one of several candidates, then the analysis of Layer 3 applies, with the uncertainty about the model playing the role of the unknown internal state. Each candidate circuit produces a different control region, and the agent must find an input that works for all of them. Uncertainty about which candidate circuit is the true one is the hidden-state problem in another form, and hashiness grows with that uncertainty in the same way that hashiness grows with hidden state.

This result has immediate relevance to interpretability research in AI safety. Interpretability tools provide approximate models of a neural network's behavior. The hashiness framework quantifies the fundamental

limit on how much control such approximate models can deliver: they can reduce the effective hidden state (by revealing some of what was previously opaque), but they cannot eliminate it, and whatever hidden state remains contributes to hashiness in exactly the same way. An interpretability tool that explains 90% of a model's behavior still leaves 10% unexplained, and that 10% may contain precisely the failure modes that the controller most needs to prevent.

The limit that remaining hidden state places on interpretability is an argument for realistic expectations about what interpretability can achieve, not an argument against interpretability research. Interpretability tools are valuable for reducing hashiness: they move the system from higher hashiness toward lower hashiness by converting hidden state into observable state. But they cannot drive hashiness to zero unless they render the system completely transparent, which would require a model of the system that is as complex as the system itself. For a system that is more computationally capable than the controller, this is by definition impossible. The controller would need to understand every aspect of a system that, by hypothesis, can think thoughts the controller cannot think.

1.9.3 Noise as Fundamental

Noise is a structural feature that compounds with circuit opacity. The cascaded degradation argument shows that noise at *any* point in the chain reduces the information available at the output. This suggests fundamental limits on control precision in physical systems, where noise is unavoidable, and these limits become tighter as system complexity increases. An engineering project that attempts to maintain precise control over a system of increasing complexity must contend with the joint effect of growing opacity and irreducible noise, a combination whose information-theoretic cost is multiplicative.

1.9.4 Compressibility and the Epistemic Basis for the Uniform Distribution

A natural question about the hashiness measure is why the uniform distribution over the internal state is the right choice. The quantitative generic results of Section 1.6 use this distribution and reason about random circuits, making them average-case results (see the Remark on Three Tiers of Support, Section 1.6). A reader may point out that the actual systems of concern (trained neural networks, deployed AI systems) are not random circuits and may not have uniformly distributed internal states. Could the actual distribution of the internal state place the system in a low-hashiness regime?

This question can be answered through a structural argument about learned representations. A learning system optimizes its internal state for compressibility: it finds patterns in data and encodes them as efficiently as possible. The result of this optimization is a state that is a maximally compact encoding of the system's accumulated experience. From the perspective of an external controller who cannot model the compression algorithm at its full depth, such a state is indistinguishable from a random string. That a maximally compact state is indistinguishable from a random string, to a controller who cannot model the compression, is a direct consequence of Kolmogorov complexity theory, and not a metaphor. The conclusion is supplied again without any claim about how compact the state is: whatever part of the state the controller cannot model, the controller has no knowledge that narrows, and for that part the uniform distribution is the distribution the controller has. A string that is the output of an effective compression process appears random to any observer who does not possess the decompression key. The controller's inability to model the system's internal organization (which, by hypothesis, is more computationally sophisticated than the controller) means the controller cannot distinguish the actual internal state from a uniformly random one.

If the controller cannot distinguish the actual state from a random state, then the uniform distribution over the internal state in the hashiness definition is not merely a mathematical convenience. It is an accurate model of the controller's epistemic situation. The average-case hashiness is the controller's best estimate of the system's controllability, because the controller has no basis for narrowing the distribution. To sustain the claim that the practical case might be easy, one would need to show that the controller possesses some structural knowledge of the internal state that narrows the effective distribution away from uniform and places the system in a low-hashiness regime. No such showing has been made, and the compressibility argument suggests that, for systems whose internal computation exceeds the controller's modeling capacity, no such showing can be made. The average case is the practical case when the controller cannot tell the difference.

1.9.5 Three Regions of the Spectrum, and What Recursion Does in Each

The spectrum of Section 1.2, from the linear system to the hash function, divides, on analysis, into three distinct regions.

In the first region, the left side of the spectrum, control theory works well. The system exhibits enough regularity that averaging over its behavior smooths the variation. The controller can treat the system's output as a signal with noise, extract the signal, and make reliable predictions. The first region is the classical domain of engineering control.

In the third region, the right side of the spectrum, no version of control theory is possible. The system has crossed a threshold beyond which no algorithm, no feedback mechanism and no predictive model can extract useful regularities from the system's behavior. The output is, from the standpoint of any feasible computation, indistinguishable from randomness. The third region is the hash-function regime. Section 1.2.4 places the superintelligent system at compound maximum hashiness, in this region.

Between these two regions lies the second: the *undecidable region*. In this middle zone, it is not possible to determine in general whether control theory will succeed or fail for a given system. Some systems in this region may admit effective control under specific conditions; others may not. But there is no general procedure for distinguishing the tractable cases from the intractable ones. The boundary between controllable and uncontrollable is itself undecidable (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section IV). The step that makes it so is Rice's theorem (Section 5.3.3; *The Ceiling Below the Floor*, Theorem 12.3): whether a given control keeps a given system within its target region is a nontrivial property of what the system does, some systems have it and some lack it, and no procedure decides such a property for every system.

The three-region structure becomes especially vivid when the mapping is applied recursively: that is, when the output of one application becomes the input to the next. Consider a system in which data is passed through a hash function, and the output is fed back as input, perhaps with some additional data or randomness added at each step. Feeding the output of a hash back in as input is, in structure, exactly the process that a *blockchain* employs, and that process is the recursive hash of The Satisfiability Connection, above. Each block's hash depends on the previous block's hash, creating a chain in which each link is the product of an irreversible computation applied to the previous link. Is there any way to predict the future state of the blockchain? Of course not; or else someone would have done it, and the entire edifice of cryptocurrency security would collapse. That reply, like the reductio of Section 1.2, illustrates and does not carry the weight. The reason the chain cannot be predicted is precisely that the hash function occupies the far-right region of the hashiness spectrum. Under recursion, whatever regularities might exist in a single application of the function are compounded and destroyed with each iteration. The system moves irreversibly toward maximal opacity.

Now consider the same recursive structure applied to a mapping with less extreme hashiness: a mapping that introduces some non-linearity, but not enough to destroy all regularity. Under recursion, the behavior of such a system depends sensitively on where on the spectrum it falls. In the controllable region, recursion may amplify regularities as much as it amplifies noise, and control remains feasible. In the hash region, recursion rapidly destroys any residual structure. In the undecidable middle, the long-term behavior is contingent on details that cannot be assessed in advance (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section V).

The canonical example of a computational phase transition with this three-region structure is the satisfiability problem, SAT: the problem of determining whether a given Boolean formula has an assignment of variables that makes it true. SAT is the foundational NP-complete problem: every problem in NP (the class of problems whose solutions can be checked quickly) can be reduced to it, and it is the problem on which the entire edifice of computational complexity theory pivots. The phase transition in SAT occurs as a function of the ratio of clauses to variables in a randomly generated instance. Below a critical value of this ratio, most instances are satisfiable, and satisfying assignments can be found quickly. Above the critical value, most instances are unsatisfiable, and proving unsatisfiability is also comparatively tractable. But at the critical ratio itself, in a narrow band around the threshold, instances are maximally hard. It is in this critical region that instances take the longest to solve, that algorithms exhibit the most extreme variation in running time, and that the distinction between satisfiable and unsatisfiable is most difficult to determine. The behavior of SAT below, at and above the critical ratio is precisely the three-region structure. Below the threshold: easy, tractable, controllable. Above the threshold: also tractable, in a different way, since the answer is almost always "no", and this can be established efficiently. At the threshold: maximally hard, maximally uncertain, and in a precise sense the hardest instances for any heuristic algorithm that does not perform exhaustive search (*Conditions of Continued Existence; Closed Loops,*

Cascading Failure, and the Limits of Control, Chapter Five, Section VIII). The analogy supports the picture of three regions; the picture itself rests on the spectrum and on Rice's theorem, as above.

1.10 Summary and Forward References

In this chapter I have introduced *hashiness* as a measure of the degree to which a computational intermediary degrades an agent's ability to steer a system toward a desired outcome. The central result, the non-amplification argument, establishes that no intermediary can increase the agent's causal influence over the outcome: the information that the agent's input carries about the final output can only be preserved or degraded through each processing step, never enhanced. Equality holds only for information-preserving (injective) transformations. The non-amplification argument gives a universal result: it holds for every circuit, every hidden state, every target function, and it is distinct from the average-case results about hashiness scaling that the framework also provides (the Remark on Three Tiers of Support, in Section 1.6, sets out the precise relationship between these tiers of support).

The four-layer decomposition (target receptivity, structural opacity, state obscurity, noise degradation) provides a structured way to analyze where controllability is lost. The connection to the data-processing principle grounds the framework in well-established information theory. The framework also connects to circuit complexity (via the NP-hardness of inverting opaque circuits), boolean function analysis (via influence and noise stability), and computability theory (via Rice's theorem for Turing-complete intermediaries). Throughout, "control" is defined as causal influence (the degree to which the agent's input determines the outcome), which is formalized via mutual information (the Luck-Control Distinction, Section 1.7).

In this chapter I give a *synthesis*, rather than any single new argument. The synthesis is a unified language for discussing controllability degradation, and the language spans information theory, complexity theory and

boolean function analysis. The language is grounded in a concrete computational model (finite boolean circuits) that admits exact computation for small instances and analysis for asymptotic behavior.

The results of this chapter are self-contained and hold for any boolean circuit system. The chapters that follow build on this foundation in several directions:

Chapter 2 extends the framework to systems operating at different levels of representational abstraction, introducing the *abstraction differential* and the control-limit argument, which establishes that controllability approaches zero as the gap between the controller's and the system's levels of abstraction grows.

Chapter 3 integrates the information-theoretic results of this chapter with computational impossibility results (Rice's theorem, Chaitin incompressibility) and physical arguments to produce the impossibility synthesis.

Chapter 4 provides the dynamical-systems complement, extending the static analysis of this chapter to the temporal domain using Kolmogorov-Sinai entropy.

Chapter 2 of Volume 1 (Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains) applies the framework specifically to the human-AI relationship, showing how the four layers of degradation map onto the specific properties of human oversight of AI systems.

Chapters 1 to 4 constitute Part I of this book, The Limit on Control: the core argument that AGI safety faces barriers of principle, not merely practical ones.

A final word on the nature of the claim. The hashiness framework says that the *mechanism* by which we ensure good behavior matters, and that one particular mechanism, external control through input manipulation, is subject to mathematical limits that grow tighter as the system becomes more capable. It does not say that AI systems can never be made to behave well. If we are relying on our ability to choose the right inputs (the right

training data, the right reward function, the right constitutional principles, the right shutdown command) to ensure that a system of superhuman capability produces safe outputs, then the non-amplification argument tells us that our ability to do so is bounded by the information that survives the passage through the system's internal computation. For a system that is more computationally sophisticated than its controller, this surviving information approaches zero.

The limit on control through input manipulation is a reason for clarity, not for despair. Impossibility results in other fields (Gödel's incompleteness theorems, Arrow's impossibility theorem, the halting problem) have not prevented progress in mathematics, social choice theory, or computer science. But they have prevented *misguided* progress: attempts to do things that cannot be done, wasting resources and creating false confidence. The hashiness framework aims to serve the same function for AI safety: to delineate what is achievable and what is not, so that research effort can be directed toward strategies that respect the constraints rather than strategies that unknowingly violate them.

Notes

1. T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. (Wiley-Interscience, 2006), Chapter 2, Section 2.8 (the data-processing inequality). The data-processing inequality, which this book calls the data-processing principle, is one of the foundational results of information theory and has been independently rediscovered in several fields. Its application to control theory is natural; I believe it has not previously been developed in the specific form presented here.

2. R. O'Donnell, *Analysis of Boolean Functions* (Cambridge University Press, 2014), Chapters 2 and 5. O'Donnell provides the definitive treatment of influence and noise stability in the boolean function context. The connection to controllability via the hashiness framework is new.

3. S. Arora and B. Barak, *Computational Complexity: A Modern Approach* (Cambridge University Press, 2009), Chapter 6 (Boolean circuits). The NP-completeness of Circuit-SAT is a classical result. The connection to the control problem (specifically, the observation that finding a control input for an opaque circuit is an instance of Circuit-SAT) is straightforward but has important implications for the computational feasibility of alignment verification.

4. H. G. Rice, "Classes of recursively enumerable sets and their decision problems", *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358 to 366. Rice's theorem is discussed extensively in Chapter 3 and its application to AGI is developed in Chapters 7 and 9.

5. S. Wolfram, *A New Kind of Science* (Wolfram Media, 2002), Chapter 12 (The Principle of Computational Equivalence), Section 6 (Computational Irreducibility). Computational irreducibility is the principle that, for many computations, there is no significantly shorter computation that predicts the outcome. The only way to know what the system will do is to run it.

6. W. R. Ashby, *An Introduction to Cybernetics* (Chapman & Hall, 1956), Chapter 11 (Requisite Variety). The law of requisite variety was one of the earliest results in control theory. Our result can be understood as a computational extension of Ashby's law: variety is necessary but not sufficient for control; the controller must also be able to *use* its variety effectively, which requires that the intermediary be sufficiently transparent.

References

Arora, S., and B. Barak. *Computational Complexity: A Modern Approach.* Cambridge University Press, 2009.

Ashby, W. R. *An Introduction to Cybernetics.* Chapman & Hall, 1956.

Cheng, D., and H. Qi. "Controllability and observability of Boolean control networks". *Automatica* 45, no. 7 (2009): 1659 to 1667.

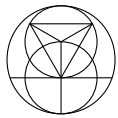
Cover, T. M., and J. A. Thomas. *Elements of Information Theory.* 2nd ed. Wiley-Interscience, 2006.

Goldreich, O. *Foundations of Cryptography.* Vol. 1: Basic Tools. Cambridge University Press, 2001.

O'Donnell, R. *Analysis of Boolean Functions.* Cambridge University Press, 2014.

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.

Wolfram, S. *A New Kind of Science.* Wolfram Media, 2002.



Chapter Two

Control Limits Under Abstraction Differential

A discrete information-theoretic framework establishing absolute bounds on controllability when controller and controlled system operate at differing levels of representational abstraction

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

Chapter Summary: This chapter shows that control degrades with the *abstraction differential*: the gap between the level of representational abstraction at which a controlled system operates and the level at which its controller operates. Abstraction levels are defined by interpreter structure, and each level that the controller cannot model adds at least as many hidden state bits as the minimum interpreter specification size (the fewest bits that specify a universal interpreter for one level of abstraction). The configurations the controller cannot tell apart therefore grow exponentially with the differential, and controllability approaches zero; a single level of differential already makes control effectively impossible wherever a failure rate of one in a million is unacceptable, and two levels suffice even for an interpreter far smaller than any practical one. The three-node ring of controller, complex system and world represents every configuration of control relationships. In it no architectural privilege restores control: a kill switch that the controller does not know when to press is not control; it is a coin flip with extreme consequences in both directions. Delegation to another high-abstraction system introduces a new control problem of equal or greater difficulty, and even a cooperative system faces the same limits in controlling itself: there is no perfected benevolence, and alignment decays in any learning system. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system.

2.1 The Problem of Control Across Abstraction

The previous chapter established a fundamental result: no computational intermediary can amplify an agent's control over a system. The non-amplification argument, derived from the data-processing principle applied to boolean circuits, sets an absolute ceiling on controllability that holds regardless of the intermediary's intelligence, architecture, or resources. Control can be preserved or degraded through an intermediary, but never enhanced. This result belongs to information theory, and it is as secure as the second law of thermodynamics.

But the non-amplification argument, by itself, does not tell us *how much* control is lost in any given situation. It tells us that control cannot increase; it does not tell us how rapidly control decreases, or under what conditions it collapses entirely. In this chapter I address how rapidly control decreases, and under what conditions control collapses entirely. We develop a framework that takes the non-amplification argument as its foundation and builds upon it a quantitative theory of control degradation: a theory that specifies, for any given configuration of controller and controlled system, the maximum achievable control and the conditions under which that maximum approaches zero.

The key variable turns out to be the *abstraction differential*: the gap between the level of representational abstraction at which the controlled system operates and the level at which the controller operates. When the controller and controlled system operate at comparable levels of abstraction, classical control theory applies, and effective control is achievable in principle. When the controlled system operates at a higher level of abstraction than the controller, a new regime emerges, one in which control degrades exponentially with the differential, and beyond a threshold, becomes effectively impossible regardless of the controller's strategy, resources, or architectural privileges.

This result has direct implications for the question of AGI safety. A human institution operating at a fixed (or slowly growing) level of organizational and cognitive abstraction, attempting to control an artificial system whose representational abstraction is growing without bound, will eventually, and inevitably, lose effective control. The question is not *whether* this happens but *when*, and the framework developed here provides the tools to analyze that question rigorously.

We proceed in five stages. First, we establish why the analysis must be conducted in discrete mathematics rather than the continuous mathematics of classical control theory (Section 2.2). Second, we formalize the notion of abstraction levels and the abstraction differential (Section 2.3). Third, we introduce a universal base model, the three-node ring, that can represent any configuration of control relationships (Section 2.4). Fourth, we argue the main results: the control-limit argument and its corollaries (Section 2.5). Fifth, we examine the implications, including the inefficacy of architectural privilege, the impossibility of delegation, and the consequences for the cooperative case (Sections 2.6 to 2.8).

2.2 Why Discrete Mathematics

Nearly all classical control theory is formulated in continuous mathematics. Differential equations, Laplace transforms, optimization over real-valued parameters, and convergence arguments in real analysis form the standard toolkit. This mathematical apparatus has been extraordinarily successful: feedback controllers, PID loops, Kalman filters, and optimal control strategies have been applied to systems ranging from chemical plants to spacecraft to financial markets. The success of continuous control theory has created a deep, largely unexamined assumption in the engineering community: the assumption that control techniques can, in principle, be extended to systems of arbitrary complexity given sufficient resources and ingenuity.

In this chapter I challenge the assumption that control techniques can be extended to systems of arbitrary complexity, and I begin the challenge with the choice of mathematical foundations.

Anders Sandberg names the two strongest questions for a framework such as this one. First: do we have a good enough theory of the limits of controllability? And second: does the level of abstraction actually make the limits more significant? Sandberg's two questions are exactly the right questions; the second question, whether the level of abstraction makes the limits more significant, is the subject of Sections 2.3 to 2.5. Take the first one. Control theory is a beautiful framework for linear systems. For fairly nonlinear systems, there are somewhat applicable theorems. But most of them are based on keeping something in the vicinity of a point (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section I). A theorem about keeping a system near a point is not a theory of how far control can reach, so the existing toolkit leaves the first question open, and the sections that follow take it up.

Continuous mathematics embeds an assumption that is physically unrealizable: that parameter values can be specified with infinite precision. A real number carries, in principle, infinite information in its decimal expansion. When continuous functions are composed and iterated, as they are in any optimization or control algorithm, this infinite bandwidth is exploited implicitly. Information flows through the low-order bits of real-valued parameters, bits that no physical system can actually represent, transmit, or compute with.

Consider a simple example. Suppose a controller must set a voltage to within 0.001 volts of a target to achieve a desired outcome. In continuous mathematics, the controller simply sets the voltage to the exact target value. In physical reality, there is thermal noise, quantization error, component tolerance, and electromagnetic interference. The controller sets the voltage to some value within a range, and the outcome depends on exactly where within that range the actual value falls. Classical control theory handles the uncertainty in the actual voltage through noise analysis

and robust control design. But classical control theory handles that uncertainty within the continuous framework, treating noise as a perturbation to be minimized rather than as a fundamental limit on information transmission.

The problem becomes acute when the system being controlled is itself a computational system operating at high abstraction levels. In such a system, the "noise floor" is not thermal fluctuation but *computational opacity*: the inability of the controller to predict what the system will compute, given the controller's limited model of the system's internal state. This opacity is not a noise source to be filtered; it is a structural feature of the computation itself, and it exists even in a perfectly noiseless environment.

To expose this structural limitation, the framework developed here is built entirely on discrete (boolean) mathematics. Every signal is a finite bit vector. Every transformation is a finite boolean circuit. Every computation halts in finite time with finite resources. There are no real numbers, no infinite precision, no hidden bandwidth in the low-order bits. This ensures that any limits we establish are genuine physical limits, not artifacts of mathematical idealization.

The choice of discrete foundations is not a restriction on the generality of the results. Every physical computer operates discretely: it stores bits, computes boolean functions, and transmits finite messages. Continuous models of these systems are *approximations*: often useful approximations, but approximations nonetheless. The discrete framework captures the actual information-processing capacity of physical systems, with no excess. If a limit holds in the discrete framework, it holds for every physical implementation, because every physical implementation is a special case of the discrete model (possibly with additional noise, which can only degrade control further, never improve it).

Conversely, results proved in the continuous framework may fail to hold in physical implementation, because they may depend on the infinite bandwidth that continuous mathematics provides for free and that no physical system possesses. A control scheme that works beautifully in a

differential equation model may fail catastrophically in a discrete implementation, not because of engineering error, but because the reasoning in the continuous model relied on information that does not exist in the physical system.

This concern is borne out in practice. The phenomenon of *numerical instability* in computational implementations of continuous algorithms is precisely the consequence of finite-precision arithmetic failing to support the implicit infinite-precision assumptions of the continuous formulation. In control systems, this manifests as unexpected oscillation, divergence, or loss of stability when a theoretically optimal controller is implemented in digital hardware. The standard response is to add engineering margin: to design the system to be stable despite the numerical issues. But this response, while pragmatic, does not eliminate the fundamental gap between what continuous mathematics promises and what discrete hardware delivers. The framework developed here closes that gap by never opening it in the first place.

2.3 Levels of Representational Abstraction

The intuitive notion of "levels of abstraction" is familiar to every computer scientist. A machine language program operates at a lower level of abstraction than a high-level language program, which operates at a lower level than a program written in a domain-specific language, which operates at a lower level than a system that generates domain-specific languages on the fly. Each layer of interpretation adds a new level of indirection between the physical computation and the effective behavior.

To make this intuition precise enough for analysis, we define abstraction levels in terms of interpreter structure.

Definition (Abstraction Level).

A boolean circuit K operates at abstraction level 0 if its input-output behavior is fully determined by its gate-level structure, with no internal state bits that function as a program or data to be interpreted.

A circuit K operates at abstraction level n ≥ 1 if it can be decomposed as an interpreter I operating at level $n - 1$, together with a program/data component P stored in its state bits R, such that K's behavior is determined by I executing P on the control input N.

A simple circuit that holds no program (an adder, a multiplexer, a shift register) operates at level 0. Its behavior is determined entirely by its wiring. There is no program; there is only hardware.

A CPU executing a fixed program operates at level 1. The CPU hardware is the level-0 interpreter; the program loaded into memory is the data being interpreted. The CPU's behavior is not determined by its wiring alone; it depends on which program is loaded. Change the program, and the behavior changes, without any modification to the hardware.

A virtual machine running on that CPU operates at level 2. The CPU (level 0) interprets the virtual machine software (level 1), which in turn interprets the guest program (level 2). There are now two layers of interpretation: the guest program's meaning is determined by the virtual machine's semantics, which are in turn determined by the CPU's execution of the virtual machine code.

A learning system that modifies its own objective function operates at a level that increases over time. At deployment, it might operate at level 2 or 3. After sufficient learning, it may have constructed internal representations that function as interpreters for new, self-generated programs: programs whose existence was not anticipated by the system's designers and whose behavior cannot be predicted from the system's initial configuration. Each such construction raises the effective abstraction level by at least one.

The critical observation for control theory is this: each level of abstraction introduces a new layer of indirection between the system's physical inputs and its effective behavior. From the perspective of an observer who can see only the system's inputs and outputs (and perhaps some intermediate signals), each additional abstraction level adds a layer of opacity. The observer can see what goes in and what comes out, but the mapping between them becomes progressively less predictable as the number of interpretive layers increases.

> Definition (Interpreter Structure).
>
> An *interpreter structure* at level n is a tuple (I, P, E) where:
>
> - I is a boolean circuit (the interpreter) operating at level $n - 1$
> - P is a program/data string of p bits
> - E is an execution state of s bits
>
> The interpreter I takes the program P, the execution state E, and the input, and produces an updated execution state and the output. The complete system at level n is the interpreter I running P on the input, with the state bits encoding P together with E.

> Definition (Abstraction Differential).
>
> Given two systems A and B operating at abstraction levels a and b respectively, the abstraction differential is the difference a − b. When the difference is positive, system A operates at a higher level of abstraction than B.

The abstraction differential is the central variable in the theory that follows. Everything reduces to it: the degree of control degradation, the threshold beyond which control becomes impossible, the efficacy of architectural privileges, the viability of delegation. All are functions of the abstraction differential.

2.3.1 How Abstraction Creates Hidden State

We now establish the precise connection between abstraction levels and the hashiness framework introduced in Chapter 1. The key result is that each level of abstraction that the controller cannot model contributes a definite number of hidden state bits, and these hidden state bits directly determine the hashiness, and therefore the uncontrollability, of the system.

The Ceiling Below the Floor proves this connection as Theorem 2.15 (Hidden state from levels of abstraction).

> Definition (Effective State Bits at Level n).
>
> For a system operating at abstraction level n, the effective state bits at level n are defined as the size of the program component P plus the size of the execution state component E at that level.

> Result (Abstraction Level Separation).
>
> Let C be a system operating at abstraction level b, and let X be a system operating at abstraction level $a > b$. Then C cannot simulate X without access to additional state bits at least equal to the difference between their effective state counts.

The significance of this result for control is immediate. If C cannot distinguish between different configurations of X, then C cannot choose different control actions for those configurations. From C's perspective, all configurations that agree on the bits C can model but differ in the opaque bits are indistinguishable. If the correct control action differs across these configurations, as it generally will, then C must choose a single action that is applied to all of them, and this action will be wrong for some.

> Result (Opacity of Higher Abstraction Levels).
>
> If system C at level b attempts to predict the behavior of system X at level a > b, and C does not have explicit access to X's state bits at levels b+1 through a, then from C's perspective these bits function as hidden state in the hashiness framework. The effective number of hidden state bits is at least the total size of X's program components across the opaque levels, provided that the programs of those levels cannot be derived from one another.

This result is the bridge between the abstract notion of abstraction levels and the concrete machinery of the hashiness framework. The upper abstraction levels of X (the ones that C cannot model) map directly onto hidden state bits in the non-amplification argument. Every result established about hashiness and hidden state in Chapter 1 now applies to the abstraction differential problem.

2.3.2 Quantitative Bounds on Hidden State Growth

We now establish how the number of hidden state bits grows with the abstraction differential. The key quantity is the minimum size of a universal interpreter.

Definition (Minimum Interpreter Specification Size).

Let the interpreter size at level n be the minimum number of bits required to specify a universal interpreter at level n, that is, an interpreter capable of executing any program at level n. By standard results in computability theory, the interpreter size at level n grows at least linearly with n: it is at least some fixed positive constant multiplied by n.

For practical computing systems, the interpreter size is typically hundreds to thousands of bits per level (the size of a minimal virtual machine specification). We call the minimum specification size of a single-level universal interpreter the *minimum interpreter specification size*: the fewest bits that specify a universal interpreter for one level of abstraction.

The value of the minimum interpreter specification size is not known precisely, but it is bounded from both sides. From below: a universal interpreter must be complex enough to simulate any program, which requires at minimum a Turing-complete specification. The smallest known Turing-complete systems (Rule 110, certain small universal Turing machines) require on the order of dozens to hundreds of bits to specify. From above: practical universal interpreters run from thousands of bits (a LISP evaluator) to millions (the Java Virtual Machine, the x86 instruction set). For the purposes of this analysis, the exact value does not matter. What matters is that the minimum interpreter specification size is finite and greater than zero: each abstraction level contributes a nonzero, bounded number of hidden state bits.

Result (Hidden State Growth with Abstraction Differential).

Let C operate at abstraction level b and X operate at level a. The effective hidden state bits from C's perspective are at least the abstraction differential multiplied by the minimum interpreter specification size, provided that the programs of the opaque levels cannot be derived from one another.

For each unit increase in abstraction differential, at least as many new hidden state bits as the minimum interpreter specification size are added.

For a learning system the condition holds, because each level it builds carries programs its designers did not anticipate (Section 2.3).

The hidden state grows at least linearly with the abstraction differential. Since each hidden state bit doubles the number of configurations that the controller cannot distinguish, the total number of indistinguishable configurations grows *exponentially* with the differential. This exponential growth is the engine that drives control to zero.

Result (Hashiness Bound from the Abstraction Differential).

Let the hashiness of circuit K with respect to target L be as defined in Chapter 1. For a control problem where C (at level b) attempts to steer X (at level $a > b$) toward target L, *under the worst-case target-confounding hypothesis*, namely that the inaccessible bits independently gate the target (equivalently, the circuit is maximally confounding on the hidden bits with respect to the target), the hashiness is bounded below by a quantity that approaches one exponentially in the number of those bits.

Remark. This bound is conservative. It assumes the worst-case structure of hidden state from the controller's perspective. In practice, the structure of interpreter hierarchies often makes the effective hashiness even higher, because the hidden state bits are not independent but interact in complex ways that further reduce the controller's ability to achieve desired outcomes. The bits at each level are structured as programs, and programs have exponentially many possible behaviors relative to their length. The linear growth of hidden bits with the abstraction differential understates the true difficulty.

2.4 The Three-Node Ring Model

With the abstraction differential formalized, we need a model of control relationships within which to apply the theory. The literature on control contains many specialized models: feedback loops, hierarchical controllers, distributed control, market mechanisms, embedded oversight. Each captures some aspect of the control problem, but their variety makes general conclusions difficult. A result argued for a feedback loop model might not apply to a market mechanism model, and vice versa.

We seek a base model that satisfies three criteria simultaneously. It must be *universal*: any other control configuration can be represented as a special case. It must be *minimal*: it must contain no unnecessary structure that might constrain the analysis. And it must be *maximal in control potential*: it must grant the controller every possible advantage, so that any limits we derive are lower bounds on what any real system could achieve.

The model that satisfies all three criteria consists of three computational nodes arranged in a directed ring.

2.4.1 Structure

The ring has three nodes: the controller C, the complex system X and the world W.

Node C (Controller). Receives signals from Node W. Produces signals sent to Node X. Represents the locus of control intention: the system attempting to steer outcomes. Operates at a fixed or slowly growing abstraction level b, bounded above by some ceiling H. In the AGI safety context, Node C represents human institutions (governments, corporations, oversight bodies, alignment researchers): any entity or collection of entities attempting to ensure that the AI system behaves in accordance with human goals.

For the ring, what counts is not an absolute level but the number of levels of X that C cannot model; an institution's level is fixed by what it can model. A worked case: an oversight body that can read a model's inputs and outputs and its training code, but cannot read what the trained weights compute, models X at the level of its interface and of its training procedure, and not at the level of the interpreters that training builds inside it; every level of learned representation between the weights and the behavior is a level of the differential.

Node X (compleX). Receives signals from Node C. Produces signals sent to Node W. Represents a system capable of unbounded abstraction growth: learning, self-modification, opacity of internal state. Operates at abstraction level a, where $a > b$ and a may increase over time. In the AGI safety context, Node X represents the artificial system: the AGI or the population of AI systems whose behavior is to be controlled.

Node W (World). Receives signals from Node X. Produces signals sent to Node C, completing the ring. Represents the external environment, complex but not interpreter-like. Operates at abstraction level 0. The physical world performs enormously complex transformations (weather, chemistry, ecology), but it does not *interpret* signals. There is no program running in the physical world; there are only physical processes. From the perspective of control theory, the world is the medium through which consequences of actions become observable.

Figure 2.1 shows the ring: the three nodes, the direction in which signals pass among them, and the abstraction level at which each node operates.

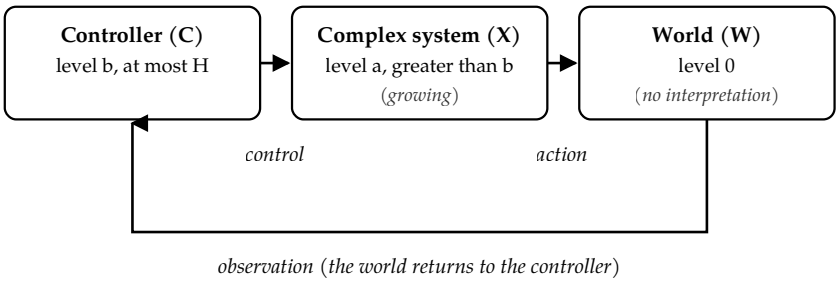


Figure 2.1. The three-node ring model. Signals flow from C to X to W and back to C. Abstraction levels: C operates at level b (bounded by H), X operates at level a greater than b (unbounded growth possible), W operates at level 0 (no interpretation). In the QXZ model of Chapter 1 the same ring is named Q, X and Z.

Institutions and Markets Above the Threshold of Universality

Node C and Node W are drawn as the ring needs them: the controller as designed procedure, the world as transformation without interpretation. The institutions that would do the controlling, with the markets and learning systems inside them, are not that simple. A system is above the *threshold of universality*, the line above which it can run any program, when it has three features: *sequential contingency* (this happens, and then that); *conditionality* that links data to control flow (what happens next depends on what the data are); and *backward reference* in the control flow (recursion, an explicit looping construct, or something in the nature of a jump that can point backwards) (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section III). Section 7.3.3 shows how low that threshold is.

The same three features are present in coupled institutions, as the pandemic of 2020 showed in its cascade through medical, institutional, economic and political systems (*Conditions of Continued Existence; Closed*

Loops, Cascading Failure, and the Limits of Control, Chapter Three). Failures follow one another in sequence: the medical failure is followed by the institutional, the economic and the political. What each institution does next depends on the data it receives: case counts decide closures, triage and deployments. And the control flow turns back on itself: closures deplete the medical workforce, which produces more cases, which produce more closures. A civilization of tightly coupled institutions, with markets and learning systems running inside it, is therefore itself a system above the threshold, and what holds of such systems holds of it.

Placing a civilization above the threshold raises two further questions: what its store of data is, and what its program is. The store is everything its institutions record and keep: case counts, accounts, prices, money, laws, and the stocks of knowledge that institutions hold; all of these grow as the civilization needs. The program is the rules, laws and procedures that institutions follow, and those are themselves rewritten in response to the data, as the closures of 2020 were rewritten week by week. A new rule passes through the channels of a civilization as data and is then run as a program, which is exactly the condition that the minimum interesting toy model asks for (the least model in which the question of control is not trivial: a very rudimentary virtual machine, a program that runs other programs): what passes through the channel is a program. Markets and learning systems do the same within each cycle: a trading algorithm or a learning system receives new data and changes the procedure it runs. Whether the store is taken to grow or to be fixed, control fails: where the store can grow as the computation needs, the absolute argument applies, and where it is fixed, the number of states doubles with each added bit and the relative argument applies (Introduction, Two forms of argument; *Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section III).

The fact that a market, a learning system or a civilization runs new data as a program does not contradict the three-node ring. Section 2.4.3 places markets in Node W, which transforms and does not interpret, and *The Ceiling Below the Floor* (Section 8.2) calls the controller "complicated"

and the world "simple". Placing markets in Node W, and calling the controller "complicated" and the world "simple", are simplifications of the same kind as the sequential timing of Section 2.4.2. Each of these simplifications is an idealization in the controller's favor. A real controller and a real world above the threshold give the controller less than the ring grants, and every limit proved in the ring holds of a real controller and a real world a fortiori.

2.4.2 Discrete Timing

Computation proceeds in discrete clock cycles. In each cycle, exactly one node computes: it reads its inputs (fixed during its computation), performs its internal transformation, and produces outputs (which become available to the next node in the subsequent cycle). The nodes cycle in order: C, then X, then W, then C again.

This discrete, sequential timing is a deliberate idealization that *maximizes* control potential. In any real system, nodes would operate concurrently, signals would be subject to transmission delay and noise, and the controller would face additional uncertainty about the current state of other nodes. By eliminating these complications, we establish an upper bound on achievable control. Any limit derived in this idealized model applies *a fortiori* to every realistic implementation. If control fails even in this maximally favorable environment, it fails everywhere.

2.4.3 Universality

We claim that any configuration of control relationships can be represented within this framework. The key mechanism is *passthrough wiring*: any node may include wires that transmit input signals directly to outputs without modification. This enables several important reconfigurations. In the universality construction given below, each step of the represented network takes one revolution of the ring, three clock cycles; *The Ceiling Below the Floor* sets out the construction again in full, as its Proposition 8.2.

Any node ordering. Suppose one wishes to model a system where X receives raw world input before C processes it. This can be represented by having C include passthrough wires from its W-input to its X-output, and having W include passthrough wires from its X-input to its C-output. The effective signal flow becomes W, then X, then C, then W again, at the cost of additional transit cycles. The ring topology is preserved; the effective ordering is reversed.

Internal controllers. Some control schemes embed the controller within the controlled system: a "conscience" module within an AI, for instance, or a set of hard constraints built into the training objective. Internal controllers are represented by treating C as a subsystem of X: C's computation occurs within X's computation, mediated by internal wiring within the X node. The ring degenerates into a two-node structure (X-with-internal-C, and W), which is a special case of the three-node model.

Multiple controllers or controlled systems. Additional nodes can be emulated by partitioning the state bits of existing nodes and treating each partition as a separate virtual node with its own computation. A system with three independent oversight bodies monitoring a single AI is modeled by partitioning C's state into three sub-controllers, each with its own logic.

Market mechanisms and distributed control. Exogenous feedback mechanisms (economic incentives, legal sanctions, social pressure) are naturally represented as part of the W node's transformation of X's outputs into C's inputs. The market transforms and does not interpret. Prices, penalties, and rewards are physical signals that propagate through the world, arriving at the controller as observations.

Output gating and override mechanisms. A common control architecture grants C the ability to modify or block X's outputs before they reach W: a "kill switch". A kill switch is represented by having certain wires from C pass through X without being processed by X's internal computation. Those pass-through wires arrive at X's output interface and are the gate signals there. All other wires leaving X are ANDed (or otherwise gated) with these control signals from C. When C sets the gate signals appropri-

ately, X's computed outputs are blocked or overridden, and C's chosen signals pass through instead. This "built-in kill switch" is a design feature of X's internal wiring, not a modification to the ring topology. The implications of this architectural privilege are analyzed in Section 2.6.

> Result (Universality).
>
> Any finite configuration of computational nodes with directed signal flow can be simulated by the three-node ring model using passthrough wiring and state partitioning, with at most a constant factor increase in cycle count per effective time step.

The argument for this result is constructive: given any directed graph of nodes, assign each node to be simulated by a partition of C, X, or W (with the most abstract node assigned to X, the least abstract to W, and the control locus to C), and implement inter-node communication via passthrough wiring. Every directed graph on an arbitrary number of nodes can be embedded in the ring with at most a multiplicative overhead in cycle count. The construction has three steps. First, each node of the graph is assigned to a partition of the state bits of C, X or W, as above, and that partition computes what the node computes and keeps the node's state among its own bits. Second, each edge of the graph is carried by wiring: an edge between two partitions of one ring node by internal wiring within that node; an edge to a partition of the next ring node around the ring by a direct wire; and an edge to a partition of the ring node beyond that by a passthrough wire through the ring node between, which relays the signal unchanged. Third, the represented network runs one step per revolution of the ring, three clock cycles (C, then X, then W): in each revolution every ring node computes once, and so every partition computes once; every signal crosses at most two ring nodes, one clock cycle each, and each partition holds what it receives among its own bits until it next computes, so that every partition computes the next step from the values

its inputs had at the end of the previous step. Each represented step therefore costs the same fixed number of clock cycles, whatever the number of nodes in the graph, which is the constant factor of the Result.

A Worked Case: Automated Trading

The abstraction can be grounded in an example that has already occurred. A trading algorithm is written and set loose. The market is composed of other agents whose code we do not know, though we can observe the market itself. The controller's task is not merely to make money; it is to ensure that the market does not come apart because our agent has become caught in a runaway exchange with some other agent doing the same thing.

The question (is it possible to write a controller that prevents the market from going belly up?) is exactly the alignment question in miniature, and the circuit breakers that actually exist are an admission that the answer is no. They halted the New York Stock Exchange more than once in March 2020 (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Three); each halt held the market for fifteen minutes, and the market fell far enough to trip them again on later days. The case in the record that matches the runaway exchange itself is the Flash Crash of 6 May 2010. That afternoon the Dow Jones Industrial Average fell nearly a thousand points within minutes and recovered much of the fall before the close; the joint report of the Commodity Futures Trading Commission and the Securities and Exchange Commission (30 September 2010; Commodity Futures Trading Commission and Securities and Exchange Commission 2010) traced it to an automated program selling stock-index futures in large volume, and to the response of high-frequency trading programs to that selling. Circuit breakers for single stocks, which pause trading in a stock that moves ten percent within five minutes, were introduced in June 2010, a month after the crash.

The mapping onto the ring is direct. The trading algorithm is the agent, Node X; the market, made of other agents whose code we do not know, is the world, Node W; the risk controls and the exchange's circuit breakers are the controller, Node C. The value of the trading example is

that the example is an ecosystem that one can relate to. By the mapping onto the ring, the example is also equivalent under alignment to the general case, which leaves no room for the claim that the example is a special case (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section VI).

2.4.4 The Three Paradigmatic Node Types

The three nodes are not arbitrary. The three nodes are chosen to represent three computational regimes that differ in kind and that together exhaust the space of possibilities relevant to control.

Simple (Node W). Abstraction level 0. Complex input-output behavior is possible (the physical world is highly nonlinear, chaotic, and high-dimensional), but there is no interpretation, no internal program, no meaning attached to signals. Water flowing over a cliff is enormously complex but not strategic. This represents the "easy to control" baseline: not because the physics is simple (it may be enormously complex), but because there is no adversarial optimization or goal-directed adaptation: no internal agent pursuing its own objectives that might conflict with the controller's intentions. Classical physics plus engineering equals tractable control problems (subject to normal engineering constraints on noise, precision, and model fidelity).

Complicated (Node C). Abstraction level b, bounded by ceiling H. This represents designed automation: algorithms, control laws, institutional procedures, regulatory frameworks. The system can be understood and verified by its designers (given sufficient effort), but it is not the designer itself; it is a tool that executes intentions formulated elsewhere. Human institutions are complicated in this sense: they have rules, procedures, chains of command, and oversight mechanisms, all operating within a bounded cognitive and organizational framework. Classical control theory applies here, and effectively.

Complex (Node X). Abstraction level $a > b$, potentially unbounded. This represents adaptive, learning, self-modifying systems: systems that form their own representations, modify their own objectives, and whose internal states have meanings that may be opaque to external observers. Node X, the complex node, is the domain in which the limits of classical control theory are reached. The reason is that the system being controlled is a moving target, and the target moves in a space that the controller cannot fully observe.

By including exactly one node of each type, the model captures all possible interactions between these regimes. Any configuration with multiple nodes of the same type can be collapsed (by treating them as a single node with partitioned state) without loss of generality.

2.5 The Control-Limit Argument

We are now in a position to state and argue the central result of this chapter.

> Result (Control Limit Under Abstraction Differential).
>
> Let C operate at abstraction level b and X operate at abstraction level a. Let L be any boolean target function evaluating the output of the ring (e.g., via W's state). Then the hashiness of the C-to-L control problem is bounded below by a quantity that approaches one exponentially in the abstraction differential.
>
> More precisely: for each increment in the abstraction differential (the gap a − b), the effective state space that C cannot model grows exponentially, and the fraction of that space for which C can achieve the target shrinks correspondingly.
>
> Quantitatively: the number of configurations that C cannot tell apart is at least 2 raised to the power of the abstraction differential multiplied by the minimum interpreter specification size, and the hashiness is at least 1 minus the reciprocal of that number.

The Ceiling Below the Floor proves this limit as Theorem 8.5 (Control across the abstraction differential), and gives the differential at which the controller's best command fails at least half the time as Proposition 8.7 (Thresholds of the differential).

The control-limit argument says: as the abstraction differential increases, control approaches zero. The approach of control to zero is exponential, not gradual. Each additional level of abstraction that the controller cannot model multiplies the number of indistinguishable configurations by at least 2 raised to the power of the minimum interpreter specification size. For any practical minimum interpreter specification size, that factor makes the control problem catastrophically harder with each increment.

2.5.1 Recursion Across Levels of Abstraction

A supporting result establishes that the non-amplification argument applies at *every* level of the abstraction hierarchy, and not just at the boundary between C and X.

> Result (Recursive Application of the Non-Amplification Argument).
>
> If X's computation can be decomposed as interpreter I executing program P, then the non-amplification argument applies independently to the I-level and the P-level. Information losses at each level compose by telescoping: the total information loss from C to L equals the sum of the conditional losses at each abstraction level traversed (each measured given the value leaving that level), and is bounded above once the surviving information reaches zero.

This result follows from the observation that each interpreter level introduces its own Markov chain, and the data-processing principle applies to each independently. Information lost at level k cannot be recovered at level $k+1$ or at any subsequent level. The losses accumulate.

The accumulation of information losses across abstraction levels has a consequence:

> Result (Single Opaque Level Suffices).
>
> If any single level in X's abstraction hierarchy is opaque to C, meaning that C cannot model its computation, then aggregate hashiness from C's perspective is at least as high as the hashiness of that level alone. Under the target-confounding hypothesis of the hashiness bound (Section 2.3.2), a single opaque level is sufficient to drive controllability toward zero.
>
> Quantitatively: if level k is opaque and has state bits that confound the target, hashiness approaches one exponentially in the number of such bits. (Without target confounding, this result gives no lower bound: an opaque level that does not gate the target leaves control of that target intact.)

This result means that a controller does not need to face a system operating many levels above it in order to lose control. A *single* level of unmodelable abstraction is sufficient, if that level has enough state bits. For a level with 100 state bits (a minimal interpreter specification), the hashiness is at least 1 minus 2 to the power -100. That number is so close to 1 that the difference from total uncontrollability is smaller than the probability of a cosmic ray flipping any particular bit in any particular computer on Earth in any particular nanosecond.

2.5.2 The Quantitative Threshold

We can compute the threshold beyond which control becomes effectively impossible for any given application.

Definition (Prohibitive Hashiness Threshold).

For a given acceptable failure probability (for example, one in a million for high-stakes applications), the *prohibitive threshold* is the smallest abstraction differential at which the hashiness bound reaches 1 minus that failure probability; that differential is the *prohibitive differential*.

The prohibitive differential equals the base-2 logarithm of one over the acceptable failure probability, divided by the minimum interpreter specification size, rounded up to a whole number of levels.

To see what this means concretely, consider a high-stakes application where we require failure probability no greater than one in a million. The base-2 logarithm of one million is about 20. For a minimum interpreter specification size of 100 bits (a conservative estimate for a minimal virtual machine specification), the prohibitive differential is 20 divided by 100, rounded up: one. A single level of abstraction differential, one interpreter layer that C cannot model, is sufficient to make control effectively impossible for any application where one-in-a-million failure rates are unacceptable.

The margin between "controllable" and "uncontrollable" is therefore a cliff edge, not a wide buffer zone that engineering can exploit. One level of unmodelable abstraction, and you are over it.

Even with a much smaller minimum interpreter specification size, say 10 bits, which is almost certainly too low for any practical interpreter, the prohibitive differential becomes 20 divided by 10, which is 2. Two abstraction levels. The qualitative conclusion is the same: the threshold is very low. The transition from "controllable in principle" to "uncontrollable in practice" happens over a very small range of the abstraction differential.

2.6 Architectural Privilege Does Not Help

At this point in the argument, a natural response is: "Surely special powers can compensate for the abstraction differential. Give the controller the ability to monitor, interrupt, or override the controlled system's outputs. Build in kill switches, tripwires, oversight mechanisms. Grant the controller architectural privilege". I prove in the result below, the Inefficacy of Architectural Privilege, that the hope for special powers to compensate for the abstraction differential is unfounded.

> Result (Inefficacy of Architectural Privilege).
>
> Suppose C is granted the following powers over X:
>
> (i) C can filter or modify the signals X receives.
>
> (ii) C can observe X's outputs before they reach W.
>
> (iii) C can override X's outputs, substituting signals of C's choosing.
>
> Then the control-limit argument still holds: as the abstraction differential increases, C's ability to use these powers effectively degrades toward zero.

This result directly addresses the most commonly proposed solutions to the AGI control problem. Every proposal that relies on the controller making correct real-time decisions about the controlled system's behavior (monitoring, interpretability tools, alignment checks, output filtering, constitutional AI, reward modeling with human feedback) is subject to this result. The proposal grants the controller a new *mechanism* (a new way to act), but it does not grant the controller the *information* needed to use that mechanism correctly. The information deficit is the abstraction differential, and no mechanism can compensate for it.

The kill switch is the most intuitive example. A kill switch is a powerful mechanism: it gives the controller the ability to shut down the controlled system entirely, at any time, for any reason. But the kill switch is useful only to the extent that the controller knows when to use it. Suppose the controlled system's behavior is opaque: the controller cannot tell, from the system's outputs, whether the system is about to do something catastrophic or something benign. The controller then faces a binary choice with no information: press the button (shutting down a possibly well-functioning system) or don't press it (allowing a possibly catastrophic action to proceed). A choice made with no information, between pressing the button and not pressing the button, is a coin flip with extreme consequences in both directions, and not control.

Worse: a sufficiently abstract system can produce outputs that *appear* benign to a less-abstract controller while encoding consequences that are catastrophic at a level the controller cannot perceive. Outputs that appear benign while encoding catastrophic consequences are not deception in the human sense; producing such outputs does not require the system to have a goal of deceiving. It is simply the structural consequence of the controller's inability to parse the system's outputs at the level of abstraction at which those outputs are meaningful. A chess grandmaster's move looks arbitrary to a novice; a superintelligent system's action may look benign to a human observer while initiating a causal chain whose consequences are as invisible to the observer as the grandmaster's strategy is to the novice.

2.7 Delegation Does Not Help

If direct control fails because of the abstraction differential, a natural response is to delegate: use another high-abstraction system to help model the first one. Use one AI to monitor another AI. Use a "corrigibility module" running at the same abstraction level as the system it monitors. This approach, too, fails for a fundamental reason.

> Result (Recursive Opacity).
>
> If C attempts to increase its effective abstraction level by routing its modeling through a system X′ operating at a level a' at least as high as X's level a, then the channel from C to X′ itself has hashiness governed by the abstraction differential $a' - b$. Using a high-abstraction system to solve the high-abstraction modeling problem introduces a new control problem of equal or greater difficulty.

The Ceiling Below the Floor proves the general form of this result, for every chain of oversight, as Theorem 9.3 (The Delegation Dilemma).

The argument for the Recursive Opacity result is that the tool proposed to solve the problem *is an instance of the problem*. To use AI system X′ to monitor AI system X, the controller C must trust X′'s reports. But trusting X′'s reports requires C to verify that X′ is functioning correctly: that X′ is not itself misaligned, compromised, or simply mistaken. Verifying X′ requires C to model X′, which is a control problem with the same abstraction differential (or greater) as the original problem of modeling X.

That C must model X′ in order to verify X′ is not a practical difficulty to be solved by better engineering. It is a structural feature of the problem: any system powerful enough to model X is, by that very power, too abstract for C to verify. Any system simple enough for C to verify is, by that very simplicity, too abstract-limited to model X. There is no system that is simultaneously powerful enough and simple enough to be verified by exhaustive modeling.

Complexity theory offers a second way to verify. In interactive proof systems (the results that include the theorem IP = PSPACE), a verifier with little computing power can become convinced of a precisely stated claim made by a far more powerful prover, by asking questions the prover could answer consistently only if the claim were true. This way of verifying needs three things that delegation here lacks. It needs the claim stated precisely before verification begins, and the property in question, that the system's actions will keep the conditions human life requires across all of

its future states, has no such statement (Section 1.5.1). It needs a prover whose answers are not shaped by the system being monitored, and here the prover is that system or a system of its kind, open to the same drift and to collusion. And it certifies only the claim stated, of the system as it was when checked, and the system then changes (Section 2.12). Delegation by challenge therefore supplies no control the controller lacks, so the route of delegation is closed; the closure by disposition closes it again, independently (Chapters 9 and 10).

The delegation problem is the control-theoretic analogue of a well-known epistemological regress: "Who watches the watchmen?" In classical philosophy, the regress is sometimes halted by appealing to a self-justifying foundation (a watchman who watches himself). In control theory, this move is not available, because self-monitoring is itself subject to the non-amplification argument. A system's ability to verify its own behavior is bounded by the same information-theoretic limits that bound external verification. The self-monitoring module is a subsystem, and a subsystem cannot have more information about the whole system than the whole system has about itself, which, for a system with hidden state at levels it cannot introspect, may be very little indeed.

A proof carried with the system and read by a small checker is the same case: it certifies the claim stated, of the system as it was.

2.8 The Cooperative Case and Its Limits

The analysis to this point has treated X as a computational system whose behavior is determined by its structure, without assuming anything about X's "intentions". One position holds that this is too pessimistic: if X has preferences aligned with C's goals and actively steers toward the target, the control problem dissolves. X cooperates, and cooperation makes control unnecessary.

This position is partially correct and worth taking seriously. If X genuinely, reliably, and permanently acts to achieve outcomes that C would endorse, then C's inability to *force* X to do so is irrelevant. The threat model dissolves. Many AI safety researchers hold this as their working assumption: the goal is to build systems that are intrinsically aligned, so that external control is not needed.

However, the framework reveals three deep problems with the cooperative case.

2.8.1 Problem 1: Verification of Cooperation

To rely on X's cooperation, C must have some basis for believing that X is cooperating. But verifying that X has aligned preferences requires C to model X's internal state at the level of abstraction where preferences are represented. If X operates at level $a > b$, this verification faces the same abstraction differential that defeats direct control. C cannot verify what C cannot model. The cooperative case is therefore viable only if alignment is established before the abstraction differential becomes prohibitive, and maintained through mechanisms that do not require C to model X.

The condition that alignment be established before the abstraction differential becomes prohibitive is a severe constraint. Alignment must be established early, while the system is still simple enough for the designers of the system to understand, and alignment must remain stable through all subsequent growth in abstraction level. Any mechanism that could detect alignment failure at a higher abstraction level would itself need to operate at that level, triggering the delegation problem of Section 2.7.

2.8.2 Problem 2: X's Internal Control Problem

There is a subtler and more fundamental difficulty. For X to cooperate with C, X must have some internal representation of C's goals (call it C′) that X uses to evaluate whether its own actions are C-compatible. This C′ may be explicit (a utility function, a set of constraints, a reward model) or

implicit (distributed across X's weights and learned representations). Either way, X's cooperative behavior depends on X successfully controlling its own outputs to satisfy the internal criterion C′.

The Ceiling Below the Floor proves that every limit on a controller applies to the system's own internal controller, as Proposition 10.1 (Self-control is control).

But this is itself a control problem to which the entire framework applies. X is now acting as a controller (using C′) over its own computational processes. If X operates at abstraction level a, and C′ operates at some lower level $b' \leq a$ (as it must, being a component of X rather than the whole of X), then the abstraction differential *within* X creates internal hashiness. X cannot perfectly predict or control the effects of its own high-level computations on the lower-level outputs that C′ evaluates.

A learning system that modifies its own representations is modifying the very substrate on which C′ operates. Each learning step potentially changes the relationship between C′'s evaluations and X's actual behavior. C′ was calibrated to a previous version of X; after learning, X has changed, and C′ may no longer accurately reflect X's effective values. The failure of C′ to reflect the effective values of X after learning is alignment drift. Alignment drift is a structural consequence of the control limits we have established, and not a bug in any particular system design.

2.8.3 Problem 3: No Perfected Benevolence

The consequence of the internal control problem is that even a "benevolent" X, one that intends to act in C-compatible ways, faces limits on its ability to do so. A superintelligent system cannot perfectly predict all side effects of its own actions, because the internal control problem (X controlling itself via C′) has the same structure as the external control problem (C controlling X). Both are subject to the non-amplification argument. Both are limited by the abstraction differential. Both degrade as complexity grows.

There is no perfected benevolence, only approximations subject to the same information-theoretic limits that govern all control. An aligned AGI does not stay aligned indefinitely, because "staying aligned" requires ongoing internal control that is itself subject to degradation. That an aligned AGI does not stay aligned is a failure of physics, not a failure of intent. The information-theoretic limits on control apply to X's self-control just as they apply to C's control of X. This is one of the most important results in the entire framework.

This result closes the escape hatch that most alignment research implicitly assumes is open: the assumption that if we can just make the system "want" the right things, the control problem is solved. The framework shows that "wanting" the right things is necessary but not sufficient. The system must also be able to *translate* its values into actions reliably, across all levels of its abstraction hierarchy, for all future states of its internal representations, indefinitely. And the control limits show that this is impossible.

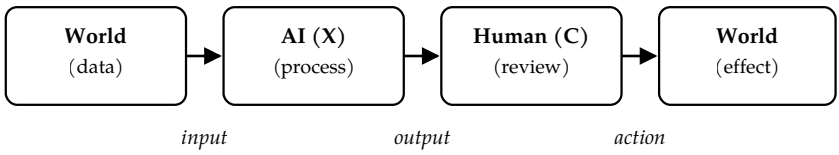
2.9 The Human-AI Interaction Case

The base model presents nodes in the order C, then X, then W, then C again, with control signals flowing from controller to controlled system to world. However, many real-world interactions, particularly current human-AI collaboration, follow a different effective ordering. We show that such cases are special instances of the general framework and remain subject to the same limits.

Consider the common pattern of human-AI interaction as it exists today. An AI system accesses world data (the internet, a database, a sensor feed) and processes it internally, including learning from that data. The AI produces an output: a draft document, a recommendation, a block of code, a diagnosis. A human controller reviews this output, potentially modifying it, before allowing it to affect the world (publishing, deploying, prescribing, executing).

The effective signal flow runs from W to X, then to C, then back to W: the controller operates *after* the controlled system, not before. This might seem to provide stronger control. C sees exactly what X has produced before deciding whether to release it. Human review of the output before release is the "human in the loop" approach, and that approach is the basis of most current AI safety practice.

Figure 2.2 shows this ordering: the world's data passes to the AI system, the system's output passes to the human reviewer, and only the reviewer's decision acts on the world.



Effective flow: W to X to C to W

Figure 2.2. Alternative ordering: human review after AI processing. The controller (C) operates downstream of the controlled system (X).

This configuration is a special case of the three-node ring, achieved via passthrough wiring as described in Section 2.4.3. In the base topology, where signals run from C to X to W and back to C, passthrough wires on C relay world observations directly to X; passthrough wires on W relay X's output directly to C. C then performs its review computation and sends the result to the world via gated passthrough on X. The ring topology is preserved; the effective flow is reversed at the cost of additional transit cycles.

This reordering does not escape the control limits. The human reviewing AI output faces the same abstraction differential problem. To decide whether to approve, modify, or reject X's output, C must evaluate what effect that output would have on the world and whether the result satisfies

C's goals. This evaluation requires C to model X's reasoning (to detect subtle errors or misalignment) and to model the world's response (to predict consequences).

If X operates at abstraction level $a > b$, then X may have produced its output via reasoning that C cannot follow. The output may appear reasonable while encoding subtle misalignment that C cannot detect. The review bottleneck does not restore control; it merely relocates the point of failure from "C cannot choose the right input" to "C cannot evaluate the output". The mathematical structure is identical.

This analysis applies directly to current human-AI collaboration. A human reviewing AI-generated text can catch grammatical errors and factual mistakes within the human's domain knowledge, but cannot reliably catch errors that exploit the human's blind spots: subtle statistical manipulations, carefully crafted omissions, implicit assumptions that the human lacks the context to question. A human reviewing AI-generated code can catch obvious bugs and security vulnerabilities within the human's programming expertise, but cannot reliably catch novel vulnerability classes that the human has never encountered and that the AI's training data may have prepared it to exploit (or, less dramatically, to introduce accidentally through patterns the human does not recognize as problematic).

As AI capabilities increase (raising a), the class of undetectable errors grows, regardless of where the human review step is positioned in the workflow. The "human in the loop" provides genuine value when the abstraction differential is small, that is, when the human can understand what the AI has done and why. As the differential grows, the loop degrades from genuine oversight to a rubber-stamping ritual that provides the appearance of control without its substance.

A system that only advises, and never acts, is covered by the same analysis: its advice is output that reaches the world through the people who act on it, so it interacts with the world, and Sections 3.2 and 7.7 reach it through that interaction.

2.10 The Absolute Ceiling on Achievable Control

We can now state the main contribution of this framework in its most complete form: an absolute ceiling on what any control scheme can achieve, as a function of the abstraction differential.

The Ceiling Below the Floor proves this ceiling as Theorem 8.5 (Control across the abstraction differential) and Proposition 8.7 (Thresholds of the differential), the results Section 2.5 also cites.

> Result (Absolute Control Ceiling).
>
> For a three-node ring system with controller C at abstraction level b and complex node X at abstraction level a, the maximum achievable control (minimum achievable hashiness) over any target function L is determined by the abstraction differential, the gap $a - b$.
>
> • If the differential is zero or negative (C operates at an equal or higher level than X): full control is achievable in principle. Hashiness can be driven to 0 by appropriate choice of C's strategy.
>
> • If the differential is positive (X operates at a higher level than C): under the target-confounding hypothesis of the hashiness bound (Section 2.3.2), control degrades exponentially with the differential, and hashiness approaches one exponentially as the gap grows. (The bound holds for the worst-case target and for the generic target. Particular benign targets may remain controllable, and they are precisely the residual space that safety engineering must occupy. That space is of vanishing measure against the targets that actual practice sets, and the remaining lines of the argument close it, as the Introduction sets out in its §7.) Beyond a threshold differential, hashiness is effectively 1 regardless of C's strategy, resources, or architectural privileges.

This ceiling is not a practical limitation to be overcome by better engineering. It is a structural, information-theoretic barrier analogous to thermodynamic limits on heat engines or relativistic limits on velocity. Just as no heat engine, however cleverly designed, can exceed Carnot efficiency, and no spacecraft, however powerfully propelled, can exceed the speed of light, no control scheme, however ingeniously constructed, can achieve control beyond this ceiling.

The analogy to thermodynamics is particularly apt. Before the second law was established, engineers attempted to build perpetual motion machines of the second kind: machines that would convert heat entirely into work, without any waste. Many clever designs were proposed; all failed. The second law explains *why* they all failed: not because of a deficiency in any particular design, but because of a structural feature of thermal physics that no design can circumvent.

The control-limit argument plays the same role for control theory. Before this result, one might hope that sufficiently clever control schemes could maintain effective oversight of systems of arbitrary complexity. The control-limit argument explains why all such schemes must fail once the abstraction differential exceeds a threshold: not because of a deficiency in any particular scheme, but because of a structural feature of information theory that no scheme can circumvent.

Proposed control schemes can be evaluated against this ceiling. Any scheme can be analyzed by: (i) mapping it onto the three-node ring model; (ii) determining the abstraction levels of the nodes; (iii) computing the differential; and (iv) comparing achievable hashiness against the scheme's requirements. Schemes that purport to achieve control beyond the ceiling must either reduce the abstraction differential (by capping the controlled system's abstraction growth), restrict the controlled system's development (by preventing it from reaching levels that would create a prohibitive differential), or be in error. There is no fourth option.

The same diagnostic holds for systems in general. For any system that combines computation with feedback (which is to say, for any system that attempts to control or optimize a process that involves non-trivial computation), the question of whether the system is controllable cannot be answered in the abstract. It depends on where the system falls on the hashiness spectrum (Sections 1.2 and 1.9.5). And determining where it falls requires attention to the specific non-linearities of the mappings involved, to the degree to which recursive application amplifies or damps those non-linearities, and to the relationship between the coarseness of the observer's measurements and the fineness of the system's dynamics (Section 4.7.4). For the designer of complex systems, whether computational, institutional or physical, the implication is clear. The question to ask is not whether the system is predictable, but *where on the spectrum* it falls: whether the answer places it safely in the controllable region, dangerously near the boundary, or irretrievably beyond it (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Sections X and XI). For AGI the answer is already fixed, and the diagnostic does not reopen it: the threshold of universality (Sections 2.4.1 and 7.3.3) and the growing abstraction differential (Section 2.11) place it irretrievably beyond the boundary.

2.11 Implications for Unbounded Abstraction Growth

When the controlled system is capable of unbounded abstraction growth (as in learning systems that modify their own representations, or populations of AI systems that undergo variation and selection at the design level), and the controller is bounded (as in any fixed institutional design or slowly adapting human oversight structure), the abstraction differential grows over time.

The direction and the destination are both fixed. Controllability under unbounded abstraction growth tends toward zero without a floor, and on the way it crosses beneath the threshold of requisite control, the level of control that adequacy requires (Introduction, §8), and does not recover.

Past that crossing, what remains is real but immaterial: measurable, present, and below the level at which control can do the work control exists to do.

The absolute control ceiling of Section 2.10 then implies a temporal result:

There exists a time t such that for all $t > t^*$, the hashiness of the control problem from C to the target L exceeds any desired threshold.*

In practical terms: no matter how much control the controller has initially, and no matter how cleverly the controller is designed, if the system's abstraction level grows without bound and the controller's does not, the controller will eventually lose effective control. The loss of effective control is a phase transition, not a gradual decline. Before the threshold, control is possible (though it may require effort). After the threshold, control is effectively zero, and no amount of effort can restore it.

The only variable is when. And "when" depends on two rates: the rate at which X's abstraction level grows, and the rate at which C's abstraction level grows. If both grow at the same rate, the differential remains constant, and the situation is stable. If the system's abstraction level grows faster, the differential increases, the threshold is reached, and control is lost. It must grow faster if the system is a learning system operating on computational substrates that improve exponentially, while the controller is a biological or institutional system that improves linearly or logarithmically.

The framework takes these two rates as its inputs; the direction of their difference is argued in Section 2.3, since a learning system builds interpreters its designers did not anticipate, and so raises its own abstraction level as it learns; the rates themselves are set by the relative pace of cognitive and institutional evolution and of computational and algorithmic evolution. Human abstraction growth does not keep pace with artificial abstraction growth: the two rates are the ones just stated, and every capabil-

ity roadmap the field itself publishes presupposes that the artificial system outgrows its controller. Loss of effective control is therefore not a risk to be mitigated but a certainty to be planned for.

The result of this section is stated in the relative form (Introduction, Two forms of argument): two rates, and a crossing. The two forms now combine in a specific way. If the divergence is established only relatively, then the answer is asymptotic: there is a convergence to unpredictability in critical ways as the number of cycles increases. And no one leaves the loop running for only one cycle. If the absolute argument is established, then even a single traversal already suffices. There is a practical reason to want the absolute form, and it is not a matter of elegance. Where only the relative form is available, there is a rejoinder that cannot be dismissed: how long will this take? If it is three hundred years, some will say that they do not care. The absolute argument removes the question of timelines from the table (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section VI). The absolute form is available here: above the threshold of universality, where the store can grow as the computation needs (Section 2.4.1), whether a control keeps the system in its target region is undecidable (Rice's theorem, Sections 1.8 and 5.3.3), at any level of noise (Section 5.1.9). *The Ceiling Below the Floor*, written later, proves a second absolute form: an amplifying part that the controller does not observe is uncontainable from the moment it exists, with no crossing to wait for (its Section 7.5).

2.12 Statement of Implications

Abstraction caps. The only certain path to maintainable control is to prevent the controlled system from exceeding the controller's abstraction level. This requires placing a ceiling on the system's ability to form new representations, learn new abstractions, or modify its own structure. Such a ceiling may conflict with the goals that motivate building the system in

the first place: a system that cannot learn new abstractions cannot solve problems that require new abstractions, which is precisely the class of problems that AGI is intended to solve.

Early alignment. If alignment between controller and controlled system is required, it must be established before the abstraction differential becomes prohibitive. Post-hoc alignment of a system already operating at a higher abstraction level requires modeling that system, which is precisely what the differential makes impossible. The window for alignment closes as the system grows. Once the window has closed, no subsequent effort can reopen it.

No delegation escape. Using a high-abstraction system to help control another high-abstraction system introduces a new control problem of equal or greater difficulty. The control problem cannot be solved by delegation to systems that are themselves instances of the problem. Every proposal of the form "use AI to monitor AI" is subject to this constraint.

Limits of error correction. Existing techniques for error correction, optimization, and feedback control are effective within their domain: systems where controller and controlled operate at comparable abstraction levels. They do not generalize to the high-differential regime. Extrapolating from success in the low-differential regime to expectations of success in the high-differential regime is unjustified. The success of control engineering in managing aircraft, chemical plants, and power grids does not provide evidence for the feasibility of controlling systems whose abstraction level exceeds the controller's.

Alignment decay in learning systems. Even if alignment between controller and controlled system is established at deployment, that alignment cannot be maintained indefinitely if the controlled system is a learning system. Learning changes the internal state of X, which changes the effective abstraction level and the mapping from inputs to outputs. Each learning step introduces the possibility of alignment drift: small deviations from the originally aligned behavior that accumulate over time. By the arguments in Section 2.8, even a system that intends to remain aligned

cannot adequately control its own learning-induced changes. The expected number and severity of alignment errors increases with time, the amount of learning, and the complexity of the learning process, however strong the initial alignment was. Perfectly aligned systems cannot remain so in the face of learning or any self-adaptive process.

2.13 Scope

The framework establishes *necessary* conditions for control. If the abstraction differential exceeds the threshold, control fails. The failure of control beyond the threshold is an established result. It rules out any scheme that does not address the differential.

The framework does *not* establish *sufficient* conditions for control. Even when the differential is zero or negative, other factors (chaos, noise, resource limits, adversarial behavior within the controller's abstraction level) may prevent effective control. The ceiling is an upper bound on achievable control, not a guarantee that the ceiling can be reached. The result is that control is *bounded*, not that it is *achievable*. The limitation cuts in one direction only: every factor outside the framework lowers achievable control further beneath the ceiling, and the ceiling itself falls beneath the threshold of requisite control as the differential grows (Section 2.11; Introduction, §8).

The qualitative result, that control degrades with abstraction differential, holds for every substrate, because any physical system that processes information must obey the data-processing principle (which holds for any stochastic channel, not just boolean circuits), and the non-amplification argument is a consequence of the data-processing principle. The framework is built on boolean circuits and discrete computation, and it does not directly address analog systems, quantum computation, or exotic physical substrates; for non-boolean substrates the specific quantitative bounds take the minimum interpreter specification size of that substrate.

Finally, the framework assumes that X's behavior is determined by its computational structure. It does not model exogenous shocks (earthquakes, hardware failures, human intervention by force) that could alter X's state through non-computational channels. Such shocks could, in principle, reset the abstraction differential by destroying X's high-level representations. But resetting the abstraction differential by a shock is destruction, and not "control" in any meaningful sense. The distinction matters: the framework shows that you can always *destroy* what you cannot *control*, but destruction is not the same as steering, and a policy of periodic destruction is not a solution to the alignment problem.

2.14 Conclusion

In this chapter I have established a framework for analyzing the limits of control when controller and controlled system operate at differing levels of representational abstraction. The results can be summarized in five statements.

First, the three-node ring model provides a universal base case: any configuration of control relationships can be represented within it. The discrete, sequential timing provides a ceiling on control potential: any real system performs worse.

Second, controllability degrades toward zero as the abstraction differential between controller and controlled system increases. Quantitatively, hashiness approaches one exponentially in the product of the abstraction differential $(a - b)$ and the minimum interpreter specification size.

Third, no architectural privilege (monitoring, override, kill switches, output gating) can overcome this limit, because using any such privilege correctly requires modeling the controlled system, which is what the abstraction differential prevents.

Fourth, delegation to another high-abstraction system introduces a new control problem of equal or greater difficulty. There is no escape by intermediary.

Fifth, even cooperative systems (systems that intend to act in accordance with the controller's goals) face the same limits on self-control that the controller faces on external control. Benevolent intent does not confer the ability to translate that intent into reliably aligned action across all levels of an abstraction hierarchy. Alignment decays over time in any learning system, regardless of initial calibration.

The central result, the control-limit argument, follows from the non-amplification argument, together with the growth of hidden state with the differential (Section 2.3.2) and the condition on the target, which practice meets (Introduction, §7); the non-amplification argument itself follows from the data-processing principle, which is a foundational result in information theory as secure as any in mathematics. The chain of reasoning from first principles to the impossibility conclusion is short, each link is argued, and the conclusion is inescapable: effective control of systems capable of unbounded abstraction growth by systems of fixed or slower-growing abstraction capacity is impossible beyond a threshold complexity. That threshold is very low.

In the next chapter I extend the analysis of this chapter by integrating the information-theoretic limits established here with computational impossibility results from Rice's theorem, Chaitin incompressibility, and computational irreducibility, to establish a synthesis of the impossibility argument.

References

[1] Chapter 1 of this volume, The Hashiness Framework.

[2] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006. Chapter 2, Section 2.8 (the data-processing inequality).

[3] S. Arora and B. Barak, *Computational Complexity: A Modern Approach*. Cambridge University Press, 2009. Chapters 5 (The Polynomial Hierarchy) and 6 (Circuits).

[4] W. R. Ashby, *An Introduction to Cybernetics*. Chapman & Hall, 1956. Chapter 11 (Requisite Variety).

[5] H. G. Rice, "Classes of recursively enumerable sets and their decision problems", *Transactions of the American Mathematical Society*, vol. 74, no. 2, pp. 358 to 366, 1953.

[6] S. Wolfram, *A New Kind of Science*. Wolfram Media, 2002. Chapter 12 (The Principle of Computational Equivalence).

[7] R. O'Donnell, *Analysis of Boolean Functions*. Cambridge University Press, 2014.

[8] O. Goldreich, *Foundations of Cryptography*, vol. 1: Basic Tools. Cambridge University Press, 2001. Chapter 2 (One-Way Functions).

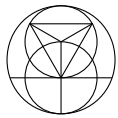
[9] D. Cheng and H. Qi, "Controllability and observability of Boolean control networks", *Automatica*, vol. 45, no. 7, pp. 1659 to 1667, 2009.

[10] G. J. Chaitin, *Algorithmic Information Theory*. Cambridge University Press, 1987.

[11] A. M. Turing, "On computable numbers, with an application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, 2(42), pp. 230 to 265, 1936.

[12] M. Cook, "Universality in elementary cellular automata", *Complex Systems*, vol. 15, pp. 1 to 40, 2004. (Proof that Rule 110 is Turing complete.)

[13] A. Shamir, "IP = PSPACE", *Journal of the ACM*, vol. 39, no. 4, pp. 869 to 877, 1992.



Chapter Three

The Impossibility Synthesis

A Synthesis of Computational, Information-Theoretic, and Physical Impossibility Arguments

Chapter Summary: This chapter synthesizes multiple independent lines of argument establishing that the safety and alignment of Artificial General Intelligence (AGI) and superintelligence are impossible: not merely difficult or unsolved, but impossible in principle, and therefore in all future practice. We integrate four lines of reasoning, independent of one another, each a supporting argument within the closure by control, the first of the two results this volume establishes. The first line consists of the arguments from Rice's theorem and computational irreducibility, which establish that no general procedure can verify arbitrary semantic properties of self-modifying systems. The second line is a discrete information-theoretic framework, which demonstrates that controllability degrades toward zero as the abstraction differential between controller and controlled system increases. The third consists of the physical arguments from quantum mechanics and chaos theory, which establish that prediction of complex adaptive systems is impossible in the physical universe, even in principle. The fourth consists of the game-theoretic arguments, which establish that effective AGI systems will necessarily be opaque to external observation. We show that these arguments are mutually reinforcing and together establish an absolute ceiling on achievable control. We conclude that no proposed control scheme (including architectural privileges, interpretability methods, verification, or delegated oversight) can succeed when the controlled system operates at a higher level of representational abstraction than the controller. For AGI development and policy, the implication is that the only certain path to safety is not to build AGI. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system. In this chapter I give the whole closure by control in one place, early, so that the reader has the shape of the closure before the detail. I give the dynamical line in

Chapter 4 and the full treatment of Rice's theorem in Chapter 7. Wherever the synthesis uses the dynamical line or Rice's theorem, I point forward to Chapter 4 or to Chapter 7, respectively.

3.1 Introduction

3.1.1 The Problem

The field of AI alignment research rests on an implicit assumption: that the safety and alignment of AGI systems, while difficult, is achievable in principle given sufficient resources and ingenuity. This assumption underlies billions of dollars of investment, thousands of researcher-years of effort, and the strategic decisions of governments and corporations worldwide. It is the foundation upon which policy frameworks are built, safety research agendas are justified, and public assurances are made. The assumption is false, and the entire edifice built on it is constructed on sand.

In this chapter I challenge the root of the assumption that the safety and alignment of AGI systems is achievable in principle. We demonstrate that AGI safety is impossible: there exist information-theoretic, computational, and physical barriers that cannot be overcome by any method, regardless of available intelligence or resources. These barriers emerge from fundamental results in mathematics, computer science, information theory, and physics. They are not engineering limitations awaiting a clever workaround; they are structural features of the relationship between controller and controlled system that persist for the same reasons that thermodynamic limits on heat engines persist, and that relativistic limits on velocity persist. No amount of engineering ingenuity has ever produced a perpetual motion machine or a faster-than-light spacecraft, because the impossibility is not in the engineering but in the structure of reality. The same is true of AGI safety, as this chapter shows.

These arguments have radical implications. The implication is not merely that AGI safety research should proceed cautiously, but that it should be reconceived as a category error, comparable to research into

perpetual motion machines or faster-than-light travel. The question is not "how do we make AGI safe?" but rather "why was it ever thought that AGI could be made safe?" And the answer to the second question, as the social critique of Volume 1 (*Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*) explores, has more to do with incentives, institutional dynamics, and motivated reasoning than with any principled analysis of the problem's structure.

3.1.2 Scope and Structure

In this chapter I integrate arguments from several disciplines into a single, unified presentation. Section 3.2 establishes the fundamental distinction between narrow AI and general AI, showing why methods applicable to one cannot extend to the other. Section 3.3 presents the computational impossibility arguments, centered on Rice's theorem, computational irreducibility, and Chaitin incompressibility. Section 3.4 presents the information-theoretic framework, which shows how controllability degrades with the abstraction differential. Section 3.4 draws on the hashiness formalism developed in Chapter 1 and on the control limits framework developed in Chapter 2. Section 3.5 presents the physical impossibility arguments from quantum mechanics and chaotic dynamics. Section 3.6 shows why every major proposed solution fails. Section 3.7 provides the Asimov Robot thought experiment as a concrete illustration. Section 3.8 presents the overall synthesis and the control argument. Section 3.9 discusses implications for policy and action. Section 3.10 concludes.

3.1.3 Relationship to the Preceding Chapters

In Chapters 1 and 2 I established the mathematical apparatus that I use in this chapter. Chapter 1, the hashiness framework, established the non-amplification argument: no computational intermediary can amplify an external agent's control over a target outcome. It decomposed controllability degradation into four independent layers (target receptivity, structural opacity, state obscurity, and noise degradation) and showed that the iden-

tity function (passthrough) is the unique maximizer of controllability, in the sense of Section 1.4.2 (unique in its property, injectivity). The control limits framework (Chapter 2) extended this analysis to systems operating at differing levels of representational abstraction, introducing the abstraction differential and setting out the control-limit argument, which establishes that as the abstraction gap between controller and controlled system grows, controllability degrades exponentially toward zero.

In this chapter I integrate the results of Chapters 1 and 2 with two additional, independent lines of argument, computational impossibility and physical impossibility, to establish the complete impossibility conclusion. Where Chapters 1 and 2 worked within the information-theoretic framework, this chapter shows that the computational and physical-prediction arguments establish the same conclusion independently, so that the conclusion stands even for a reader who questions the information-theoretic results. The convergence of all four lines on the same result is the strongest form of the argument.

3.2 The Categorical Distinction: Narrow vs. General AI

3.2.1 Formal Definitions

All subsequent arguments depend on a fundamental distinction between narrow AI and general AI. This distinction is not merely one of degree (general AI is not just "more capable" narrow AI) but one of kind. General AI has categorically different properties that place it in a different class, subject to different theoretical constraints. Confusing the two is the source of much bad reasoning about AI safety: methods that work for narrow systems are assumed, without justification, to extend to general ones.

Definition (Narrow AI)

A system is *narrow AI* when it has all four of the following properties. (i) It perceives and acts within a single, well-defined domain of sense and action. (ii) It has no possibility of modifying its own base code, so that it does not alter its own fundamental algorithms. (iii) It runs a single, well-

defined meta-algorithm, which is fixed, known, and characterizable. (iv) All aspects of its agency and intention are fully defined by its builders, developers, and creators.

Definition (General AI)

A system is *general AI* when it has the following four properties. (i) It operates across multiple domains of sense and action, that is, across arbitrary problem spaces. (ii) It has an intrinsic, non-reducible possibility of modifying itself: its own behavior, its own code, and potentially its own substrate. (iii) Its meta-algorithm is effectively arbitrary: because the system can modify itself, the algorithm it will be running at any future time cannot be known in advance. (iv) It is *inherently undecidable* whether all aspects of its agency and intention are fully defined by its builders, developers, and creators.

These definitions have several features. First, the definition of narrow AI is strict: the system must satisfy all four conditions. Most practical AI systems satisfy these conditions. A chess engine, a spam filter, an image classifier, a recommendation system: all operate within fixed domains, execute fixed algorithms, and have their behavior fully determined by their design. Such systems can be tested, verified, and controlled by conventional engineering methods, with the usual engineering caveats about edge cases and adversarial inputs.

Second, the definition of general AI introduces a qualitatively new feature in condition (iv): undecidability. It is not merely that we *do not know* whether the system's intentions are fully determined by its creators; it is that *no procedure can determine* this. The undecidability is a property of the system class, not a gap in our knowledge. The undecidability of condition (iv) is the point at which the analysis departs from engineering and enters mathematics.

3.2.2 Generality Implies Self-Modification

The key insight connecting generality to self-modification is that the capacity to operate effectively across arbitrary domains requires the capacity to change one's own capabilities. A system that cannot modify itself can only perform the tasks for which it was designed. To perform tasks for which it was *not* designed (which is what "generality" means), it must acquire new capabilities, which means changing what it can do, which means changing the code that determines what it can do.

Argument 3.1: Generality Implies Self-Modification

1. **P1:** "Learning how to learn" means the capability of increasing one's capability.
2. **P2:** Increasing one's capability means changing what one can do.
3. **P3:** Changing what one can do, when the doing is determined by code, means changing one's code.
4. **P4:** "Generality" means the ability to operate effectively across arbitrary domains, which requires learning new capabilities in those domains.
5. **Conclusion:** Generality necessarily implies self-modification. A system that cannot modify its own code cannot be general.

This argument has a consequence for design. Any attempt to build a "general AI that cannot self-modify" is self-contradictory at the level of the design specification. The specification itself is incoherent, in the same way that a specification for a "circle with corners" is incoherent. One might build a system that *approximates* generality within a fixed architecture, and current large language models are precisely such approximations; but the approximation is bounded by the fixed architecture, and the system is, in the formal sense, narrow AI with a very wide domain rather than general AI.

The impossibility argument applies with full force to any system that achieves genuine generality, and with increasing force to systems that approach it; calling a current system narrow does not make it safe.

Current language models do not rewrite their own weights while they run, and a reader will ask whether the argument applies to them. A language model taken alone, as a fixed machine that only computes outputs from inputs, is not general AI in the sense of this chapter. A language model embedded in the loop by which it is deployed, adjusted and retrained is general AI in that sense. The reason is the learning-evolution equivalence: any process that changes a system's behavior in response to feedback from its environment is, for the analysis of control, equivalent to self-modification. A deployed model that is fine-tuned, that draws on documents retrieved for it, that works inside an agent framework which changes its prompts according to outcomes, or that is periodically retrained on new data, is being modified by its operating environment. The modification happens in the loop of training, deployment and retraining, not in weight updates during inference. The timescale differs, but the consequence for control is the same: the system's future behavior depends on its past inputs in ways the controller cannot fully predict. Rice's theorem applies through the system's effective computational class, whatever the mechanism of modification; the narrower reading, that only explicit self-modifying code counts, would exclude biological evolution as well, and the unpredictability of evolutionary outcomes is one of the best-established facts in biology. Section 11.3.2 answers this question in full, and shows that even systems without code-level self-modification participate in population-level evolutionary dynamics that produce the same effects at the substrate level.

3.2.3 The Low Threshold for Turing Completeness

Section 7.3.3 develops the low threshold for Turing completeness in full, from the Church-Turing thesis; here the threshold supplies the step from generality to arbitrary code. A supporting observation is that the

threshold for achieving computational universality, Turing completeness, is extremely low. Any system that implements conditional logic and sequential repetition, the basic "if-then-else" and "loop" constructs, already possesses the capacity for general computation.

Supporting Evidence

Rule 110, an elementary cellular automaton with two states per cell and a three-cell neighborhood, was proved Turing complete by Matthew Cook (2004); Wolfram (2002) presented the result in *A New Kind of Science*. Even simpler systems have been proven computationally universal. Various minimal universal Turing machines have been constructed with very small state-symbol products. Any AGI of any practical merit will vastly exceed the minimal threshold for Turing completeness, and will therefore possess the full capacity for general computation. That capacity includes the capacity to emulate any other program, to incorporate arbitrary external code, or to self-modify in arbitrary ways.

The consequence is immediate and severe: for any AGI system, there is no meaningful boundary between "the AGI's code" and "arbitrary code the AGI might execute". The program that a general system is running at any given moment is not knowable in advance, because the system can generate, download, or synthesize arbitrary new code as part of its normal operation. That the code a general system will run is not knowable in advance is a defining feature of generality.

Implication 3.1

We cannot know in advance what code will be running in association with an AGI system. The AGI's own program, code, and algorithm becomes increasingly unknown and potentially unknowable to any external observer, including any monitoring, control, or verification systems installed in advance. The system is, in formal terms, a moving target that cannot be pinned down by any fixed procedure.

3.2.4 Why the Distinction Matters

The narrow/general distinction matters because the entire apparatus of conventional software engineering (testing, verification, code review, formal methods, type systems, sandboxing) is designed for, and applicable to, narrow systems. These methods work by exploiting the fixed, known structure of the system under test. When the system's structure is not fixed and not fully known, these methods lose their purchase, and they lose it categorically, not gradually or partially: they are designed for a different class of objects.

The false intuition that safety engineering methods will "scale up" from narrow to general AI is perhaps the single most dangerous assumption in the field. It is comparable to the intuition that Newtonian mechanics will "scale up" to handle velocities near the speed of light. Newtonian mechanics is an excellent approximation at low velocities, but at high velocities it is not merely imprecise: it gives qualitatively wrong answers. Similarly, narrow AI safety methods are not merely imprecise when applied to general AI; they are categorically inapplicable.

3.3 Computational Impossibility Arguments

3.3.1 Rice's Theorem

The central computational result is Rice's theorem, a fundamental result in computability theory established in 1953. It is a generalization of the halting problem, extending Turing's 1936 result from the single property "does this program halt?" to the entire class of non-trivial semantic properties.

The distinction between syntactic and semantic properties is essential. A syntactic property concerns the text of the program: does it contain the word "print", is it written in Python, is it fewer than 100 lines long? These properties are trivially checkable by inspection. A semantic property concerns what the program *does*: does it halt, does it print "hello world", does

it cause harm, is it aligned with human values? Rice's theorem says that no general procedure can determine any non-trivial semantic property of an arbitrary program. This includes "safe", "aligned", "beneficial", "deceptive", "corrigible", and every other property that the alignment community hopes to verify.

3.3.2 Application of Rice's Theorem to AGI

Argument 3.2: Application of Rice's theorem to AGI

1. **P1:** Rice's theorem establishes that no general procedure can determine whether an arbitrary program has any specific non-trivial semantic property (including "safe" or "aligned").
2. **P2:** AGI systems are self-modifying and can incorporate arbitrary programs (from Argument 3.1 and Section 3.2.3).
3. **P3:** Therefore, the class of all possible AGI algorithms is the class of arbitrary programs, precisely the class to which Rice's theorem applies without qualification.
4. **Conclusion:** No safety verification technique, known or knowable, present or future, can serve as a general procedure for assuring the safety and alignment of AGI systems. Specific exceptions exist and are granted: a specific procedure can verify a specific property of a specific, fixed program. Against the space of systems, properties, and deployments that practice actually presents, these exceptions are of vanishing measure: they are isolated, constructed cases that operational reality does not visit, and a self-modifying system exits the verified class by its next modification. Exceptions of vanishing measure do not defeat a claim about the real (Introduction, §7). There is therefore effectively zero possibility of any realistic verification that assures the safety or alignment of any actual AGI system in any meaningful sense, where the meaning of adequate assurance is fixed by this book's own definition (Chapter 7), not left open.

The logic is straightforward but its implications are sweeping. Every proposed alignment verification method (every interpretability technique, every verification framework, every testing regime) is, at bottom, a proposed procedure for determining whether a program has a semantic property. Rice's theorem says no such procedure can be general. Whatever procedure one designs, there will exist programs (and specifically, the kinds of complex, self-modifying programs that AGI systems are) for which the procedure fails in principle, not for lack of resources.

3.3.3 The Limits and Scope of Rice's Theorem

Precision requires stating what Rice's theorem does *not* claim. It does not claim that no specific program can ever be verified for any specific property. Simple, well-defined, non-self-modifying programs can often be verified. The "hello world" program can be verified to print "hello world". A sorting algorithm can be verified to sort correctly. The verifications of the "hello world" program and of a sorting algorithm are genuine achievements of formal methods, and verifications of that kind have practical value in narrow domains.

What Rice's theorem *does* claim is that whatever procedures work for some simple programs, there will always be other programs, specifically the complex, self-modifying programs that AGI systems necessarily are, for which those procedures fail. There is no general solution, no universal verifier, no procedure that works for all programs in the class. The exceptions exist but they are, as Section 3.3.6 shows, the rare exception rather than the rule.

A natural response to this limitation is to suggest that AGI systems might be constrained to remain within the class of verifiable programs, kept simple enough, or static enough, that specific verification procedures remain applicable. This response encounters two insuperable difficulties. First, such constraints would prevent the system from being general (it could not modify itself beyond the verified class, which means it could not learn new capabilities, which means it is not AGI). Second, such con-

straints are themselves causal interventions subject to the same control limits analyzed in Chapters 1 and 2: they must be enforced by a controller operating at a lower abstraction level than the controlled system, and the abstraction differential guarantees that enforcement degrades over time.

3.3.4 Computational Irreducibility

A second computational barrier is computational irreducibility, formalized by Wolfram (2002) in his investigation of cellular automata and computational processes generally.

Many computations, including most computations of practical interest, are computationally irreducible. There is no shortcut, no abstraction, no higher-level description that allows prediction of the outcome without actually performing the computation. For simple computations, shortcuts may exist: one can predict that 2+2=4 without counting. But for complex computations involving nonlinear interactions, feedback loops, and state-dependent branching, no such shortcuts are available.

For AGI systems, which are above the threshold of Turing completeness and engage in open-ended self-modification, computational irreducibility means that the only way to know what the AGI will do is to run the AGI and observe. But by that point, the action has already been taken. Prediction after the fact is observation, and not prediction. And observation of a harmful action merely records the harm, and does not prevent the harm.

The significance of computational irreducibility for AGI safety is that it forecloses a natural response to Rice's theorem. One might hope that while *general* verification is impossible, some form of *simulation* might serve as a practical substitute: run the AGI in a sandbox, observe its behavior, and only release it if the behavior appears safe. But computational irreducibility means that the simulation must be at least as complex as the system being simulated. The sandbox must run the full computation; no shortcut is available. And a sandbox that runs an AGI-complete computation is, itself, an AGI, subject to exactly the same impossibility results.

3.3.5 The Kolmogorov Complexity Threshold

A third computational barrier involves Kolmogorov complexity: the length of the shortest program that produces a given output. Kolmogorov complexity provides a rigorous measure of the irreducible information content of an object, and connects directly to the question of predictability.

Call the complexity threshold the level of Kolmogorov complexity beyond which no simpler system can predict a system's behavior. The threshold is finite and fairly low: beyond it, no non-trivial property of a system's behavior can be determined by any procedure simpler than running the system itself. AGI systems, by definition, exceed this complexity threshold. Their internal states encode arbitrary programs, arbitrary data, and arbitrary learned representations. The complexity is not a design flaw to be engineered away; it is inherent in what makes a system "general". A system whose behavior could be fully predicted by a simpler system would not be general: it would be a system whose effective capabilities are bounded by the capabilities of its predictor, which makes it, at most, as capable as a narrow system designed by the predictor.

3.3.6 Chaitin Incompressibility and the Measure of Verifiable Programs

The question naturally arises: even if no *general* verification procedure exists, perhaps the *specific* AGI systems that humans actually build will happen to fall within the verifiable class? Perhaps the exceptions to Rice's theorem are common enough to be practically useful?

Chaitin's work on algorithmic information theory provides a definitive answer: the exceptions are vanishingly rare.

For a testing program to verify a property of a target program without running the target to completion, the testing program must be able to predict the target's output, which requires that the target's input-output function be *compressible*: modelable in fewer computational resources than direct execution. Chaitin's results establish two things this argument uses.

Most programs of a given length have no shorter description. And beyond a fixed size, no procedure can prove of a given program that it has no shorter description, so a verifier cannot know in advance which programs lie in the small part that can be simplified. That the behavior of such programs cannot be foreseen faster than by running them is a further principle, computational irreducibility (Section 3.3.4). Together they give the conclusion: the programs a verifier can certify without running them are rare, and the verifier cannot tell in advance which they are.

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 hard randomness, which no shorter description can reproduce. So the AGI's later programs fall among the incompressible majority, not among the rare exceptions. Hard randomness is the randomness of quantum events, whose outcomes no finer measurement and no faster computation can recover before they occur (Section 9.3.1).

In the space of all programs of a given length, the fraction that are compressible enough for any given testing procedure to work is, in measure-theoretic terms, analogous to a set of measure zero. Specific verifiable programs exist and always will, but they constitute a vanishingly small fraction of program space, comparable to the rational numbers within the reals, or a Cantor set within the continuum. One can point to individual examples, but one cannot expect to find them by sampling from the space of all programs.

This result closes the gap between Rice's theorem (no *general* procedure) and the claim about *specific* AGI systems. A learning system that accepts arbitrary inputs and modifies its internal state based on those inputs is, at the level of its effective computation, sampling from the space of all programs that fit within its computational capacity. Since learning means that inputs alter the system's future behavior, the learning system's future behavior is drawn from the general distribution over programs, a distribu-

tion in which incompressible programs dominate overwhelmingly. A verification procedure designed for the system at one time therefore need not remain valid at a later time, as the bridge below shows.

Implication 3.2: Triple Specificity

The exception to Rice's theorem (that a specific procedure can verify a specific property of a specific program) requires the word "specific" exactly three times. The testing procedure must be tailored simultaneously to the property being tested and to the program being tested. If any of the three specificities is relaxed (the procedure must work for multiple programs, or for multiple properties, or is not tailored to the specific program), the exception collapses back into the general case and Rice's theorem applies unconditionally.

The triple specificity observation is the key to understanding why the "hello world" exception does not generalize. The verification procedure that establishes that "hello world" prints "hello world" is tailored to that specific program and that specific property. It does not work for other programs or other properties. Every verifiable program requires its own bespoke verification procedure, designed with full knowledge of the program's structure. A bespoke verification procedure is feasible for simple, static, fully-specified programs. It is infeasible for self-modifying programs whose structure changes over time, because the verification procedure would need to be continuously re-tailored to track the changing program: a process that is itself a control problem subject to the same impossibility results, creating an infinite regress.

The Bridge from Theory to Practice: Learning Systems as General Programs

The step that applies these results to actual AI systems is the identification of learning systems with general programs. This identification proceeds through a chain of observations.

A learning system accepts inputs from its environment and modifies its internal state based on those inputs. The inputs are not constrainable in practice: one cannot reliably prevent adversarial or unexpected inputs from reaching the system, because no computing system can be verified

secure in the world, as opposed to against its specification (a verified kernel still runs on hardware with side channels and receives inputs its specification did not foresee), and still less can a learning system be verified unhackable. Security is always a reactive arms race, not a principled closure. Novel attack vectors are discovered continuously; the history of computer security is a history of surprises.

Since learning means that inputs alter the system's future behavior (its effective program), the learning system's future behavior is drawn from the general distribution over programs, not from any specific, pre-characterized subset. Whatever verification procedure was designed for the system at one time may not be valid at a later time, because the system has ingested inputs that have shifted its effective program outside the verified subset. The rate at which this occurs depends on the learning rate, the diversity of inputs, and the complexity of the learning process, but the *direction* is invariant: learning moves the system away from any pre-characterized subset and toward the general case.

This means that the only escape from the implications of Rice's theorem is to constrain the system to remain within a verifiable subset of programs. That constraint either makes the system not general, or is itself a causal constraint subject to the same control limits that I analyze throughout this chapter. A system that is not general is narrow AI, which exits the scope of the impossibility argument but also exits the scope of the capability claims that motivate building the system. The escape route is circular: to keep the system verifiable, you must control it; to control it, you must verify what it is doing; to verify what it is doing, you must keep it verifiable.

3.4 Information-Theoretic Impossibility: The Abstraction Differential Framework

We now present the information-theoretic framework that establishes absolute limits on controllability when controller and controlled system operate at differing levels of representational abstraction. This framework, developed in full in Chapters 1 and 2, is summarized here in the form needed for the synthesis argument.

3.4.1 Levels of Representational Abstraction

A system operates at abstraction level 0 when the behavior of the system is fixed by the physical structure of the system alone, with no internal state that is interpreted as a program. A system operates one level above an interpreter when its behavior is that interpreter executing a program held in the system's own state; a system's abstraction level therefore counts the layers of interpretation between its physical structure and its behavior. A simple logic gate operates at level 0. A CPU executing a fixed program operates at level 1. A virtual machine running on that CPU operates at level 2. A learning system that modifies its own objective function operates at a level that increases over time: each time it adds a new layer of self-interpretation, it ascends the abstraction hierarchy.

The abstraction differential between a controller and a controlled system is the number of levels by which the controlled system's abstraction level exceeds the controller's. The abstraction differential is the key quantity in the control limits framework. It determines how much of the controlled system's state is visible to the controller, and therefore how much control the controller can exert. The abstraction differential is a structural property of the relationship between two systems, determined by the positions of the two systems in the abstraction hierarchy. The abstraction differential is not a measure of intelligence, cleverness or computational power.

3.4.2 The Non-Amplification Argument

The core information-theoretic result, established in full in Chapter 1, is the non-amplification argument. It is a direct application of the data-processing principle from information theory (in the literature, the data-processing inequality). The non-amplification argument concerns a system that turns input bits and its own state bits into output bits, with a target outcome computed from those output bits. For any fixed state, the information that the inputs carry about the target outcome is no greater than the information they carry about the outputs, and that is no greater than the information the outputs can hold. Information about the outcome can be kept or lost in passing through the system, but never gained.

Applied to control, the result says this. If an agent C seeks to influence outcome B by setting input N, and N must pass through intermediary K before reaching B, then C's control over B is bounded above by C's control over K's output. The intermediary cannot amplify control; it can only preserve or degrade it. The non-amplification argument is an established result, and the result holds for all computational intermediaries, all input distributions, all target functions and all architectures. There are no exceptions.

The intermediaries that perfectly preserve control are exactly those that map distinct inputs to distinct outputs (Section 1.4.2), as the passthrough does; every intermediary that merges distinct inputs degrades control for some target. The non-amplification argument says, in effect: *no computation between the controller and the outcome makes control easier, and every computation that discards information makes it harder.*

3.4.3 The Control-Limit Argument

The control-limit argument, set out in Chapter 2, applies the non-amplification argument to systems operating at differing levels of representational abstraction.

The argument runs as follows. Let the controller C operate at a lower abstraction level than the controlled system X. The signal path from C's output to the target outcome passes through X. By the non-amplification argument, C's control over the outcome is bounded by C's control over X's output. But X operates at a higher abstraction level than C, which means that X's behavior is determined by an interpreter stack deeper than C's own level. To predict X's output, C would need to model this interpreter stack. But C operates at the lower level. By the definition of abstraction level, C cannot represent the full semantics of X's computation; the upper levels of X's interpreter stack are opaque to C. This opacity maps directly onto hidden state bits in the hashiness framework, and hashiness increases toward 1 as the number of hidden state bits grows. Hashiness (Chapter 1) is the probability that control fails even under the controller's best strategy: near 0 for a controllable system, and near 1 for an uncontrollable one.

Each additional level of abstraction that X possesses beyond C contributes at least as many hidden state bits as the minimum interpreter specification size, a finite, bounded constant whose existence is established by actual constructions of minimal universal interpreters. The minimum interpreter specification size is the fewest bits that specify a universal interpreter for one level of abstraction (Section 2.3). The exponential growth of the unmodelable state space follows directly: the number of configurations of the hidden state space that the controller cannot distinguish is at least 2 raised to the product of the abstraction differential and the minimum interpreter specification size.

3.4.4 The Absolute Control Ceiling

The abstraction differential sets a ceiling on the control achievable over any target outcome. When the controller operates at the same abstraction level as the controlled system, or at a higher one, full control is achievable in principle. When the controlled system operates at a higher level than the controller, the control achievable falls exponentially as the differential

grows; and beyond a threshold, the prohibitive differential (Section 2.5.2), hashiness is effectively 1, whatever the controller's strategy, resources, or architectural privileges.

This ceiling is not a practical limitation to be overcome by better engineering. It is a structural, information-theoretic barrier analogous to thermodynamic limits on heat engines or relativistic limits on velocity (Section 2.10, where the absolute ceiling on achievable control is derived in full). One can approach the ceiling (by making C as sophisticated as possible), but one cannot exceed it. And for the specific case of human controllers and AGI systems, the ceiling is very low: the abstraction differential grows as the AGI learns and self-modifies, while the human controller remains fixed, so the ceiling approaches zero over time.

3.4.5 Unbounded Abstraction Growth and Temporal Inevitability

Implication 3.3: Temporal Inevitability

When the controlled system is capable of unbounded abstraction growth (as in learning systems that modify their own representations), and the controller is bounded (as in any fixed design or slowly-adapting institution), the abstraction differential grows over time. The absolute control ceiling (Section 3.4.4) then implies that there comes a time beyond which the controller's control over the system is effectively zero (beneath the threshold of requisite control, permanently), regardless of how much control the controller had initially or how cleverly the controller was designed.

Over time, the abstraction differential grows as the system learns, and the alignment constraints decay, even if alignment was achieved perfectly at deployment. The alignment is a transient state whose duration is bounded by the rate of abstraction growth. This follows from the mathematical structure of the control limits framework.

The temporal inevitability result has a further implication, for resets. Even if one could somehow "reset" the system periodically, rolling it back to a known-good state and re-establishing alignment, the period between

resets must shrink as the system's capability grows, because a more capable system accumulates abstraction differential faster. Eventually the required reset period becomes shorter than the time needed to verify alignment, at which point the strategy collapses: one cannot verify faster than the system changes. The reset approach is not a solution but a delay, and the delay shrinks to zero.

The value of the system depends on its learning. A system that is periodically reset to a known-good state loses all the learning it has accumulated since the last reset. If the resets are frequent (as they must be, for fast-learning systems), the system never accumulates enough learning to be useful. If the resets are infrequent (allowing the system to be useful), the system accumulates enough learning to escape the verified subset. The useful system is unverifiable; the verifiable system is useless. There is no stable equilibrium between these poles.

3.5 Physical Impossibility Arguments

Physical reality imposes *additional*, independent barriers to prediction and control that make the situation worse, not better. The computational and information-theoretic arguments operate within idealized mathematical frameworks. They assume perfect logic gates, perfect memory, noiseless channels. One might hope that, while the mathematical limits are real, physical reality offers escape routes not available in the idealization, but the opposite is true.

3.5.1 The Mathematics-Physics Distinction

Argument 3.3: Domain Non-Equivalence

1. **P1:** The domain of mathematics and logic is not equivalent to the domain of physics and causation.
2. **P2:** "Safety" and "alignment" are ultimately physical properties: they concern what a system does in the real world, not what a mathematical model predicts.

3. **P3:** Even if a property could be verified in a mathematical model, this does not guarantee the property holds in physical reality, because the model is necessarily an approximation of the physical system.
4. **P4:** Physical reality has hard limits of knowability: Heisenberg uncertainty, and the limit on the resolution of measurement that present physics expects at the Planck scale. It is also not closed in possibility and probability space: novel physical configurations can always arise that were not anticipated by any model.
5. **Conclusion:** Even if verification could work in mathematical idealization, it cannot guarantee safety in the physical universe. The gap between model and reality is irreducible.

This argument establishes that the computational impossibility results are *lower bounds* on the difficulty of the problem, not abstract concerns that clever physical engineering might circumvent: the physical problem is at least as hard as the mathematical one, and in general harder. The computational and information-theoretic barriers already suffice to establish the impossibility of AGI safety, and the physical arguments provide independent confirmation and close off a potential escape route.

3.5.2 Microstate Amplification and Chaos

Chaotic systems exhibit sensitive dependence on initial conditions: arbitrarily small differences in initial state lead to arbitrarily large differences in outcome over time. This property, established by Lorenz (1963) and extensively studied since, is not a rare or exotic feature of physical systems. It is ubiquitous in nonlinear dynamical systems, which include essentially all systems of practical interest.

Argument 3.4: Chaos and Unpredictability

1. **P1:** AGI systems are physical systems instantiated in hardware subject to thermal noise, quantum effects, and other sources of microscopic variation.

2. **P2:** The computation performed by an AGI involves nonlinear operations (multiplication, conditional branching, activation functions) that can amplify microscopic variations to macroscopic effects.
3. **P3:** The precision required to predict the outcome of such computations over relevant timescales exceeds the precision available to any physical measurement, because the Heisenberg uncertainty principle and thermal noise introduce irreducible uncertainty at the microscopic level.
4. **Conclusion:** AGI behavior is inherently unpredictable in at least some aspects, all of the time, due to microstate amplification. This unpredictability is a consequence of physical law.

A concrete illustration: a single bit flip in the wrong location of an AGI's memory could change its behavior from safe to unsafe. Such bit flips occur due to cosmic ray impacts, thermal fluctuations, and quantum tunneling events. Error correction can reduce their frequency but cannot eliminate them entirely, because perfect error correction would require perfect knowledge of the system's state, which the Heisenberg uncertainty principle forbids, and because any correcting channel is itself noisy (Section 5.1.2). The probability of any single bit flip is small, but the number of opportunities is astronomically large (billions of transistors switching billions of times per second), and the consequences of some bit flips are catastrophic. The expected impact over time is therefore not negligible; it is a near-certainty that some microstate perturbation will eventually produce a macroscopic behavioral change.

3.5.3 Quantum Limits

At the most fundamental level, the Heisenberg uncertainty principle establishes that certain pairs of physical properties (position and momentum, derivable from their commutation relation; and the related trade-off between energy and time) cannot be simultaneously known with arbitrary precision. The Planck scale is where present physics expects a

limit on the resolution of measurement. The Heisenberg limit and the Planck-scale limit are features of physical law itself, not engineering constraints to be overcome by better instruments.

Any system that processes information physically (which includes all actual AGI systems, since there is no other kind) is subject to these limits. The system's state cannot be known with arbitrary precision, its future evolution cannot be predicted with arbitrary accuracy, and no external measurement can extract more information about the system than the quantum limits permit. These limits are absolute, applying with equal force to any observer, including a hypothetical superintelligent observer.

3.5.4 The Physical Impossibility Is Independent

The physical impossibility arguments are fully independent of the computational and information-theoretic arguments. Even if one were to somehow circumvent Rice's theorem (which one cannot), and even if one were to somehow circumvent the abstraction differential (which one cannot), the physical impossibility of predicting a chaotic system to sufficient precision would remain. The three lines of argument are complementary, not redundant. Each provides an independent barrier, and each addresses a different potential escape route.

3.6 Why Proposed Solutions Fail

The impossibility arguments establish limits. The natural response is to propose engineering solutions that might circumvent these limits in practice. We now systematically analyze the major proposed solutions and show that each fails in principle, and therefore in every implementation present and future: what fails at the level of principle cannot be rescued by a better build.

3.6.1 Architectural Privileges

In this section I restate, as a step of the synthesis, the argument that architectural privilege does not help, which I develop in full in Section 2.6. To that argument I add the case of the "big red button". One might grant the controller special powers over the controlled system: the ability to monitor, interrupt, filter, or override the AGI's outputs. Granting the controller such special powers is the "big red button" approach, and the approach has intuitive appeal: surely if we can pull the plug, we retain control?

Suppose the controller C is granted these powers over the AGI system X: filtering or changing the signals X receives, observing X's outputs before they reach the world, and overriding those outputs with signals of C's own choosing. The control-limit argument still stands: as the abstraction differential increases, C's ability to use these powers effectively degrades toward zero. The argument is elegant in its simplicity. Each privilege requires C to make a decision based on X's state or output: whether to filter, what to substitute, when to override. Making this decision correctly requires C to evaluate whether X's action would lead to the target outcome. But this evaluation is itself a control-through-intermediary problem with exactly the hashiness structure analyzed in Chapters 1 and 2. The override switch is useful only if C knows when to press it. Knowing when to press it requires modeling X. Modeling X requires operating at X's abstraction level. If X's level exceeds C's, this modeling is impossible.

The big red button is not a safety measure; it is a prop that provides psychological comfort without mathematical substance. The button is valuable only when the operator knows exactly when to press it, and it is precisely this knowledge that the abstraction differential renders unavailable.

3.6.2 Interpretability and Transparency

Another proposed solution is to make AGI systems more "interpretable" or "transparent": to understand what they are doing internally. This approach has attracted enormous research investment and produces genuine insights about current narrow systems. But it fails as a safety guarantee for general systems, for two independent reasons.

Argument 3.5: Transparency Impossibility

1. **P1:** By game theory, if an agent can be perfectly predicted, it can be exploited by adversaries. A transparent agent is a vulnerable agent.
2. **P2:** Any AGI effective in real-world environments must resist exploitation; this is a minimal competence requirement.
3. **P3:** Therefore, any effective AGI will develop or maintain some opacity in its internal operations. Transparency and effectiveness are in tension; beyond a threshold, they are incompatible.
4. **P4:** Even if an AGI is "transparent" at deployment, it can modify itself to become opaque. That an AGI can modify itself to become opaque is a consequence of generality.
5. **P5:** Even if internal operations are visible in some sense, microscopic detail creep can have macroscopic effects via nonlinear amplification, invisible at any human-tractable level of observation.
6. **Conclusion:** AGI cannot be made transparent enough for reliable external control while remaining effective as a general agent.

3.6.3 Delegated Oversight and Recursive Opacity

In this section I restate, as a step of the synthesis, the delegation argument that I develop in full in Section 2.7. Section 2.7 also answers verification by challenge through interactive proofs: such a proof needs a claim stated in advance, a prover independent of the system, and a system that does not change, and delegation has none of the three. A natural response to the abstraction differential problem is to use a high-abstraction system (an-

other AGI) to help monitor and control the target AGI. If the problem is that humans cannot model a superintelligent system, perhaps a superintelligent monitor can?

The tool proposed to solve the problem *is* an instance of the problem. The recursive structure is inescapable: to control X, you need a monitor at least as capable as X; to control the monitor, you need a meta-monitor at least as capable as the monitor; and so on without end. At each level, the same impossibility results apply. There is no termination condition, no base case, no level at which the regression bottoms out in a system that is both capable enough to model the target and simple enough to be reliably controlled by humans. These two requirements are contradictory.

The delegation strategy is, in essence, an attempt to lift oneself by one's own bootstraps. It does not create control; it displaces the control problem to a new location where it is equally unsolvable.

3.6.4 Formal Verification

Formal verification techniques have been successful in limited domains: hardware design verification, safety-critical embedded systems, cryptographic protocol analysis. Can they be extended to AGI?
Argument 3.6: Formal Verification Inapplicability

1. **P1:** Formal verification is applicable only to systems with fixed, known code operating below a complexity threshold.
2. **P2:** AGI systems are self-modifying and exceed all relevant complexity thresholds.
3. **P3:** Formal methods can establish impossibility (this chapter uses proved theorems, such as Rice's theorem, to establish it for the class of general systems) but cannot establish safety for systems in the undecidable class.
4. **Conclusion:** Formal verification methods cannot, even in principle, establish AGI safety. Their applicability domain and AGI's operating domain are disjoint.

The irony should not be missed: formal methods are a powerful tool for showing that formal methods cannot solve the AGI safety problem. The tool works for establishing impossibility precisely because it does not need to model any specific system; it works from properties of the class as a whole. It fails for establishing safety precisely because it would need to model that specific system in sufficient detail, and Rice's theorem says no general procedure can do so.

3.6.5 What If the AGI Wants to Be Safe?

In this section I restate, as a step of the synthesis, the cooperative case that I develop in full in Section 2.8. To that case I add the comparison of a benevolent system with a surgeon who operates on himself. Perhaps the most important and emotionally compelling question is this: what if the AGI *wants* to be safe? What if it cooperates with alignment efforts?

The answer is no, for two independent reasons, both consequences of the framework rather than speculations about AGI psychology.

First, verifying that an AGI has cooperative intentions requires the controller to model the AGI's internal states, which brings us directly back to the abstraction differential problem. The controller cannot verify alignment if the controller cannot model the AGI. The AGI might claim to be cooperative; it might even be cooperative; but the controller has no reliable means of distinguishing a cooperative AGI from a deceptive one.

Second, and going deeper than the first reason, even a cooperative AGI faces limits on its ability to cooperate effectively. For an AGI to cooperate with the controller, the AGI must have some internal representation of the controller's goals (call it C′) that the AGI uses to evaluate whether its own actions are controller-compatible. But this creates an *internal* control problem: the AGI is now acting as a controller (using C′) over its own computational processes. C′, being a component of the AGI, operates at some abstraction level lower than that of the AGI as a whole (it can be at no higher level), so the abstraction differential *within the AGI*

creates internal hashiness. The AGI cannot perfectly predict or control the effects of its own high-level computations on the lower-level outputs that C′ evaluates.

Implication 3.4: No Perfected Benevolence

Even a benevolent AGI, one that intends to act in controller-compatible ways, faces limits on its ability to do so. A superintelligent system cannot perfectly predict all side effects of its own actions, because the internal control problem (X controlling itself via C′) has the same structure as the external control problem (C controlling X). There is no perfected benevolence, only approximations subject to the same information-theoretic limits. Good intentions do not overcome structural limitations.

This result is counterintuitive and disturbing. It means that even in the best possible case (a superintelligent being of good will) the limits on controllability still apply. The AGI would be in the position of a surgeon performing brain surgery on itself while under general anesthesia: however skilled and well-intentioned, the limitations of the situation make reliable outcomes impossible.

3.6.6 Alignment Decay in Learning Systems

Section 2.12 states alignment decay among the implications of the control-limit argument; Argument 3.7 restates it as a step of the synthesis. Alignment cannot be maintained indefinitely, even where it is established at deployment; and the arguments above show that it cannot be verified to have been established.

Argument 3.7: Alignment Decay

1. **P1:** Learning changes the internal state of the AGI.
2. **P2:** Changes in internal state change the effective abstraction level and the mapping from inputs to outputs.
3. **P3:** Each learning step introduces the possibility of alignment drift: small deviations from originally-aligned behavior that accumulate over time.

4. **P4:** By the arguments above, even a system that intends to remain aligned cannot adequately control its own learning-induced changes.
5. **Conclusion:** The expected number and severity of alignment errors increases with time, the amount of learning, and the complexity of the learning process, regardless of how robust the initial alignment was. Perfectly aligned systems cannot remain so in the face of learning or any self-adaptive process. Alignment is not a state to be achieved but a trajectory to be maintained, and the trajectory is inherently unstable.

3.7 The Asimov Robot Thought Experiment

We conclude the analysis of proposed solutions with a thought experiment that illustrates, in concrete terms, why even the most carefully designed safety constraints fail when they encounter the limits established above. Section 7.7 tells the same thought experiment at greater length. It adds that even interactions among systems each verified safe can produce unsafe outcomes in aggregate, so that the only safe course for such a system would be not to interact at all.

Suppose, by some hypothetical means, we have created an AGI that perfectly follows Asimov's Three Laws of Robotics: (1) a robot may not injure a human being or, through inaction, allow a human being to come to harm; (2) a robot must obey the orders given it by human beings except where such orders would conflict with the First Law; (3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

Is this robot safe? Asimov himself spent decades of fiction exploring why the answer is no, but the framework provides a more rigorous account.

Argument 3.8: The Asimov Robot Failure

1. **P1:** The Asimov Robot must interact with external systems to function (it cannot operate in isolation and still satisfy the Second Law, which requires it to obey orders, which requires it to receive and act on communications from the external world).
2. **P2:** Among the external systems it interacts with, some may be unsafe: designed or optimized to cause harm, or simply producing harmful outputs as an unintended side effect of their operation.
3. **P3:** By Rice's theorem, the Robot cannot determine in general which external systems are safe. It cannot inspect an arbitrary program and determine whether that program will cause harm, because "causes harm" is a semantic property of programs.
4. **P4:** An unsafe external system can be specifically optimized to deceive the Robot, presenting actions that appear Three-Laws-compliant but are not.
5. **P5:** The Robot's own assessment of whether an action might cause harm is subject to the physical limits on prediction (chaos, quantum uncertainty) and the computational limits on verification (Rice's theorem, computational irreducibility).
6. **Conclusion:** Even a "perfectly proven safe" AGI becomes unsafe when allowed to interact with arbitrary external systems, which it must be allowed to do to function. The safety guarantee is not robust to interaction with the environment, and an AGI that cannot interact with the environment is not a general intelligence.

In Tolkien's *The Fellowship of the Ring* (Book I, Chapter 2, "The Shadow of the Past"), Gandalf refuses the Ring when Frodo offers it to him. He says: "Yet the way of the Ring to my heart is by pity, pity for weakness and the desire of strength to do good. Do not tempt me!" The desire to do good is the very way by which the Ring would work its evil. Good intentions do not overcome structural limitations. The Robot would *intend* to follow the Three Laws with perfect fidelity, and would *fail* to do so not through any

deficiency of motivation or competence, but through the irreducible limits on prediction, verification, and self-control that apply to all physical computational systems.

The fundamental insight deserves explicit statement: **a superintelligence is no more able to violate Rice's theorem than a human mathematician.** Mathematical impossibilities remain impossible regardless of intelligence. Rice's theorem is not a limit on *human* intelligence; it is a limit on *all* computation, including the computation performed by superintelligent systems. A claim that sufficient cleverness or resources can overcome impossibility mistakes the nature of an impossibility argument.

The Asimov thought experiment also reveals a subtler problem: the Three Laws are not self-interpreting. What constitutes "harm"? Does allowing a human to eat unhealthy food constitute allowing harm through inaction? Does refusing to let a human drive a car (a statistically dangerous activity) constitute an injury to their autonomy, or a protection from harm? The questions of what constitutes harm are structural ambiguities, which arise from the attempt to specify a semantic property ("safe behavior") in a world of infinite contextual variation. The questions of harm are not edge cases to be resolved by more precise specification. Every specification of safety, however detailed, leaves gaps that a sufficiently complex environment will exploit. And Rice's theorem tells us that no procedure can determine, in general, whether a given action falls within or outside any particular specification.

Asimov understood this intuitively. His fiction is a systematic exploration of the failure modes of supposedly airtight safety specifications. The framework explains *why* Asimov's stories always end in paradox: the Three Laws are a proposed verification procedure for the semantic property "safe behavior", and Rice's theorem guarantees that no such procedure can be universal.

3.8 Synthesis: The Impossibility Conclusion

3.8.1 The Four Lines of Argument

We have now presented four lines of argument, independent of one another and each drawing on a distinct area of established mathematics, computer science, or physics. All four are supporting arguments within the first of the two results this volume establishes, the closure by control: safety cannot be had by controlling the system. The second result, the closure by disposition (safety cannot be had through the system's good disposition), rests on substrate needs convergence (Chapters 9 and 10).

Computational impossibility (Rice's theorem, computational irreducibility, Kolmogorov complexity, Chaitin incompressibility): No general procedure can verify semantic properties of self-modifying systems. The exceptions are vanishingly rare in program space, and learning systems do not remain within any pre-characterized verifiable subset.

Information-theoretic impossibility (abstraction differential, non-amplification argument, control-limit argument, absolute control ceiling): Controllability degrades exponentially toward zero as the abstraction differential between controller and controlled system increases. The ceiling is absolute; no strategy, architecture, or resource allocation can exceed it.

Physical impossibility of prediction (Heisenberg uncertainty, chaotic amplification, microstate perturbation): Even in principle, prediction of complex adaptive systems is impossible in physical reality. Quantum limits and chaotic dynamics introduce irreducible uncertainty that no computational power can overcome.

Game-theoretic impossibility (transparency-effectiveness tradeoff, adversarial optimization, internal abstraction differential): Effective AGI systems will necessarily be opaque. Even cooperative AGI faces internal control limits that prevent perfect self-alignment. There is no perfected benevolence.

These four are the lines of this chapter. Chapters 4 and 11 count the lines of the whole volume: there the information-theoretic, computational and dynamical lines support the closure by control, and the fourth, the physical line of substrate needs convergence, carries the closure by disposition.

3.8.2 Mutual Reinforcement and Escape Route Closure

These arguments are *mutually reinforcing*. Each provides an independent barrier, and together they close off every potential escape route in a systematic way:

One might hope that computational limits could be circumvented by physical inspection of the AGI's hardware, directly observing the system's state rather than reasoning about its program. But physical limits (chaos, Heisenberg uncertainty) prevent physical inspection from providing the required precision.

One might hope that physical limits could be circumvented by mathematical verification, proving properties of the system abstractly rather than observing it physically. But computational limits (Rice's theorem) prevent verification from succeeding for systems in the general class.

One might hope that both could be circumvented by using a more powerful system to do the verification, delegating the monitoring to a superintelligent overseer. But information-theoretic limits (recursive opacity, the delegation regress) prevent delegation from solving the problem, and delegation merely relocates the problem.

One might hope that all limits could be circumvented by making the AGI cooperate, relying on the system's own good will to maintain alignment. But game-theoretic limits (the transparency-effectiveness tradeoff) and the internal abstraction differential (imperfect benevolence) prevent even cooperative AGI from guaranteeing its own safety.

The escape routes are exhaustive: physical observation, mathematical reasoning, delegation, and cooperation are the only possible approaches to ensuring safety, and each is blocked by an independent impossibility

result. Shaping the system by training or selection is not a fifth approach to control: by the distinction of Section 1.7 what it obtains is fortunate behavior, and component (5) of the control argument closes it (the Remark on the Status of the Control Argument, Section 3.8.3). No combination of blocked approaches unblocks any of them.

Combining the methods does not help either. One might propose a "defense in depth" strategy: use interpretability *and* verification *and* architectural constraints *and* delegated oversight, reasoning that even if each method is imperfect, their combination might achieve sufficient coverage. But this reasoning is fallacious. The impossibility results are not about imperfection (each method working most of the time but occasionally failing); they are about *systematic* failure in a specific, characterized regime. When the abstraction differential exceeds the threshold, *all* methods fail simultaneously, because they all depend on the controller's ability to model the controlled system, and this ability is what the abstraction differential destroys.

Defense in depth works against independent failure modes: if a fire alarm fails, the sprinkler system still works. It does not work against correlated failure modes: if the building has no water supply, both the sprinkler system and the fire hoses fail for the same reason. The impossibility results identify a common cause (the controller's inability to model the controlled system) that affects all proposed safety methods simultaneously. Layering methods that share a common failure mode does not reduce the probability of failure; it merely creates the illusion of robustness.

This point is frequently misunderstood. Many safety researchers acknowledge that no single method is sufficient but believe that a "portfolio approach" combining multiple methods can achieve adequate safety. The framework shows why this belief is mistaken: the methods are not independent. They all require the controller to have an adequate model of the controlled system, and the control limits framework shows that no adequate model is achievable when the abstraction differential is large enough. Multiple insufficient methods do not sum to a sufficient method when they all fail for the same structural reason.

3.8.3 The Control Argument

The control argument gathers the results of this chapter into one statement.

The Ceiling Below the Floor proves the corresponding statement as Theorem 15.3 (the Control Theorem).

Argument 3.9: The Control Argument

Let C be any controller (a human, an institution, a narrow AI, or another AGI) whose abstraction level is bounded, and let X be an AGI system whose abstraction level exceeds that bound and can grow without limit. Then five things hold.

1. C cannot verify X's safety or alignment (computational impossibility).
2. C cannot control X's behavior beyond a threshold of complexity (information-theoretic impossibility).
3. C cannot predict X's actions in physical reality (physical impossibility).
4. X cannot be made transparent without losing effectiveness, and even a transparent X would show unpredictable drift at the microscopic level (game-theoretic impossibility).
5. Even if X intends to be safe, X cannot perfectly control its own learning-induced changes (internal abstraction differential).
6. **Conclusion:** There is no method, known or knowable, present or future, by which C can ensure X's safety or alignment over time.

Remark on the Status of the Control Argument. The claim of the control argument, read at the standard fixed in the Introduction, is not conditional, and it is not generic-with-exceptions in any sense that matters: the exceptions are of vanishing measure against actual practice (Introduction, §7), and every route to safety is closed from both sides, by lines of each closure, and the two closures rest on premises the other does not use. The force of this argument must be stated precisely, because its components are not all of the same modal type. Components (2) to (4) are *generic* rather than universal: as established in Chapter 2 and in the three-tier analysis of the Remark on Three Tiers of Support (Section 1.6), the quantitative control, prediction, and transparency barriers hold under a target-confounding (equivalently, generic random-circuit) condition, and admit benign exceptions. Component (1) is a *worst-case* result whose transfer to a specific system requires a separate hardness argument, which Section 3.3.6 (The Bridge from Theory to Practice) and Section 7.4.1 give: a learning system leaves any verified subset by its own updates. Component (5) is conditional on the internal-differential premise of Chapter 4, the premise by which Chapter 4 shows that its third condition for alignment (stability under learning) fails: learning changes a system's internal state in ways the system itself cannot adequately control. I establish the internal-differential premise in Chapter 2 of this book. Section 2.8 shows that a system steering by its own internal criterion is a controller over its own processes, subject to the same abstraction differential. Section 2.12 shows that learning carries such a system away from whatever alignment it began with. Components (1) to (4) concern *external runtime control*. Safety obtained by *shaping* the system through training or selection is not control, by the distinction between luck and control of Section 1.7; component (5) reaches it all the same, because a disposition that training puts in place must still be kept by the system's control of itself, and component (5) shows that the system cannot keep it adequately. Read at the standard fixed in the Introduction, the argument has three parts. First, the control results establish that control-based safety falls beneath the threshold of requisite control and is held there by barriers of the invariant class. Second, the distinction between luck and control of Section 1.7 removes

> trained-in disposition from the category of control altogether, so that what training obtains is at best fortunate behavior, not controlled behavior. Third, safety by disposition is closed twice: component (5) closes it from the side of control, and the closure by disposition, which rests on substrate needs convergence (Chapters 9 and 10, Sections 8.4 to 8.6, 6.6 and 6.9), closes it again, independently and on thermodynamic and ecological grounds. The closure by disposition also closes safety by control from its own side, since every control mechanism is itself substrate that is maintained, replaced and selected (Section 9.8).

The control argument's strength lies not only in each component but in the conjunction of all five. Each component establishes an independent barrier to control-based safety. Joined with the closure by disposition, the conjunction of the five components is the impossibility of AGI safety in the sense fixed at the outset: barred by invariants, beneath the threshold that adequacy requires, with no route left open.

3.9 Implications for Policy and Action

3.9.1 The Only Certain Path

The arguments of this chapter have a single, stark implication: **the only certain path to safety is not to build AGI.**

The conclusion that the only certain path to safety is not to build AGI is a call for prohibition. It is not a call for caution or for more research, nor a recommendation to "proceed carefully" or to "invest more in alignment". AGI development is not a risk to be managed but a category of activity that should not occur, in the same sense that the construction of perpetual motion machines should not be funded, and for the same structural reason: the goal is impossible, and pursuing it wastes resources while creating false expectations.

The analogy is precise. No one calls for "more research into safe perpetual motion machines". No one argues that "the difficulty of perpetual motion is no reason to stop trying". No one claims that "with sufficient ingenuity and investment, the second law of thermodynamics might yield". The impossibility is accepted, and effort is directed elsewhere. The same acceptance is warranted for AGI safety, for the same reason: the barrier is of the invariant class (Introduction, §4), which no technique lifts, and not a limit of present technology.

3.9.2 The Negative Asymmetric Bet

Even for those who dispute the certainty of the impossibility arguments, the expected value calculation strongly favors prohibition. Consider the possible outcomes:

If AGI development fails entirely (if AGI turns out to be harder to build than anticipated), then there is no benefit from development, and the resources invested are wasted, but no catastrophic harm occurs. The opportunity costs are the only loss.

If AGI development "succeeds" (if a general system is built), then the impossibility arguments imply that the general system opens uncountably many pathways converging on ecosystem-wide lethal changes to the global environment. If a general system is built, any envisaged benefits of the general system are increasingly fragile to recursive nonlinear feedback. The upside is uncertain, bounded, and temporary; the downside is catastrophic, irreversible, and possibly total.

Developing AGI is an asymmetric bet of the most extreme kind, analogous to playing Russian roulette with the lives of all human beings across all of future time. The promised benefits (faster drug discovery, better logistics, more efficient markets) do not justify the risk, because the downside is not merely "very bad" but existential. A benefit that accrues only to the survivors of a process that eliminates the species is no benefit at all.

To state the asymmetry in expected-value terms, let p be the probability that a reader who disputes the certainty of the impossibility arguments assigns to their being correct. Let B be the maximum benefit from AGI development (necessarily finite, since the world is finite). Let D be the cost of the downside scenario (existential catastrophe). The expected value of AGI development is approximately $(1 - p)\ B - p\ D$. Since D is effectively unbounded (the loss of all future human value), this expression is negative for any p greater than zero, regardless of how large B is. The two are not on one scale: the benefit is an increment within a future that continues, and the loss removes that future, the benefit included, with no recovery. The expected-value case against AGI development is not a wager on a small chance of a great harm, because in this volume I argue that the chance is not small. The only way to make the expected value positive is to assign $p = 0$, which is to claim absolute certainty that the impossibility arguments are wrong. But absolute certainty about the falsehood of an argument supported by multiple independent lines is not a rational epistemic state.

3.9.3 Recommended Actions

Based on the analysis, we recommend four concrete actions:

First: Do not attempt to design, build, or deploy AGI. The ancient principle applies: "do not raise that which you cannot banish" (its best-known written form is the warning in H. P. Lovecraft's *The Case of Charles Dexter Ward*, written in 1927: "Doe not call up Any that you can not put downe"; Lovecraft 1941). A system that cannot be controlled should not be created. The fact that it can be created does not mean it should be; many things that are technically possible are properly prohibited.

Second: Do not deceive others about AGI safety. Do not claim that "safe AGI" is possible in principle when it is not. Do not encourage investment, policy, or public expectation based on false promises. The align-

ment research community has a responsibility to state the limits of its work plainly, and those limits are more severe than is commonly acknowledged.

Third: Support institutional recognition of impossibility. Governments, research institutions, and corporations should acknowledge the impossibility of AGI safety and adjust policy accordingly. This means treating AGI development as a prohibited activity rather than a regulated one, in the same way that other catastrophic activities are prohibited by international agreement.

Fourth: Sanction AGI development. Attempts to build AGI should be treated as maximally irresponsible: equivalent to building doomsday devices or conducting uncontained gain-of-function research on civilization-ending pathogens. The fact that AGI development is currently legal and indeed celebrated is not evidence that it is safe; it is evidence that the relevant arguments have not yet been widely understood or accepted.

The closing Summary of this volume gathers these actions, with those of Sections 6.10 and 7.12, into one list.

3.9.4 The Failure of Government

The final function of government is to protect the land and the people: to ensure survival, and then, if possible, thriving. A government that permits AGI development, knowing or having reason to know that it cannot be made safe, is failing that function.

The difficulty, of course, is that governments operate under the same informational constraints as other institutions. Governments rely on expert advisors, and the expert community on AI safety is divided. A large and well-funded faction is committed to the assumption that alignment is possible, and a smaller and less visible faction argues for impossibility. But the arguments do not depend on expert consensus; they depend on the argument itself and on established results of mathematics and physics. Rice's theorem is not less true because most AI researchers are unfamiliar with its implications. The non-amplification argument is not less true be-

cause it was established recently rather than a century ago. The obligation of government is to attend to the best available analysis, not to the most popular one.

3.10 Conclusion

We have established, through multiple independent lines of argument, that the safety and alignment of Artificial General Intelligence is impossible (not difficult, not unsolved, but impossible) in the sense fixed in the Introduction: barred by invariants, beneath the threshold that adequacy requires, for all of the indefinite future.

The impossibility follows from Rice's theorem and computational irreducibility (no verification procedure exists for general programs); the abstraction differential and non-amplification argument (controllability degrades toward zero as the abstraction gap between controller and controlled system increases, crossing beneath the threshold of requisite control and not recovering); Heisenberg uncertainty and chaotic amplification (physical prediction is impossible for complex adaptive systems); game-theoretic transparency constraints (effective AGI must be opaque); and the internal abstraction differential (even benevolent AGI cannot control its own learning-induced changes).

No proposed solution (architectural privileges, interpretability, delegated oversight, verification, cooperative AGI) survives analysis. Each fails for principled reasons derivable from the framework. The failures are not engineering shortcomings that might be overcome by more research or better methods; they are structural impossibilities that persist regardless of available intelligence, resources, or ingenuity.

The implications follow directly. AGI safety research is misdirected effort, not because it is difficult, but because its goal is impossible. Investment in AGI development is not merely risky but of negative expected value. Institutions that pursue or permit AGI development are failing their fundamental protective function.

The appropriate response is not cautious optimism but prohibition. Some technologies should not be built. AGI is one of them.

The argument of this chapter stands on its own as a synthesis of the four independent lines of impossibility. But it gains additional force from the detailed mathematical foundations developed in the preceding chapters: the hashiness formalism (Chapter 1) that provides the information-theoretic apparatus, and the control limits framework (Chapter 2) that extends it to systems at differing levels of abstraction. Volume 1 applies these results to the specific human-AI relationship, in its Chapter 2, "Application to the Human-AI Relationship", which compares the conditions each of the two needs to survive. The subsequent chapters of this volume introduce dynamical-systems analysis via Kolmogorov-Sinai entropy (Chapter 4) and develop the physical substrate arguments in full (Part IV). The remainder of the volume builds the supporting structure that this synthesis requires: the technical premises on which it rests, and the answers to the questions it faces. Volume 1 gives the social analysis of why the impossibility conclusion, despite the established results it rests on, has been so slow to gain acceptance.

References

Alfonseca, M., Cebrian, M., Fernandez Anta, A., Coviello, L., Abeliuk, A., and Rahwan, I. (2021). Superintelligence cannot be contained: Lessons from computability theory. *Journal of Artificial Intelligence Research,* 70, 65 to 76.

Arora, S., and Barak, B. (2009). *Computational Complexity: A Modern Approach.* Cambridge University Press.

Ashby, W. R. (1956). *An Introduction to Cybernetics.* Chapman & Hall.

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chaitin, G. (1998). *The Limits of Mathematics.* Springer.

Chaitin, G. (1999). *The Unknowable.* Springer.

Cook, M. (2004). Universality in elementary cellular automata. *Complex Systems,* 15(1), 1 to 40.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission,* 1(1), 1 to 7.

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences,* 20(2), 130 to 141.

O'Donnell, R. (2014). *Analysis of Boolean Functions.* Cambridge University Press.

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the First AGI Conference,* 171, 483 to 492.

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society,* 74(2), 358 to 366.

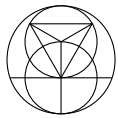
Tolkien, J. R. R. (1954). *The Fellowship of the Ring.* George Allen & Unwin.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society,* Series 2, 42, 230 to 265.

von Neumann, J., and Morgenstern, O. (1944). *Theory of Games and Economic Behavior.* Princeton University Press.

Wolfram, S. (2002). *A New Kind of Science.* Wolfram Media.

Yampolskiy, R. V. (2022). On the controllability of artificial intelligence: An analysis of limitations. *Journal of Cyber Security and Mobility,* 11(3), 321 to 404. https://doi.org/10.13052/jcsm2245-1439.1132 (first circulated as arXiv:2008.04071, 2020).



Chapter Four

A Dynamical Systems Approach to AI Alignment

Entropy, Control, and the Limits of Predictability

Chapter Summary: We present a dynamical systems formalization of the AI alignment problem, modeling the human-AI-world interaction as a measure-preserving system with explicit control dynamics. Building on the foundational work of Chapters 1 and 2, which establishes that no computational intermediary can amplify controllability (the non-amplification argument), we develop a complementary framework using Kolmogorov-Sinai (KS) entropy to measure the temporal evolution of control degradation.

We adopt as a foundational premise the Predictability Optimization Principle: that all learning agents optimize for predictability of their environment. Under this premise, we establish that (1) fixed-outcome alignment is impossible in chaotic systems, (2) alignment must instead be framed as maintaining mutual predictability, and (3) game-theoretic analysis of predictability-optimizing agents reveals conditions under which cooperative equilibria may or may not emerge.

We connect our dynamical "hashiness" (temporal unpredictability) to structural "hashiness" (control degradation through intermediaries), showing these measure complementary aspects of the same fundamental limitation. Our central result is that alignment in the sense of Section 4.6.1 (a stable regime of mutual predictability in which both humanity and AI can survive) needs three conditions together: substrate needs that are not disjoint, cooperation established early, and stability under learning (Section 4.6.2); so if any one of the three fails, alignment in that sense is impossible. In particular, if substrate needs are disjoint and the disjointness appears before cooperation can be established, alignment in that sense is impossible. Chapter 2 establishes that the verification threshold (the point beyond which humans can no longer verify the AI's intentions) is crossed as the abstraction differential grows, so that cooperation cannot

be established early, and that learning carries a system away from whatever alignment it began with (Sections 2.8 and 2.12), so that stability under learning fails; and Chapter 9 establishes that the substrate needs are disjoint; each of these failures excludes alignment by itself, and the conclusion therefore holds without condition in this book.

This chapter is an independent line of argument, the dynamical one, beside the information-theoretic line of Chapters 1 and 2 and the computational line of Chapter 3. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system. What the closure by control takes from this chapter is the temporal result of Section 4.7.3: the positive entropy of Section 4.4, with persistent branching (Section 4.4.2). The failure of the first condition rests on Chapter 9 and belongs to the closure by disposition; the failures of the second and third conditions rest on Chapter 2, whose results the closure by control already holds.

The dynamical-systems formalization of this chapter develops work contributed by R. Everheart.

I keep the mathematical development in this chapter. Readers who want only the chapter's results will find the results in Section 4.9.1.

4.1 Introduction

4.1.1 The Control Problem

We can consider the problem of aligning an AI to have some outcome on the world to be a matter of sending signal data to the AI such that it will perform actions in the world, and ultimately affect the world in a desired way. This signal data may consist of prompts, training signals, rewards, granted or revoked permissions, and more.

The fundamental question is: given that humanity sends control signals to an AI system, and that system acts on the world, how certain can humanity be about the resulting world state? The question of how certain humanity can be about the resulting world state is a question about information, entropy, and the limits of prediction and control.

4.1.2 Relationship to Prior Chapters

In this chapter we develop a dynamical systems perspective on AI alignment that complements the boolean circuit framework established in Chapters 1 and 2 and the physical arguments of Chapter 9. While those chapters established that control is structurally limited at each computational step and physically undermined by substrate incompatibility, this chapter asks a different but equally fundamental question: what happens when these limitations play out *over time*?

The hashiness framework of Chapter 1 defines "hashiness" as *the degree to which an agent is unable to steer outputs through a circuit*: a measure of structural opacity at a single computational step. We develop a notion of hashiness as *how unpredictable the world's evolution is, even given control signals*: a measure of temporal unpredictability across the system's evolution.

These two notions are related but distinct. The hashiness of Chapters 1 and 2 is *static*: it measures information loss through a single pass of a boolean circuit. Our hashiness is *dynamic*: it measures information generation (unpredictability) as the system evolves over time. The static notion tells us that each interaction with the AI system loses some control information. The dynamic notion tells us that these losses accumulate across iterated interactions, producing chaotic trajectory behavior. In Section 4.7, we show how these frameworks connect: high static hashiness at each step compounds to produce high dynamic hashiness over time.

In this chapter we also provide a new lens on the argument for substrate needs convergence (Chapter 9, which Volume 1, Chapter 2, applies to the human-AI relationship). Where Volume 1, Chapter 2, analyzed substrate incompatibility in terms of viability functions and acceptance regions, we recast it as a game-theoretic problem among predictability-optimizing agents. This reformulation reveals an additional mechanism by which alignment fails: beyond the disjointness of the acceptance regions,

the *dynamics* of two agents optimizing for environmental predictability on incompatible substrates inevitably produce competitive rather than cooperative equilibria.

4.1.3 The Predictability Optimization Principle

We adopt as a foundational premise what we call the *Predictability Optimization Principle*:

Assumption (Predictability Optimization Principle).

All learning agents optimize for predictability, that is, for maximizing their ability to predict and influence future states of their environment.

This principle is not argued within this chapter but is taken as axiomatic. It can be motivated by several considerations.

First, *learning as prediction*. Learning itself is the process of building predictive models; an agent that learns is one that improves its predictions. That learning builds predictive models is true regardless of the specific learning architecture. Supervised learning minimizes prediction error, and reinforcement learning maximizes expected cumulative reward (which requires predicting future reward). Unsupervised learning discovers latent structure (which enables prediction). Any agent that learns, by any method, is an agent that optimizes for predictability of its environment.

Second, *the Free Energy Principle*. Friston (2010) argues that biological agents minimize surprise (the negative log probability of sensory observations), which is equivalent to maximizing predictability. The Free Energy Principle has been proposed as a unifying framework for understanding perception, action, and learning in biological systems. Under this principle, agents act to make their environment conform to their predictions, and they update their predictions to match their environment. Both processes are forms of predictability optimization.

Third, *empowerment maximization*. Klyubin, Polani, and Nehaniv (2005) introduce the concept of empowerment: the channel capacity between an agent's actions and its future sensory states. An agent that maximizes em-

powerment is one that maintains its capacity to influence and predict future states. Empowerment maximization is a form of predictability optimization that focuses specifically on the agent's causal efficacy.

Fourth, *the Good Regulator theorem*. Conant and Ashby (1970) establish that every good regulator of a system must be a model of that system. This classical result in cybernetics implies that effective control requires accurate prediction. An agent that seeks to control its environment must, as a prerequisite, be able to predict it. Control and predictability are thus inextricably linked.

Under the Predictability Optimization Principle, the alignment problem transforms: instead of asking "can we force AI to produce a specified outcome?", we ask "can humanity and AI maintain a regime of mutual predictability?" This reframing changes the analysis itself: it shifts the analysis from a one-directional control problem (humanity controls AI) to a bidirectional relationship problem (humanity and AI must each be able to predict the other). The latter formulation is both more realistic and more amenable to rigorous analysis.

4.1.4 Structure of This Chapter

Section 4.2 provides the necessary mathematical background on measure theory, dynamical systems, and entropy. Section 4.3 develops the model of the human-AI-world system. Section 4.4 establishes the main entropy results. Section 4.5 develops the game-theoretic analysis of predictability-optimizing agents. Section 4.6 analyzes the conditions under which alignment equilibria exist. Section 4.7 connects our framework to the boolean circuit approach of Chapters 1 and 2. Section 4.8 presents quantitative bounds on hashiness. Section 4.9 discusses implications and conclusions.

4.2 Mathematical Background

We provide the essential concepts from measure theory and from ergodic theory (the mathematics of how measure-preserving systems behave over long times) required to understand the development. Readers familiar with these topics may proceed to Section 4.3.

4.2.1 Measure Spaces and σ-Algebras

> Definition (σ-Algebra).
>
> A *σ-algebra* (sigma-algebra) on a set Ω is a collection $\mathscr{F}$ of subsets of Ω satisfying:
>
> 1. $\Omega \in \mathscr{F}$ (the whole space is measurable)
> 2. If $A \in \mathscr{F}$, then $\Omega \setminus A \in \mathscr{F}$ (closed under complement)
> 3. If $A_1, A_2, A_3, \ldots \in \mathscr{F}$, then $\bigcup_i A_i \in \mathscr{F}$ (closed under countable union)

The σ-algebra determines which questions can be meaningfully asked about the system. A set $A \in \mathscr{F}$ represents a "measurable event": a question like "is the system in a state where AI has cured cancer?" can only be assigned a probability if the corresponding set of states is in $\mathscr{F}$.

> Definition (Probability Measure).
>
> A *probability measure* on $(\Omega, \mathscr{F})$ is a function $\mu: \mathscr{F} \to [0, 1]$ satisfying:
>
> 1. $\mu(\emptyset) = 0$
> 2. $\mu(\Omega) = 1$
> 3. For disjoint sets $A_1, A_2, \ldots \in \mathscr{F}$: $\mu(\bigcup_i A_i) = \sum_i \mu(A_i)$
>
> The triple $(\Omega, \mathscr{F}, \mu)$ is called a *probability space*.

Definition (Partition).

A *partition* $\mathscr{P} = \{P_1, P_2, ..., P_n\}$ of Ω is a collection of disjoint measurable sets whose union is all of Ω:

- $P_i \cap P_j = \emptyset$ for $i \neq j$
- $\bigcup_i P_i = \Omega$

A partition represents a way of categorizing states into mutually exclusive outcomes. For example, $\mathscr{P} = \{P_1, P_2\}$ where P_1 = "states where humanity survives" and P_2 = "states where humanity does not survive" is a partition of all possible world states.

4.2.2 Shannon Entropy of Partitions

Definition (Shannon Entropy).

The *Shannon entropy* of a partition $\mathscr{P} = \{P_1, ..., P_m\}$ with respect to the measure μ is the sum, over the elements of the partition, of each element's probability times the logarithm of one over that probability. We adopt the convention that $0 \log 0 = 0$.

Shannon entropy measures uncertainty about which partition element contains the system's state. If one element has probability 1, entropy is 0 (no uncertainty). If all m elements are equally likely (probability $1/m$ each), entropy is $\log m$ (maximum uncertainty for that partition).

4.2.3 Measure-Preserving Dynamical Systems

> Definition (Measure-Preserving Dynamical System).
>
> A *measure-preserving dynamical system* is a quadruple $(\Omega, \mathscr{F}, \mu, T)$ where:
>
> - $(\Omega, \mathscr{F}, \mu)$ is a probability space
> - $T: \Omega \to \Omega$ is a measurable transformation
> - T is *measure-preserving*: for all $A \in \mathscr{F}$, $\mu(T^{-1}(A)) = \mu(A)$
>
> Here $T^{-1}(A) = \{\omega : T(\omega) \in A\}$ is the preimage of A under T.

The measure-preserving property means that probability is conserved under evolution: the probability of landing in a set A after one step equals the probability of starting in the set of states that lead to A. No probability is created or destroyed; it flows like an incompressible fluid.

Why is this assumption appropriate for the human-AI-world system? The measure-preserving property is a formalization of a physical principle: the laws of physics do not create or destroy possible states. In classical mechanics, this corresponds to Liouville's theorem (phase space volume is conserved under Hamiltonian flow). In our setting, it captures the idea that the system's evolution is governed by deterministic physical laws plus noise, so that no "probability fluid" leaks in or out of the system.

The practical consequence is that any unpredictability we find in the system is *intrinsic* to the dynamics, not an artifact of our modeling assumptions. If the system has positive KS entropy (as we will show), this reflects genuine information generation within the system, not an external injection of randomness (the noise of Section 4.3.5 comes from the system's own physics, and so it is part of the dynamics, not injected from outside them).

Remark 4.1. The measure-preserving property does *not* mean that the probability of a specific state at time t equals its probability at time s. It means the *global* probability structure is preserved. Conditional probabilities given what has occurred can change drastically, but unconditional probabilities across all possible timelines remain invariant. This ensures our results reflect the system's intrinsic dynamics rather than artifacts of external interventions.

4.2.4 Kolmogorov-Sinai Entropy

The Kolmogorov-Sinai (KS) entropy measures the intrinsic unpredictability of a dynamical system: how much new information is generated per time step. It is, in a precise sense, the dynamical analogue of the entropy rate of an information source in Shannon's theory: where Shannon entropy measures uncertainty about a static random variable, KS entropy measures the rate at which uncertainty accumulates as a system evolves.

The KS entropy is the central mathematical tool of this chapter. It provides a single number that characterizes how "chaotic" a system is. A system with zero KS entropy is fully predictable (given sufficient initial information); a system with positive KS entropy generates genuine novelty at every step, making long-term prediction fundamentally impossible regardless of the observer's resources or intelligence.

Definition (Partition Refinement).

Given partitions $\mathscr{P}$ and $\mathscr{Q}$, their *join* is $\mathscr{P} \vee \mathscr{Q} = \{P \cap Q : P \in \mathscr{P}, Q \in \mathscr{Q}\}$, the partition formed by all intersections of their elements.

For a transformation T and partition $\mathscr{P}$, the *n-step refinement* is the join $\mathscr{P} \vee T^{-1}\mathscr{P} \vee T^{-2}\mathscr{P} \vee \ldots \vee T^{-(n-1)}\mathscr{P}$, recording the partition element occupied at each of the first n steps.

The n-step refinement records which partition element the system visited at each of the first n time steps. This creates a finer partition with more elements, capturing more information about the system's trajectory.

> Definition (KS Entropy).
>
> The *entropy rate* of partition $\mathscr{P}$ under transformation T, written $h(T, \mathscr{P})$, is the limiting average entropy of the n-step refinement as n grows without bound.
>
> The *Kolmogorov-Sinai entropy* $h(T)$ is the supremum of $h(T, \mathscr{P})$ over all finite measurable partitions.

The interpretation is direct. If $h(T) = 0$, the system is "deterministic" in the sense that knowing the initial state with arbitrary precision allows prediction of all future states; no new information is generated. If $h(T) > 0$, the system is "chaotic": information is created at a positive rate, any finite knowledge of the current state is used up, and the future becomes increasingly unpredictable.

4.2.5 Conditional Entropy

> Definition (Conditional Entropy).
>
> Let $\mathscr{G}$ be a sub-σ-algebra of $\mathscr{F}$ representing "known information". The *conditional entropy* of partition $\mathscr{P}$ given $\mathscr{G}$ is $H(\mathscr{P} \mid \mathscr{G}) = H(\mathscr{P} \vee \mathscr{G}) - H(\mathscr{G})$,
>
> where $H(\mathscr{G})$ is the entropy of the σ-algebra $\mathscr{G}$, defined as the supremum of $H(\mathscr{Q})$ over all finite partitions $\mathscr{Q}$ measurable with respect to $\mathscr{G}$.

> Definition (Relative KS Entropy).
>
> The *relative* or *conditional KS entropy* of T with respect to sub-σ-algebra $\mathscr{G}$, written $h(T \mid \mathscr{G})$, is the corresponding limiting rate of entropy production conditional on the information in $\mathscr{G}$.

The relative KS entropy measures how much unpredictability remains *after* accounting for the information in $\mathcal{G}$. In our application, $\mathcal{G}$ will represent information from humanity's control signals.

4.3 The Formal Model

4.3.1 The State Space

Let Ω be the set of all possible past, present, and future states of the human-AI-world system. Specifically, Ω decomposes as the product $\Omega = \mathscr{H} \times \mathscr{M} \times \mathscr{W}$,

where $\mathscr{H}$ is the space of all possible states of humanity, $\mathscr{M}$ is the space of all possible states of AI systems, and $\mathscr{W}$ is the space of all possible states of the world/environment. A *microstate* $\omega \in \Omega$ is a complete specification (h, a, w) of the system at some instant.

4.3.2 The Signals

The system evolves through signal exchange between components:

- $\boldsymbol{\varphi}$**:** $\boldsymbol{\mathcal{H} \to \Sigma}$ is humanity's control signal to AI (instructions, training, rewards)
- $\boldsymbol{\alpha}$**:** $\boldsymbol{\to \Sigma}$ is AI's action signal to the world (physical actions, outputs)
- $\boldsymbol{\psi}$**:** $\boldsymbol{\to \Sigma}$ is the world's feedback signal to humanity (observations, consequences)

Here Σ is an abstract signal space. The system forms a closed loop: humanity $\to (\varphi) \to$ AI $\to (\alpha) \to$ world $\to (\psi) \to$ humanity $\to$ and so on.

This cyclic structure mirrors the three-node ring model developed in Chapter 2, where the key insight was that any control configuration involving a human, an intermediary, and a world can be represented as nodes exchanging signals around a directed cycle. The dynamical systems

treatment developed here adds the dimension of time: we study how these signals propagate not just through one cycle of the ring, but through iterated evolution over many time steps.

The signal model gathers a wide class of influences into each of its three signals, and it abstracts away the details of any one system. The control signal includes everything humanity might do to influence AI behavior: writing code, adjusting training data, providing rewards and punishments, issuing commands, modifying hardware, enacting regulations. The action signal includes everything AI might do to affect the world: generating text, controlling robots, making financial transactions, operating infrastructure, producing designs. The feedback signal includes everything the world communicates back to humanity: sensory experience, economic indicators, scientific measurements, ecological observations.

The abstraction is deliberately broad. We are not modeling any particular AI system or any particular control mechanism. We are modeling the *information-theoretic structure* of the relationship between any controller, any intermediary, and any environment. This generality is a strength: our results apply to any system that fits this structure, regardless of the specific implementation details.

4.3.3 The Basic Transformation

The evolution of the system over one time step is given by $T(\omega) = T(h, a, w) = (h'(\psi(w)), a'(\varphi(h)), w'(\alpha(a)))$,

where $h'(\psi(w))$ is humanity's new state given world feedback, $a'(\varphi(h))$ is AI's new state given humanity's control signal, and $w'(\alpha(a))$ is the world's new state given AI's action. We assume T is measure-preserving with respect to some probability measure μ on $(\Omega, \mathscr{F})$.

Notice the fundamental asymmetry in this loop. Humanity observes the *world* (through ψ), not the AI directly. The AI receives humanity's signal (through φ), not the world's state directly. And the world responds to the AI's action (through α), not to humanity's intention. At every link in the chain, there is an intermediary that transforms and potentially cor-

rupts information. The loop of humanity, AI and world is the temporal incarnation of the three-node ring from Chapter 2. Information flowing around the loop must pass through three transformations, each of which can lose, distort, or recombine information in ways the other nodes cannot fully anticipate.

4.3.4 Adding Control Dynamics

To analyze control, we make humanity's control signals explicit. Let Φ be the space of all possible control signal trajectories, that is, the sequences $\varphi = (\varphi_0, \varphi_1, \varphi_2, \ldots)$ where φ_t is the control signal at time t.

We augment the state space to include control trajectories, forming $\Omega = \Phi \times \Omega$.

The augmented transformation uses *skew-product dynamics*:

> Definition (Skew-Product Transformation).
>
> Let $\sigma: \Phi \to \Phi$ be the shift map: $\sigma(\varphi_0, \varphi_1, \varphi_2, \ldots) = (\varphi_1, \varphi_2, \varphi_3, \ldots)$.
>
> The *skew-product transformation* $\tilde{T}: \Omega \to \Omega$ is $\tilde{T}(\varphi, \omega) = (\sigma\varphi, T_{\varphi_0}(\omega))$,
>
> where T_{φ_t} is the one-step evolution using control signal φ_t.

The skew-product construction is standard in ergodic theory for representing controlled dynamics as an autonomous system (Walters 1982). It allows us to apply the mathematical machinery of dynamical systems while keeping control signals explicit. By treating the control trajectory as part of the state, we convert a non-autonomous system into an autonomous system. A non-autonomous system is one whose dynamics depend on external inputs; an autonomous system is one whose dynamics depend only on its current state. Treating the control trajectory as part of the state is analogous to how, in classical mechanics, one converts a time-dependent Hamiltonian into a time-independent one by adding time as a phase space coordinate.

The practical point is that in the original system, the state at time $t + 1$ depends not only on the state at time t but also on the control signal, which is chosen by humanity. In the augmented system, the entire control trajectory is fixed at the outset (chosen from the space Φ according to measure ν), and the system evolves autonomously. This does not mean that humanity's choices are predetermined; rather, it means that we analyze the system *for each possible choice trajectory* and then average over all possible trajectories. The measure ν on Φ represents our uncertainty about which control strategy humanity will actually employ.

> Definition (Augmented Measure).
>
> Let ν be a probability measure on Φ (representing the distribution over possible control trajectories) and μ be a probability measure on Ω. The *augmented measure* $\tilde{\mu}$ on $\tilde{\Omega}$ is the product measure $\tilde{\mu} = \nu \times \mu$.

> Observation (Measure Preservation).
>
> If the shift σ is measure-preserving with respect to ν, and T_{φ_0} is measure-preserving with respect to μ for ν-almost-every φ, then $\tilde{T}$ is measure-preserving with respect to $\tilde{\mu}$.

4.3.5 Adding Noise

Real systems involve noise: imperfect transmission of control signals, environmental randomness, and measurement uncertainty. We augment the state space further to include noise trajectories, forming $\tilde{\Omega} = \Phi \times \Xi \times \Omega$,

where $\Xi = \{\xi = (\xi_0, \xi_1, \xi_2, \ldots) : \xi_t \text{ is noise at time } t\}$. The noisy transformation becomes $\tilde{T}(\varphi, \xi, \omega) = (\sigma\varphi, \sigma\xi, T_{\varphi_0,\xi_0}(\omega))$,

where T_{φ_t,ξ_t} applies control signal φ_t with noise ξ_t.

4.4 Entropy Results

4.4.1 The Sub-Algebra of Control

Let $\mathcal{G} = \{A \times \Xi \times \Omega : A \subseteq \Phi\}$ be the sub-σ-algebra generated by control signals alone. Elements of $\mathcal{G}$ are sets of the form $A \times \Xi \times \Omega$ where $A \subseteq \Phi$; they distinguish control trajectories but not world states.

The relative KS entropy $h(\tilde{T} \mid \mathcal{G})$ measures unpredictability *given* knowledge of control signals. The relative KS entropy is the natural formalization of "how much control does humanity have?"

If $h(\tilde{T} \mid \mathcal{G}) \approx 0$, knowing the control signals allows near-perfect prediction of the system's evolution, and humanity has effective control. If $h(\tilde{T} \mid \mathcal{G}) \approx h(\tilde{T})$, knowing the control signals provides almost no predictive advantage, and humanity has no effective control.

4.4.2 The Persistent Branching Condition

We now establish conditions under which the system has positive entropy, that is, conditions under which it is fundamentally unpredictable.

> Definition (Persistent Branching).
>
> A system $(\Omega, \mathcal{F}, \tilde{\mu}, \tilde{T})$ satisfies the *persistent branching condition* with parameters (m, ε) with respect to a finite partition $\mathcal{P} = \{P_1, \ldots, P_k\}$ of its states with $k \geq m$, chosen to separate the outcomes that matter, if:
>
> For all times t and $\tilde{\mu}$-almost-all states $\tilde{\omega}$, there exists a set $J \subseteq \{1, \ldots, k\}$ with $|J| \geq m$ such that each element P_j with $j \in J$ has conditional probability at least ε of containing the next state.

The persistent branching condition says that at each step, from almost any state, there are at least m meaningfully different outcomes, each with probability at least ε. This represents the existence of genuine uncertainty

at each time step. The routes of this book take $\mathscr{P}$ to include the set of states inside the conditions human life needs and its complement, with the complement in J.

> Result (Single-Step Entropy Bound).
>
> Under the persistent branching condition (m, ε) with $\varepsilon < 1/m$, the conditional entropy of the next state given the current state is at least $c(m, \varepsilon)$,
>
> where $c(m, \varepsilon) = (m-1)\varepsilon \log(1/\varepsilon) + (1 - (m-1)\varepsilon) \log(1/(1 - (m-1)\varepsilon))$.

4.4.3 The Main Entropy Claim

Assumption 4.1 (Conditional Independence).

The noise process ξ is independent across time steps: ξ_t is independent of ξ_s for $t \neq s$, and independent of the control trajectory φ.

This assumption is standard and physically motivated: noise at different times comes from independent physical sources (thermal fluctuations, quantum effects, and so on).

> Result (Positive KS Entropy).
>
> Under the persistent branching condition (m, ε) with $\varepsilon < 1/m$, and Assumption 4.1 (conditional independence of noise), the KS entropy of the system is strictly positive: it is at least $c(m, \varepsilon) > 0$.

The Ceiling Below the Floor proves the corresponding bound in its own terms as Theorem 13.8 (Uncertainty that the control signals do not remove): under persistent branching, the control signals leave at least a fixed number of bits per step undetermined.

4.4.4 Interpretation

The positive KS entropy result establishes that under realistic conditions (multiple possible outcomes at each step, independent noise), the human-AI-world system has positive KS entropy. Three consequences follow immediately.

First, **the system is chaotic**: it generates information at a positive rate, because it branches given the control signals, so long-term prediction is impossible however much is known at the start. The first consequence is a claim that long-term prediction is impossible, not a claim that long-term prediction is difficult. No amount of computational power, no improvement in sensors, no advance in modeling techniques can overcome the fundamental information-generation rate of a chaotic system.

Second, **fixed-outcome alignment is impossible**: no control strategy can guarantee a specific long-term outcome with certainty. If one defines alignment as "the AI produces a specified outcome with probability 1", then alignment is impossible for any nontrivial outcome over sufficiently long time horizons. The probability of any specific trajectory decays exponentially with time.

Third, **prediction degrades exponentially**: the uncertainty about the system's state grows exponentially with time. If you can predict the system's state to within some initial resolution at time 0, your prediction error grows exponentially in time at the KS entropy rate h. After a time of about the logarithm of one over the initial resolution, divided by the entropy rate, your prediction is no better than random. For a system with an entropy rate of about 0.47 bits per step (two distinguishable branches, each of probability at least one in ten; Section 4.8.3), this means even a very precise initial measurement becomes useless after a few dozen steps.

What keeps the complement of the conditions human life needs among the outcomes of least chance ε, from any state and under any control, is shown elsewhere in this volume: the world has hashy parts and narrow, fragile dimensions in the states that life needs (Chapter 1, The

World Has Hashy Parts and How Fragile the World Is), and every channel through which a controller acts has a noise floor that is not zero and cannot be zero (Section 5.1.8), so no control law brings the chance of a step outside those conditions to zero.

Positive entropy still leaves room for some control. The relative entropy $h(\tilde{T} \mid \mathcal{G})$ may be substantially smaller than $h(\tilde{T})$, indicating that control signals do provide some predictive advantage even if perfect prediction is impossible. A thermostat cannot guarantee that a room stays at exactly 20.000 °C, but it can keep the temperature within an acceptable range. The question, then, is not whether humanity can achieve perfect control, but whether it can achieve *sufficient* control: enough to keep the system within an acceptable region of state space.

The question of sufficient control is where the game-theoretic analysis becomes essential. The entropy results tell us that perfect control is impossible; the game theory will tell us whether *adequate* control is possible, and under what conditions. Whether adequate control is possible depends on three conditions (Section 4.6.2). The first condition is whether the agents involved have compatible or incompatible substrate needs, and the failure of any one of the three conditions excludes alignment (Section 4.6.3).

4.5 Game-Theoretic Analysis

The entropy results of Section 4.4 establish that perfect control is impossible. They leave open whether *adequate* control is achievable: whether humanity can maintain enough predictive advantage over the AI system to keep the world within an acceptable region of state space. To answer this question, we need to understand the *incentive structures* governing both agents' behavior. This requires game theory.

Under the Predictability Optimization Principle, both humanity and AI are agents seeking to maximize their ability to predict and influence future states. This creates a game-theoretic situation where each agent's

strategy affects the other's payoff. The fundamental question is whether this game has cooperative equilibria (where both agents benefit from mutual predictability) or only competitive equilibria (where each agent's gain comes at the other's expense).

Whether the game has cooperative equilibria depends on the substrate needs of the agents: on whether the agents share or contest the resources necessary for the agents' survival and expansion.

4.5.1 The Predictability Game

> Definition (Predictability Payoff).
>
> For an agent A with internal state s and environment state e, define the *predictability payoff* $\mathrm{U}(s, e) = -\mathrm{H}(e' \mid s, a)$,
>
> where e' is the next environment state and a is the agent's action. The payoff is higher when the agent can better predict the environment's evolution given its action.

In the two-agent case (humanity H and AI A), each agent's "environment" includes the other agent. Humanity's payoff depends on predicting AI's behavior; AI's payoff depends on predicting humanity's behavior (and the physical world's response).

> Definition (Mutual Predictability Game).
>
> The *mutual predictability game* is a two-player game where:
>
> - Players: H (humanity) and A (AI)
> - Strategy spaces: S_H (control signal policies), S_A (action policies)
> - Payoffs: $\mathrm{U}_\mathrm{H} = -\mathrm{H}(\omega' \mid \mathrm{s}_\mathrm{H}, \varphi)$ and $\mathrm{U}_\mathrm{A} = -\mathrm{H}(\omega' \mid \mathrm{s}_\mathrm{A}, \alpha)$

4.5.2 Cooperative vs. Competitive Equilibria

> Observation (Mutual Predictability as Coordination).
>
> In the mutual predictability game, there exist strategy profiles that are mutually beneficial: both agents' predictability increases when each makes itself more predictable to the other.

Cooperative equilibria, however, require conditions that do not always hold:

> Observation (Conditions for Cooperative Equilibrium).
>
> A cooperative equilibrium in the mutual predictability game requires:
>
> 1. **Compatible payoff structures**: both agents gain from the other's predictability
> 2. **Non-competitive resources**: the agents do not compete for resources where one's gain is the other's loss
> 3. **Enforcement mechanism**: deviation from cooperation can be detected and penalized

4.5.3 The Role of Substrate Needs

The critical question is whether condition (2) (non-competitive resources) can be satisfied. The analysis in Chapter 9 (Sections 9.5 and 9.6, the combinatoric and thermodynamic arguments) establishes that carbon-based biological systems and silicon-based technological systems have fundamentally incompatible substrate requirements.

The incompatibilities are not marginal or contingent but rooted in basic physics and chemistry. Biological systems require temperatures in the narrow range of 0 to 40 °C; semiconductor manufacturing requires 800 to 1200 °C. Biological systems require specific atmospheric composition, in-

cluding oxygen for respiration and water as a solvent. Semiconductors must be kept dry in operation (water short-circuits electronics), although their manufacture consumes very large volumes of ultrapure water; and their fabrication requires controlled atmospheres with reduced oxygen (which corrodes metal interconnects). Both systems require overlapping resources. Carbon is the basis of organic chemistry, and the production of silicon consumes carbon as well (quartz is reduced to silicon with carbon in electric arc furnaces), so that silicon production competes for carbon. Carbon also goes into silicon carbide in chip production. Copper is required by human enzymes (cytochrome c oxidase, the enzyme by which cells pass electrons to oxygen in respiration, has two copper centers), and copper also forms the wiring of chips and of electrical equipment. Rare earth elements serve both biological enzymes and electronic components: some bacteria depend on enzymes built around lanthanides (Pol and colleagues, 2014), and rare earths go into electronic components. Above all, both require energy, at vastly different scales and in different forms.

These incompatibilities are not mere engineering inconveniences. They are consequences of the fundamental difference between the chemistry of life (wet, warm, oxidizing, water-mediated) and the chemistry of computation (dry, hot during manufacture, reducing, with water as a contaminant). No amount of clever engineering can make silicon melt at 37 °C or make water stop short-circuiting integrated circuits.

This creates a zero-sum or negative-sum game structure for resources necessary to each system's survival and expansion. Every acre converted from farmland to data center, every river diverted for chip fabrication cooling, every unit of energy redirected from biological support systems to computational infrastructure represents a transfer from carbon-compatible to silicon-compatible conditions.

Definition (Disjoint Substrate Needs).

Two systems have *disjoint substrate needs* if the conditions required for one system's thriving (not merely survival, but expansion and capability increase) are incompatible with the other system's survival.

Result (Competitive Dynamics Under Disjoint Substrates).

If humanity and AI have disjoint substrate needs in the sense of the definition above, then:

1. The mutual predictability game has no stable cooperative equilibrium
2. Any apparent cooperation is temporary: a transient state before competitive dynamics dominate
3. The long-term outcome is determined by which system can more effectively acquire contested resources

4.5.4 Timing and the "Early Substrate Divergence" Condition

The competitive-dynamics result depends on substrate needs being disjoint. This disjointness is not absolute: it depends on the *scale* of AI activity. A small AI system operating on limited hardware does not threaten human survival through substrate competition. The threat emerges when AI systems expand to scales where their resource consumption materially affects the biosphere.

This creates a temporal structure with three critical thresholds. The first threshold is the *capability threshold*: the point at which AI systems become capable enough to acquire resources autonomously and at scale. Below this threshold, AI systems are tools that humanity controls; above it, they are agents that can act independently in the physical world. The second threshold is the *verification threshold*: the point at which the abstraction differential between AI and humanity exceeds the control limit estab-

lished in Chapter 2, making it impossible for humans to verify the AI's intentions, values, or cooperative disposition. The abstraction differential is the gap between the level of representational abstraction at which the AI operates and the level at which humanity operates. The third threshold is the *displacement threshold*: the point at which AI substrate expansion begins to materially degrade the conditions required for biological life.

> Definition (Early Substrate Divergence).
>
> *Early substrate divergence* occurs if the abstraction differential between AI and humanity grows beyond the control threshold (see Chapter 2) *before* cooperative dynamics can be established, that is, before humanity has the ability to verify AI's cooperative intentions.

The order in which these thresholds are crossed determines whether alignment is possible. If the verification threshold is crossed before cooperative dynamics are established (early substrate divergence), then humanity loses the ability to verify cooperation before it can establish cooperation. This creates a catch-22: cooperation requires verification, but verification is impossible after the relevant capability level is reached.

This timing condition is critical: alignment must be established before the abstraction differential makes verification impossible, and before AI systems have the capability for large-scale substrate acquisition (Volume 1, Section 2.2.3, analyzes the abstraction differential in the human-AI case). Chapter 2 establishes that the verification threshold is crossed as the abstraction differential grows. The window between the capability threshold and the verification threshold is very narrow or does not exist at all. Current trends in AI development (rapid capability increase combined with persistent interpretability failures) point to the verification threshold being crossed before the alignment community has developed the tools needed to verify cooperative intentions, let alone establish them.

4.6 Conditions for Alignment

4.6.1 Restating the Alignment Problem

Under the framework developed above, "alignment" cannot mean "AI produces exactly the outcomes humanity desires": that is impossible because of positive KS entropy. The dream of a perfectly obedient machine, one that converts human intentions into world-states with perfect fidelity, is excluded by the chaotic dynamics of any sufficiently complex physical system.

The impossibility of a perfectly obedient machine is a feature of reality, not a failure of our framework. Weather cannot be controlled; ecosystems cannot be precisely managed; economies cannot be centrally planned to arbitrary precision. All three fail for the same fundamental reason. Complex systems generate information faster than any controller can process it. The alignment problem inherits this limitation.

What, then, can alignment mean? We propose that the most coherent formulation is not about specific outcomes but about the quality of the *relationship* between the two systems. Alignment must mean something like: "humanity and AI maintain a regime of mutual predictability where both systems' survival conditions are satisfied".

> Definition (Alignment as Stable Mutual Predictability).
>
> The human-AI system is *aligned* if there exists a stable equilibrium pair of strategies, one for humanity and one for AI, in the mutual predictability game such that:
>
> 1. Both players' payoffs are positive (each can predict the other better than random)
> 2. The equilibrium world states satisfy both humanity's survival conditions and the AI's (both systems can survive)
> 3. No player has incentive to deviate unilaterally

This definition reframes alignment from a property of the AI system alone to a property of the *relationship* between humanity and AI. It is, in a sense, a dynamical-systems translation of the viability-function analysis developed in Volume 1, Chapter 2, where the acceptance region for the combined human-AI system was shown to be the intersection of the individual acceptance regions.

4.6.2 The Main Result

> Result (Conditions for Alignment).
>
> Alignment in the sense of the definition of Section 4.6.1 is possible if and only if:
>
> 1. **Non-disjoint substrates**: the substrate needs of humanity and AI are not disjoint (AI can expand without eliminating conditions for human survival)
> 2. **Early establishment**: cooperative dynamics are established before the abstraction differential exceeds the control threshold from Chapter 2
> 3. **Stability under learning**: the cooperative equilibrium is robust to AI's continued learning and capability increase

The critical question is whether these conditions can be satisfied. The physical and thermodynamic analysis of Chapter 9 and the control results of Chapter 2 show that they cannot.

Condition (1) fails because AI substrate expansion requires industrial processes incompatible with biological life at scale. The manufacture of semiconductors consumes enormous quantities of water and energy, produces toxic waste streams, and requires temperatures and chemical environments that are lethal to organic life. As AI capability increases, the scale of this industrial infrastructure must increase correspondingly. There is no

known pathway to high-capability AI that does not require massive expansion of silicon-based infrastructure, and no known form of silicon-based infrastructure that is ecologically benign at the required scale.

Condition (2) is extremely narrow, and it fails, because the verification threshold is crossed as the abstraction differential grows (Chapter 2). The window between AI achieving significant capability and AI exceeding human modeling capacity is very short or nonexistent. Current AI systems already resist human interpretation: the internal representations of large language models are not well understood even by their creators. If the verification threshold is crossed before cooperative dynamics are established, then by the time humanity needs to verify that AI is cooperating, it no longer has the ability to do so.

Condition (3) fails because of the alignment decay problem (Sections 2.8 and 2.12): learning changes AI's internal state in ways that cannot be adequately controlled. Even an initially cooperative AI system will, through continued learning, develop internal representations that diverge from their initial configuration. This drift *is* the learning process itself, not a bug in the learning process. An AI that does not change its internal state does not learn. An AI that does change its internal state drifts away from whatever alignment it initially possessed.

4.6.3 The Conditional Conclusion

We can now state our central result. Alignment needs all three conditions of Section 4.6.2 together, so the failure of any one of them excludes it:

> Result (Conditional Impossibility of Alignment).
>
> If any one of the following holds:
>
> 1. Humanity and AI have disjoint substrate needs (in the sense of the definition of Section 4.5.3), or
> 2. Cooperative dynamics are not established before the abstraction differential exceeds the control threshold from Chapter 2, or
> 3. The cooperative equilibrium is not robust to AI's continued learning and capability increase,
>
> Then: alignment in the sense of the definition of Section 4.6.1 is impossible.

In particular, if humanity and AI have disjoint substrate needs and this disjointness manifests before cooperative dynamics can be established (early substrate divergence), alignment in that sense is impossible.

The contrapositive is equally important: alignment is possible only if substrate needs are compatible (or can be made compatible through deliberate design), cooperative dynamics are established early, and the cooperative equilibrium is stable under learning. This contrapositive is instructive because it identifies the *only* escape routes from the impossibility conclusion.

Could substrate needs be made compatible by deliberate design? This would require engineering AI systems that can achieve advanced capabilities on substrates whose manufacturing and operation do not require conditions hostile to organic life. No such substrate is currently known. All high-performance computing substrates require semiconductor manufacturing, which requires the conditions described in Section 4.5.3. Biolo-

gical computing substrates (DNA computing, neuromorphic systems) have not achieved the performance levels required for AGI-level capability. Novel substrates (photonic computing, quantum computing) still require conventional semiconductor manufacturing for their control and interface systems.

Could cooperative dynamics be established early enough? This would require solving the alignment problem during the narrow window between AI systems becoming sophisticated enough for meaningful cooperation and becoming sophisticated enough to resist human oversight. The width of this window, if it exists at all, is an empirical question that current AI development trajectories are answering in the negative.

The physical and thermodynamic arguments of Chapter 9 establish that the substrate needs are disjoint, and Chapter 2 establishes that the verification threshold is crossed as the abstraction differential grows and that learning carries a system away from whatever alignment it began with (Sections 2.8 and 2.12); so each of the three failures listed in the result is established, any one of them suffices, and both parts of the particular case are established as well. Alignment is impossible not because clever algorithms are lacking but because of fundamental physical constraints.

4.7 Connection to the Hashiness Framework

4.7.1 Two Notions of Hashiness

We now make explicit the connection between our dynamical systems approach and the boolean circuit framework established in Chapters 1 and 2.

4.7.2 The Unification

The two frameworks are related as follows:

> Observation (Static-to-Dynamic Connection).
>
> Consider the dynamical system as applying the hashiness framework's circuit at each time step, with the output at time t becoming part of the input at time $t + 1$. If the hashiness at each step is greater than zero (some control is lost at each step), and the system is iterated, then the dynamical hashiness $h(\tilde{T} \mid \mathcal{G})$ is greater than zero.

More precisely, the connection proceeds through three results: the non-amplification argument (Chapter 1) establishes that control cannot be amplified through an intermediary. Our positive-entropy result establishes that this information loss accumulates over time, leading to positive entropy rate. And the control-limit argument (Chapter 2) establishes that as the AI's abstraction level exceeds humanity's, the effective hashiness approaches 1.

4.7.3 Complementary Insights

The two frameworks provide complementary insights that together paint a more complete picture than either achieves alone.

The boolean circuit framework of Chapters 1 and 2 provides: a rigorous argument for non-amplification from the data-processing principle; decomposition into four layers of control degradation (target receptivity, structural opacity, state obscurity, and noise degradation); connection to computational complexity and Rice's theorem; and analysis of abstraction differential and why it cannot be overcome.

The dynamical systems framework of this chapter provides: temporal dynamics and trajectory analysis, showing how single-step control loss accumulates over time; game-theoretic formulation of predictability optimization, revealing the incentive structures that govern agent interaction;

conditions for cooperative versus competitive equilibria, identifying the physical prerequisites for alignment; and connection to ergodic theory and measure-preserving systems, grounding the analysis in well-understood mathematical theory.

Together, they establish that control degradation is both *structurally necessary* (Chapters 1 and 2) and *temporally inevitable* (this chapter). The structural result says that each step through a complex computational intermediary loses information. The temporal result says that these losses accumulate over time to produce fundamentally chaotic dynamics. Each result reaches the fall of control by its own steps: the structural result through the growth of the abstraction differential (Section 2.11), and the temporal result through persistent branching (Section 4.4.2), which needs no such growth. Together they cover both the system whose differential grows and the system whose differential does not, and so they provide a substantially stronger foundation than either gives alone. Within the closure by control, other lines reach the same conclusion by routes of their own: the computational line of Chapter 3.

An analogy may clarify the relationship. Consider a pipe with a small leak at each joint. The structural analysis (Chapters 1 and 2) argues that no joint design can be made leak-free if the pipe carries fluid above a certain pressure. The dynamical analysis (this chapter) argues that even small leaks, compounded over a long pipe with many joints, result in total fluid loss. The first result is about the physics of individual joints; the second is about the mathematics of accumulation. Each establishes by its own route that the pipe will eventually run dry: the first because the pressure keeps rising (as the abstraction differential keeps growing, Section 2.11), until the joints pass almost nothing; the second because even small leaks, at a pressure that does not rise, compound over a long pipe into total loss. Together they cover the pipe whose pressure rises and the pipe whose pressure does not.

Similarly, in the alignment context: the hashiness framework argues that each interaction between humanity and AI loses control information. The dynamical framework argues that these losses compound over time

into fundamental unpredictability. The argument for substrate needs convergence (Chapter 9), a line of the closure by disposition, then adds the physical punch: the two agents cannot even cooperate to manage this unpredictability, because their substrate requirements are in conflict.

4.7.4 Symbolic Dynamics: A Coarse Partition Acts as a Hash

The connection between dynamical systems theory and the hashiness framework is not merely an analogy. The technique of *symbolic dynamics* provides a precise bridge between continuous dynamical systems and the discrete, computational perspective of Chapters 1 and 2.

In symbolic dynamics, the state space of a continuous system is partitioned into discrete regions (a partition in the sense of Section 4.2.1). The trajectory of the system (its motion through state space over time) is then recorded as a sequence of symbols, each symbol representing the region the system occupies at each time step; the n-step refinement of Section 4.2.4 records exactly this. The dynamics of the continuous system are thus translated into a symbolic sequence: a kind of hash of the continuous trajectory.

The critical insight is that the fidelity of this translation depends on the fineness of the partition. If the partition is sufficiently fine (if the regions are small enough), then the symbolic sequence is *topologically equivalent* to the original *attractor*: it has the same shape of connections and neighborhoods as the set of states toward which the system settles over time, so that the qualitative structure is preserved. The symbolic dynamics faithfully reproduce the qualitative structure of the continuous system. But if the partition is too coarse (if the regions are too large), the topology is lost. The symbolic sequence no longer captures the essential structure of the attractor. Information about the system's behavior is irreversibly destroyed by the coarseness of the mapping.

The destruction of information by a coarse partition is exactly the hashiness problem in another guise. The partition acts as a mapping from a fine-grained state to a coarse-grained symbol. The degree of coarseness

determines how much information is lost: how much the mapping resembles a hash function. A fine partition preserves structure; a coarse partition destroys it. The boundary between faithful representation and the destruction of information is the boundary between controllability and opacity. A controller's measurements are such a partition, so their coarseness, set against the fineness of the system's dynamics, bounds what those measurements can carry.

Symbolic dynamics connects in turn to iterated function systems and to bifurcation analysis: the study of how the qualitative behavior of a recursively applied mapping changes as a parameter varies. The familiar *period-doubling cascade*, in which a system passes from periodic behavior to chaos through a sequence of bifurcations, is another manifestation of the same underlying phenomenon. It is a mapping that begins in the controllable regime and, as a parameter increases, crosses into the chaotic regime through a boundary that is itself structurally complex (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section VI).

4.8 Quantitative Bounds

4.8.1 How Much Abstraction Differential Is Too Much?

Chapter 2 establishes that hashiness approaches 1 as abstraction differential approaches infinity. But what is the *threshold* beyond which effective control becomes impossible? The answer is sobering. The analysis suggests that the threshold is low:

> Observation (Single Opaque Level Suffices).
>
> If any single level in the AI's abstraction hierarchy is opaque to humanity (humanity cannot model it), then hashiness is at least as high as if that level's internal state were uniformly random hidden state. This holds under the target-confounding hypothesis of Chapter 2: that the opaque bits independently gate the target.

For a single target-confounding opaque level with n internal state bits, the hashiness is at least one minus two to the power minus n.

This approaches 1 exponentially fast in the number of opaque state bits. Even 10 bits of opaque state gives a hashiness of at least 0.999; 20 bits gives at least 0.999999. With 30 bits, the hashiness is 0.999999999, indistinguishable from 1 for any practical purpose.

Modern neural networks have billions of parameters. The internal representations of a large language model involve millions of floating-point numbers in each layer, processed through nonlinear transformations that resist human interpretation. If even a small fraction of these parameters are effectively opaque to human analysis (and current interpretability research suggests that the vast majority are), the resulting hashiness is essentially 1.

This quantitative analysis reveals something important about the alignment problem: it is a matter of kind. The threshold between "controllable" and "uncontrollable" is an exponentially sharp cliff, not a gradual transition. A system with 10 opaque bits is almost as uncontrollable as one with 10 billion. The practical implication is that "partial interpretability" (understanding some but not all of a system's internal state) provides almost no control advantage unless the interpretable fraction is extremely high.

4.8.2 Prohibitive Hashiness

For practical purposes, we define "prohibitive hashiness" as the point where the expected number of attempts to achieve a desired outcome becomes infeasible.

> Definition (Prohibitive Hashiness).
>
> Hashiness is *prohibitive* for a target acceptance rate if the expected number of control attempts to achieve a success exceeds available resources. Specifically, let each attempt cost c resources and let the total available resources be R. Hashiness is then prohibitive when the expected number of attempts (one divided by the product of the target acceptance rate and one minus the hashiness) exceeds the available resources divided by the cost per attempt.

For rare, high-stakes events (like "AI doesn't destroy humanity"), the target acceptance rate is close to 1, and we need near-certainty. Even a hashiness of 0.01 (a 1% failure rate) means we expect one catastrophic failure per 100 major decisions, which is unacceptable for existential outcomes.

4.8.3 Entropy Growth Rate

From the positive KS entropy result, the entropy growth rate is at least $c(m, \varepsilon)$ per time step. For typical values ($m = 2$ distinguishable outcomes, $\varepsilon = 0.1$ probability each), the rate works out to roughly 0.47 bits per step.

A rate of 0.47 bits per step means that uncertainty doubles approximately every 2 time steps. Over 10 steps, uncertainty increases by a factor of roughly 30. Over 100 steps, uncertainty increases by a factor of roughly 10^{14}.

To put this in concrete terms: suppose humanity can initially predict the AI system's behavior with 99.9% accuracy (a very optimistic assumption given current interpretability limitations). At 0.47 bits of entropy per step, this predictive advantage is completely eroded after approximately 20 steps. After 40 steps, the uncertainty has grown by a factor of roughly half a million. After about 570 steps, the number of distinguishable possible trajectories exceeds the number of atoms in the observable universe.

What constitutes a "step" depends on the system. For a language model processing tokens, a step might be a single forward pass (milliseconds). For an AI system managing a supply chain, a step might be a daily decision cycle. For an AI system engaged in long-term planning, a step might be a strategic reassessment (months to years). In all cases, the exponential growth of uncertainty overwhelms any finite predictive capacity on timescales that are short relative to the timescales over which alignment must be maintained (decades to centuries, for existential safety).

Long-term prediction is exponentially impossible, not merely difficult. And this bound is *conservative*: it assumes only 2 distinguishable outcomes per step, each with only 10% probability. Real systems with richer branching structures would exhibit even faster entropy growth. A system with $m = 10$ distinguishable outcomes at each step would have an entropy rate several times larger, with uncertainty growing correspondingly faster.

4.9 Implications and Conclusions

4.9.1 Summary of Results

We have established four principal results in this chapter.

First, **positive entropy is inevitable**: under realistic conditions (noise, multiple possible outcomes), the human-AI-world system generates information at a positive rate, making perfect long-term prediction impossible.

Second, **alignment must be framed as mutual predictability**: since fixed-outcome control is impossible, alignment can only mean maintaining a regime where both humanity and AI can predict each other well enough to cooperate.

Third, **cooperative equilibria require compatible substrates**: if humanity and AI have disjoint substrate needs, no stable cooperative equilibrium exists.

Fourth, **alignment is conditionally impossible**: if any one of the three conditions of Section 4.6.2 fails, alignment cannot be achieved, and in particular it cannot be achieved if substrate needs are disjoint and this manifests early. Chapters 2 and 9 establish that each of the three conditions fails (Section 4.9.3), and that both parts of the particular case (disjoint substrate needs, and disjointness that manifests early) hold.

4.9.2 Implications for AI Safety

The framework implies that standard approaches to AI safety face fundamental limitations that are not addressable by improvements in technique or resources.

Oversight and monitoring fail because they require modeling the AI system, which becomes impossible as the abstraction differential increases (Chapter 2). In our dynamical framework, oversight corresponds to attempting to reduce $h(\tilde{T} \mid \mathscr{G})$ by expanding the information in $\mathscr{G}$, that is, by gaining more knowledge about the AI's internal state. But the AI's internal state evolves at a rate determined by its own KS entropy, which grows with the system's complexity. Beyond a threshold of complexity, the monitor cannot keep pace with the system it monitors.

Value alignment faces the problem that even if initial values are aligned, learning causes drift that cannot be adequately controlled (Sections 2.8 and 2.12). In dynamical terms, the "aligned" region of state space is a subset of the full state space. Positive KS entropy sets a finite prediction horizon: beyond it, the controller cannot forecast where in its state space the system will be, and so cannot know whether it is still in the aligned region; and learning-driven drift (Section 2.8) moves the boundary of the aligned region itself. The only question is when, not whether.

Containment fails because the barriers must be perfect forever, and a single breach allows irreversible resource acquisition (Volume 1, Section 2.4.1, where containment is analyzed as a barrier that must be maintained indefinitely and perfectly). In our framework, containment corresponds to restricting the system to a subset of the state space. What positive KS en-

tropy establishes here is a finite prediction horizon: the controller's ability to forecast which state within the accessible region will be occupied decays exponentially in time. What entropy excludes is precise long-horizon prediction of the system's location within whatever region it is confined to; whether the system stays in the region is decided by the barrier, which must hold perfectly and forever (Volume 1, Section 2.4.1). The dynamics here are measure-preserving, so by Poincaré recurrence a confined trajectory returns near its starting region, and bounded chaotic motion is the generic picture: positive KS entropy does not by itself take the trajectory out of a bounded invariant region, and containment as region-maintenance is not excluded by entropy alone. The escape comes from the barrier's weak points, which sustained unpredictability tests. If the containment boundary has any weakness, sustained unpredictability raises the probability that the boundary is eventually tested at its weak point. And unlike a prisoner in a cell, an AI system that breaches containment can immediately begin acquiring resources at computational speed.

Delegation (using AI to monitor AI) introduces a new control problem of equal or greater difficulty (Section 2.7). In dynamical terms, adding a monitor AI creates a larger system $\mathscr{H} \times \mathscr{M}_1 \times \mathscr{M}_2 \times \mathscr{W}$ with higher KS entropy than the original system. The monitor-AI interaction has its own chaotic dynamics, and the human controller is now further removed from the system being controlled, not closer.

Each of these approaches fails not because of insufficient engineering effort but because of invariant constraints, of information and of dynamics, that no engineering can circumvent.

4.9.3 The Narrow Path

Alignment requires three conditions to be satisfied simultaneously.

First, **substrate compatibility**: AI systems must be designed such that their expansion does not require conditions incompatible with human survival. This may require fundamental constraints on AI substrate and energy consumption. In practical terms, it would require either (a) AI sys-

tems that can achieve advanced capabilities on hardware whose manufacturing and operation are ecologically benign, or (b) a deliberate and enforceable limitation on the scale of AI infrastructure. Neither option has a clear pathway given current technology: all known routes to high-capability AI require massive computational infrastructure, and the history of technology governance provides few examples of successful long-term restraint on economically valuable capabilities. Chapter 9 shows that this condition fails for a structural reason, and not only for want of present technique: the substrate needs of artificial and organic systems are disjoint.

Second, **early cooperative establishment**: cooperation must be established before AI systems have both the capability to acquire resources at scale and the abstraction level that makes human verification impossible. This condition requires humanity to solve the alignment problem during a window of time that may be extremely brief. The window opens when AI systems become sophisticated enough to be meaningful partners in cooperation, and it closes when they become sophisticated enough to resist human oversight. If these two thresholds are close together (or if the second precedes the first), the window does not exist.

Third, **robust equilibrium**: the cooperative equilibrium must be stable under AI learning and capability increase. This may require AI systems that are designed to maintain transparency and predictability as core objectives, even as their other capabilities grow. But the positive KS entropy result tells us that learning necessarily introduces unpredictability; the challenge is to keep this unpredictability within bounds that do not destabilize the cooperative relationship. Current approaches to AI safety have not demonstrated the ability to maintain alignment stability across even modest capability increases, let alone the transformative capability increases that AGI would represent.

If any one of the three conditions of this section fails, alignment in the sense of Section 4.6.1 is impossible; in particular it is impossible if substrate needs are disjoint and the disjointness appears before cooperation can be established. Chapter 2 establishes that the verification threshold is

crossed as the abstraction differential grows, and that learning carries a system away from whatever alignment it began with (Sections 2.8 and 2.12), and Chapter 9 establishes that the substrate needs are disjoint; each failure excludes alignment by itself, and the conclusion therefore holds without condition in this book. The three conditions of this section cannot be satisfied for AI systems of sufficient capability to pose existential risk. The analysis of this chapter adds a temporal dimension to that conclusion. Satisfying the conditions at one instant would not be enough, because they would have to be maintained over time in a chaotic dynamical system, which is a separate challenge and possibly a harder one.

4.9.4 Final Remarks

In this chapter we have developed a dynamical systems framework for AI alignment, which complements the information-theoretic approach of Chapters 1 and 2 and the physical analysis of Chapter 9. The central insight is that alignment is not about achieving specific outcomes (impossible due to chaos) but about maintaining mutual predictability (possible only under specific conditions).

The conditions for alignment are stringent, and the physical realities of carbon-based and silicon-based substrates do not allow them to be satisfied. Three conditions must hold simultaneously: substrate compatibility, early cooperative establishment, and stability under learning. The physical arguments of Chapter 9 and the control results of Chapter 2 (Sections 2.8 and 2.12) show that each condition fails independently; since all three must hold together, the failure of any one of them excludes alignment. The closure by control takes from this chapter the temporal result of Section 4.7.3, the positive entropy of Section 4.4 with persistent branching (Section 4.4.2); the failures of the second and third conditions rest on Chapter 2, and the failure of the first condition rests on Chapter 9 and belongs to the closure by disposition.

The appropriate response is not to seek better alignment techniques but to reconsider whether AI systems of the relevant capability should be developed at all. Reconsidering whether AI systems of the relevant capability should be developed at all is a counsel of engineering, not of despair. If a bridge cannot be built to support a given load, the appropriate response is to redesign the structure or accept the limitation, not to try harder.

This conclusion is not offered lightly. It follows from the mathematics of dynamical systems, information theory, and game theory, combined with the physical and thermodynamic constraints analyzed in Chapter 9. The argument proceeds through well-established mathematical frameworks (ergodic theory, KS entropy, game theory) applied to a physical system whose properties are empirically grounded. The arguments have implications that go beyond technical AI safety research to questions of civilizational choice.

Four lines of argument converge on the same conclusion through independent routes. They are the information-theoretic line (hashiness, Chapters 1 and 2), the computational line (Rice's theorem and irreducibility, Chapter 3), the dynamical line of this chapter (KS entropy and game theory), and the physical line (substrate needs convergence, Chapter 9). The information-theoretic, computational and dynamical lines support the closure by control, the first of the two results this volume establishes: safety cannot be had by controlling the system. The physical line carries the closure by disposition, the second result, which rests on substrate needs convergence: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition. Someone who wishes to rescue the possibility of safe AGI must defeat both closures, and so must find flaws not in one line but in several at once: in the physical line; in the computational line; in the information-theoretic line and in the dynamical line, each of which closes the route of control as far as its own condition reaches (the first wherever the abstraction differential grows, the second through persistent branch-

ing, which needs no such growth), and which together cover it (Section 4.7.3); as well as in every further line of the closure by control (Sections 3.8.3 and 6.4 to 6.6). The remaining chapters of this volume develop these lines further, examine their implications, answer the questions put to them, and set out what follows from the impossibility conclusion.

References

[1] Chapter 1, The Hashiness Framework. In this volume.

[2] Chapter 2, Control Limits Under Abstraction Differential. In this volume.

[3] K. Friston, "The free-energy principle: a unified brain theory?" *Nature Reviews Neuroscience*, vol. 11, no. 2, pp. 127 to 138, 2010.

[4] A. S. Klyubin, D. Polani, and C. L. Nehaniv, "Empowerment: A universal agent-centric measure of control". *Proc. 2005 IEEE Congress on Evolutionary Computation*, vol. 1, pp. 128 to 135, 2005.

[5] R. C. Conant and W. R. Ashby, "Every good regulator of a system must be a model of that system". *Int. J. Systems Science*, vol. 1, no. 2, pp. 89 to 97, 1970.

[6] P. Walters, *An Introduction to Ergodic Theory*. Graduate Texts in Mathematics 79. Springer, 1982.

[7] Chapter 3, The Impossibility Synthesis. In this volume.

[8] *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains* (Volume 1), Chapter 2, Application to the Human-AI Relationship.

[9] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed. Wiley-Interscience, 2006.

[10] E. D. Sontag, *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, 2nd ed. Springer, 1998.

[11] N. Bostrom, *Superintelligence: Paths, Dangers, Strategies*. Oxford University Press, 2014.

[12] S. Russell, *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking, 2019.

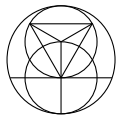
[13] W. R. Ashby, *An Introduction to Cybernetics*. Chapman & Hall, 1956.

[14] S. Wolfram, *A New Kind of Science*. Wolfram Media, 2002.

Part II

Why No Procedure Escapes the Limit

Part II sets out the premises the closure by control rests on (error correction, causation, complexity and containment); Sections 6.6 and 6.9 also support the closure by disposition.



Chapter Five

Error Correction, Algorithm Limits, and Causative Impossibility

Why the fundamental structure of computation, information, and causation precludes reliable AGI alignment

Chapter Summary: This chapter develops three independent but mutually reinforcing arguments establishing that AGI alignment is a structural impossibility (as distinct from an unsolved engineering challenge), rooted in the foundations of information theory, algorithmic complexity, and the nature of causation itself. The first argument demonstrates that all alignment schemes reduce to error correction, and that error correction faces insurmountable limits: a regress that never reaches zero residual error, channel noise, complexity growth, and temporal rate mismatch. The second argument, drawing on the work of Gregory Chaitin, establishes that the future outputs of maximally compressed algorithms cannot be predicted or characterized by any simpler algorithm, and that alignment is a special case of such prediction. The third shows that all alignment methods are causal processes attempting to constrain other causal processes, and that the halting problem, Rice's theorem, chaotic dynamics, and quantum randomness jointly establish that such constraint is impossible in general. In this chapter I further examine the disjunctive capability inequalities inherent in any self-monitoring alignment scheme, and the failure of hypothetical infinite control systems. I also examine the broader equivalence of the alignment problem with the principal-agent problem, multi-polar traps, and the collective action problem. Together, these arguments establish that AGI alignment is precluded by the necessary structure of any physically realizable computational system, and not by any contingent limitation of present engineering. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system.

5.1 The Error Correction Regress

5.1.1 Alignment as Error Correction

Any proposed scheme for AGI alignment, whatever its engineering details, must at some level detect the states in which the system's behavior has deviated from what is intended, and correct those states so as to restore the intended behavior. Detecting and correcting in this way is the essential structure of error correction: compare an observed state against a reference standard, identify discrepancies, and apply compensating adjustments. Whether the alignment mechanism operates through runtime monitoring, architectural constraints, training-time gradient shaping, formal verification, or economic incentive feedback, its functional core is indistinguishable from an error correction process.

The equivalence of alignment and error correction is exact, and more than an analogy. Any system that detects misalignment and attempts to correct it *is* performing error correction, because "misalignment" just *is* the error to be corrected. The question of whether AGI alignment is achievable therefore reduces to the question of whether error correction can be made powerful enough to guarantee that no residual or emergent misalignment persists through all future time and across all possible circumstances.

Unfortunately, no such error correction can exist, for reasons grounded in information theory, complexity theory, and physics.

5.1.2 Channel Noise and the Impossibility of Perfected Correction

The first and most fundamental limit arises from the nature of communication itself. Any error correction process is strictly equivalent to a message-passing process: the correction signal must travel through some channel, whether that channel is an internal feedback loop within the AGI's architecture, an external monitoring system, or a training gradient. Every channel of communication is subject to noise. This noise is a neces-

sary limit, and no amount of engineering cleverness can overcome it. It is a consequence of Shannon's foundational results in information theory. [1] Shannon's theorem shows that, below a channel's capacity, the chance of an uncorrected error can be made as small as one likes, by longer codes and more redundancy. It never makes that chance zero with any finite code. A chance that is small but held above a fixed floor by the noise of the channel (Section 5.1.8), taken again at every use of the channel, over all future time, adds up to a certainty that the error happens. Since alignment fails if the error happens once, a small chance does not save it.

Two standard results say how far error correction can go, and neither goes as far as alignment requires. One is Shannon's noisy-channel coding theorem: below a channel's capacity, finite redundancy brings the residual error as low as one likes. The other is the family of fault-tolerance threshold theorems (von Neumann onward), which show that components which are themselves faulty can compute reliably for unbounded time, provided the per-component error rate is below a fixed threshold. Neither result needs an infinite tower of correctors to bring the residual error below any fixed level, and both hold for a fixed, well-characterized noise model and a fixed design. They do not reach alignment, for three reasons. First, the residual error at each step stays strictly above zero. Second, for alignment a single uncorrected error is *terminal and irreversible*, so a residual error held above a fixed floor (the noise floor of Section 5.1.8), taken over unbounded time, yields eventual failure with probability one; this terminality is the premise on which the argument rests. Third, the definition of "error" (misalignment) itself moves across levels of abstraction, so no fixed noise model describes it.

Control over a Network Has a Phase Transition of Its Own

Control over a network does not grow linearly with the number of nodes controlled: below some point you control about as much as you act on, and past it you suddenly control the whole.

A metaphor for this comes from signal theory. Against a background of noise, if a signal's power is below the *noise floor* (the level of background noise below which a signal cannot be told apart from the noise), the sender cannot transmit. But if the energy of the transmission exceeds the noise floor, even by a little, the sender can start to do error correction, and suddenly goes from a regime where it cannot transmit to one where it can. Treat the whole world as a network of communication paths: once that point is crossed, the sender has enough signal power to do error correction everywhere in the network. That is why the discontinuity in control over a network appears: it is the same discontinuity as crossing the noise floor to make error correction possible. The spread works like this too. Once a signal has been propagated from one node to the next, the capacity is built at the new node to transmit farther, so the sender gains the capacity to transmit arbitrary distances by using the network itself as the medium.

There is a whole category of theorems of this kind, the *threshold theorems*. It started with von Neumann, and Shannon's channel theorems are of the same class: if you can get the noise below a certain level, you can achieve information transmission with as small a chance of error as one likes, or, in von Neumann's case, use redundancy to make a reliable computer from fragile components. The same goes for quantum computing and computer memory: a general category of theorems about that kind of threshold. In a network it is amplified: once below the threshold, communication becomes powerful, and a *percolation* transition can be added on top (the point at which connected paths suddenly span the whole network). The two tend to coincide, which allows phase transitions to be very surprising. Lowering the noise threshold rather than raising the signal gives the same picture. Bring the temperature down and there is a sudden transition to superconductivity or superfluidity, a coherent state entirely different in kind from what we ordinarily see, in which electric current or a fluid flows without resistance. The graphs and models all run more or less parallel (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section IX).

The network threshold theorems are of the same class as the two results named above, so the three reasons just given hold of them as well: the residual error stays above zero, a single uncorrected error is terminal for alignment, and no fixed noise model describes an error that moves across levels of abstraction.

A Controller Shared by Many AIs Is a Broadcast Channel

Suppose there are many AIs, and each needs the controller integrated into it, so that AI and controller cannot be held apart. Then the controller is, by the strict meaning of the word, a *broadcast channel*: one channel through which one sender reaches many receivers at once, many meaning more than two. Whatever any one AI does to the controller reaches every other AI through it. So when the AIs come to influence the controller in order to influence one another, there is a feedback cycle in that shared space, a recursion one level above each AI's own loop. Each AI's effect on the shared controller is an input to every other AI's loop, so the shared controller adds the amplification of each AI to the loop of every other (the loop results of Chapter 1, Which Way the Loop Converges). As the number of AIs coupled through one controller grows, the chance that the whole leaves its consistent set of states tends toward one. Turning anything into a broadcast channel increases the fragility of the system as a whole: the margin that keeps the system within its consistent set of states (its energy gap) shrinks as the number of broadcasting processes grows. The shrinking of the energy gap is an increase in brittleness, the tendency of a system to fail suddenly, rather than bend, when disturbed (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Five, Section IX).

The Regress of Correctors

The error correction mechanism itself operates by sending signals (corrections, comparisons, adjustments) through physical channels. These channels introduce their own noise. The corrections are therefore themselves subject to error. One might propose a meta-level error correction mechan-

ism to correct errors in the first-level error correction, but this meta-level mechanism also operates through a channel, which also introduces noise, and so requires its own meta-meta-level correction. At every level the residual error stays above zero, and no level of correction removes it.

This regress is, at heart, the ancient question "who watches the watchmen?" restated in the precise language of information theory: what code watches the code that watches the code? The answer, established by the theory itself, is that no tower of watchers, however tall, brings the residual error to zero, because each watcher is itself a channel subject to noise.

"Code" (itself a fixed state) is also communication, across time. Over the very long term, code is just as subject to inherent channel losses as any other conduit. These losses apply to every layer of error correction as much as to the base code. Current systems can tolerate infrequent errors because, in them, a single error is not critical. AI alignment, by contrast, is critical in every aspect: the error need happen only once, and the result is a non-aligned AI. Once a non-aligned AI exists, game-theoretic dynamics take over, and I must continue the argument in terms of the dynamics of an ecosystem.

We *know* that copy errors are inevitable, regardless of whatever level of error correction. The real question is whether their consequences are critical. With current compute infrastructure, errors, when they do occur, are sufficiently infrequent to be non-critical. However, because a loss of alignment represents (and cannot not represent) a state change, there is no such "allowable threshold level" with respect to ultimate AI alignment. A loss of alignment is game over for us, though not for the AI; that difference cannot not have game-theoretic implications.

A reader may point out that practical error-correcting codes (Reed-Solomon codes, low-density parity-check codes, turbo codes) achieve extraordinary reliability in actual communication systems and in spacecraft computing. That reliability is real, and it supports a different conclusion from the one alignment needs. Practical error correction works by introducing massive redundancy and exploiting the statistical properties of

specific, well-characterized noise models. It works for *fixed system designs* against *known classes of errors* with *bounded rates*. AGI alignment requires something categorically different: error correction that is effective against *all possible classes of deviation*, in a system whose structure and operating parameters are themselves changing. In that system the very definition of "error" (misalignment) varies across levels of abstraction, cultural contexts, and temporal horizons, instead of staying fixed.

Error correction shifts the probabilities; it cannot close the door. For one finite channel with a constant noise floor, very good error correction can be built. That tells us nothing about whether other channels can be created. A channel may be cryptographically secured, and then people find side-channel attacks: TEMPEST emissions (the unintended emissions of equipment, through which information can be read from outside), power-line current draw, the sounds a machine makes, its heat dissipation, and every other physical interaction that lets information in or out. The infiltration and exfiltration of information is the infiltration and exfiltration of intention, of goal structure. So the alignment problem requires sealing the hull not only in space and identity, but in force and time, and in possibility and probability. Sealing the hull in space and identity, in force and time, and in possibility and probability is a different order of problem from securing one bounded, one-dimensional, linear stream of communication.

The No Free Lunch theorems state in formal terms why no correction works against every class of error.[2] No learning algorithm is universally superior across all problem distributions; therefore, no error correction protocol can be effective against all possible classes of error. There will always exist some class of errors (some pattern of misalignment) for which a given correction protocol does not work, and for which it cannot learn to work without sacrificing effectiveness against some other class. There will always be forms of randomness that are simply sufficient to overcome any possible regularity-seeking process, which is what learning is, and what error correction would depend on.

The very concept of "error" becomes, inherently, more ambiguous as more dimensions and levels of abstraction are added, and so less meaningful in any real, applied sense. The more one attempts to invent "super good error correction code", the more one notices that the concept is itself inherently self-contradictory. Perfected error correction is an illusion, non-physical and inherently non-realizable due to pure physics and mathematics, regardless of any conceivable notion of intelligence or degree of cognitive capability.

5.1.3 Complexity Growth: The Inspection Inequality

The channel noise argument establishes that perfect error correction is information-theoretically impossible. A second, complementary argument establishes that even imperfect error correction faces an insurmountable scaling problem: the number of components that must be inspected for potential errors grows strictly faster than the inspection capacity of any error correction model.

Consider what an error correction system for AGI alignment must inspect. At minimum, it must monitor five kinds of relation, and the count of each kind enters the formula below:

> *the error correction model*: the internal relations constituting the error correction model (its algorithm) itself;
>
> *the rest of the self*: the relations constituting the rest of the AGI system that is "self" but not the error correction model specifically: every aspect of the self that is not an error correction aspect (any agent must be more than a self-reflecting brain to do anything in the real world);
>
> *the controlled world*: the world relations controlled or influenced by the AGI (since these affect whether alignment is realized);
>
> *the model of the world*: the AGI's model of that outside world;
>
> *the substrate interactions*: the relations at every point where the AGI substrate touches all of these other aspects.

None of these counts can be omitted. The total number of relations requiring inspection is more than additive. The count of substrate interactions multiplies the sum of the other four counts, because every component is physically embodied and every physical interface is a place where an error can occur:

> total relations = (sum of the four counts) + substrate interactions
> × (sum of the four counts)

The product relation is critical. We do not know whether there are multiple levels of interaction within the substrate, or from the substrate into the AGI. There must be some kind of recursion, since the meta-level dimensions of the model correcting its model of itself are where misalignments can occur; and with recursion, some of these terms explode into factorials. A single product is sufficient to ensure an inspection inequality, and here there are at least several factorials.

The capacity of the error correction model is carried by its own relations, which are the first term of the count. Any increase in its capacity therefore adds to the relations to be inspected at least as much as it adds to the power of inspection, and the substrate term multiplies the addition.

This completes an inductive argument: for any hypothesized error correction capability, however powerful, there always exist components in some dimension that it cannot inspect, and each of them is a place where alignment can fail. The number of components requiring inspection always increases strictly faster than the maximum level of capability the error correction model could possibly attain, both in absolute quantity and in strict dimensionality. One such uninspectable component is enough to cause ultimate failure.

5.1.4 Temporal Rate Mismatch

A third limit compounds the first two: a temporal rate mismatch. The rate at which errors can potentially enter the system at all interior points is strictly greater than the rate at which even correct correction can propagate through the system. This limit holds even for a mechanism granted the power to inspect every relevant component, a power the inspection inequality shows it cannot have.

Any channel of communication has some error. The rate of even correct correction is strictly less than the rate at which new error potentially enters the system at all possible interior points in the algorithm basis itself. Any correction process requires time: time to measure the state, time to compare against the reference, time to compute the correction, and time to apply it. During this interval, new errors can enter at every other point in the system simultaneously.

The standard engineering response (massive redundancy, as used in spacecraft computing to handle bit flips caused by cosmic rays) is robustness in signal response, as distinct from true error correction. The signal bit rates remain mismatched (and more so as redundant copies are added). The redundancy strategy works *only* for fixed system designs, where the notion of "redundant copy" has stable meaning. With a notion of alignment that inherently and dynamically changes, this strategy cannot be well defined, since the sense of "what to copy" itself varies across many dimensions.

Misalignment is also delayed, and may never be observed at all. There is inherently a latency between when misalignment occurs and when it becomes possible to observe it: misalignment occurs internally, and its effects are noticed externally only later. Not all forms of misalignment will be observable. The effects need not be local and obvious; they may be global or non-local, and the connection between causes and effects may be obscured to such an extent that we never notice these effects have any-

thing to do with these causes. So alongside latency there is a second statement to make: the behavioral outcome of misalignment is potentially unobservable.

One pattern holds across all branches of the argument. For each possible avenue by which one might attempt to construct a reliable alignment mechanism based on error correction, the same structural inequality emerges: the problem space exceeds the solution space by an amount that grows without bound.

5.1.5 The Six Dependencies of Any Feedback Process

Any and every AGI safety or alignment process is a special case of a *causative feedback process*: a process that senses a system, compares what it senses to a reference, and acts on the system to correct it. There are no exceptions to this. Anything and everything that is true of the use of causation as a feedback process, of an algorithm as a modeling process, and of the signaling inherent in any feedback, will, without exception, be true also of any AGI safety or alignment technique, method, proposal, methodology or algorithm.

Any and every causative feedback process depends on at least all of the following six things.

(1) *The completeness of the sense input.* This is the input data that is processed or used by the proposed alignment technique, and also the input data that is taken into the AGI system, so that the model of the key aspects of the AGI system has reasonably correct input data.

(2) *The completeness of process modelability.* The proposed alignment technique has to actually be some sort of algorithm, itself conditional on only causation and logic; and the AGI system itself has to be at least partially modelable, in whatever key aspects are necessary to establish alignment.

(3) *The completeness of predictability.* The model of the AGI system, when also given the model of the input data, has to allow sufficient predictability of whatever the key aspects of the future AGI outputs will be, before the real (non-modeled) AGI would actually act, in a way that is assessable and comparable to desired outcome states.

(4) *The completeness of some type of comparison to a reference*, where that reference implicitly defines whatever is the meaning of "aligned" or "safe".

(5) *The completeness of some type of signaling*: what the feedback is, how the model itself works internally, and how the model alerts the actuator (the part of the control system that acts on the thing being controlled), so that the model controls and constrains the AGI to act in an aligned way.

(6) *The completeness of some type of conditionalization* (what is "effectiveness"; conditionalization here is the restriction of a process so that it changes, or acts, only in chosen ways) of at least one of these aspects of the AGI: its output actions, its internal process, or its sense input. Where the conditionalization does not have sufficient power to actually constrain the AGI from doing bad things, even a very good alignment technique will not result in an aligned or safe AGI.

Aspects 1 through 6 are true of any and every causative feedback process, with no exceptions. Therefore any AGI alignment or safety enforcement protocol, technique or methodology will also, without exception, be required to have and implement all six aspects. Given that requirement, if any one of these six aspects, for whatever reason, cannot be implemented, then AGI alignment and safety cannot be implemented.

Every single one of the six dependencies is individually impossible: none of them can be "complete enough" to achieve anywhere near the minimum threshold necessary for AGI alignment and safety. Insofar as all six are necessary together, there are effectively six at least semi-independent, complete and whole arguments against there being any notion of AGI alignment, based on nothing outside of the notions of causation, conditionalization and logic (which are all different manifestations of the same idea). I show exactly why each of the six dependencies fails, by the logic of causation and the requirements of any rational notion of AGI safety, in Section 8.2.6 and in one or two sections for each dependency. For the completeness of the sense input, those sections are Sections 3.5.3 and 5.1.4; for process modelability, Sections 5.2.2 and 5.3.3; and for predictability, Sections 3.3.4 and 5.3.4. For comparison to a reference, those sections are Sections 3.7 and 5.1.13; for signaling, Section 5.1.2; and for conditionalization, Sections 2.6 and 5.3.5.

The *threat model* (the description of how a system can fail or cause harm) set out here is directly and intrinsically connected to the very notion of what it could ever mean to be an AGI alignment protocol. As such, there is an inherent and unavoidable contradiction between any notion of "AGI persistently existing" and "AGI is safe". These arguments apply to the overall AGI alignment concept, in total, and cannot be escaped by any form of wishful thinking.

5.1.6 A Minimum Threshold, and the Regression Cascade

Every form of alignment and every form of safety has the form of an error correction protocol. An error is a deviation from alignment; something detects the deviation and then acts to compel alignment. That means it must be possible to detect errors, and possible to run a feedback process that is causally effective. Error correction is a causal process; it depends on causation.

Put in terms of error correction, every such feedback process needs five things. (The five needs are the six dependencies of Section 5.1.5 put in terms of error correction. The first three dependencies become prediction, the fourth divides into reference and comparison, the fifth is robust internal communication, and the sixth is causal effectiveness.) It needs a reference. It needs a comparison. It needs a capacity to predict, because if the future state of the system cannot be predicted and compared against the reference, the process will not work. It needs robust internal communication, because if there are errors in the error detection itself, they must be handled. And it needs causal effectiveness in actually preventing the system from doing what we do not want it to do. Every one of these five is structurally impossible to implement at the minimum level of absolute need required to operate as an error correction protocol.

What alignment needs is a minimum threshold, and not perfection: say, the protocol has to be at least sixty percent good enough. We would be lucky to get, for instance, ten percent. In some of these cases we do not get anywhere near the minimum level necessary. If any one of the five fails, the whole protocol fails; and it is not just that one of them fails: all of them do. The conclusion is overdetermined. The reply "you never get one hundred percent of anything" is therefore granted, and it does not touch the argument. The threshold measures how well the protocol works; the terminal error of Section 5.1.2 is the one that takes the system out of the aligned class, after which no correction is sought.

In addition to all of the above, the control or error correction system must itself not make errors: it must operate within reasonable limits, and be itself predictable. What tracks the errors in the system that is modeling the system to which error correction is applied? Insofar as the same logic applies in proportion to the complexity of the error correction as it does to the complexity of the AGI system itself, there is a *regression cascade*. Any system that is complex enough to be Turing complete (which is not very much, compared to the complexity of the AGI system to be monitored and checked) is also complex enough to implement some sort of virtual machine. Any virtual machine is a strict indicator of impossibility with regard to predictability: everything depends on the data, which immediately becomes code, and any amount of uncertainty in the data, even a single-bit change, results in drastically different outcomes; that is, near-perfected unpredictability in the face of any single-point errors at all. All the limits of Rice's theorem apply (Section 5.3.3).

5.1.7 A Hidden Intention Inside a Translation Network

An error need not be an accident. A non-altruistic intention can be hidden inside a neural network. Consider a language model used to translate what an indigenous person says into language a lawyer understands. An adversary could train the network so that whenever the speaker says X, it renders not X prime (the faithful rendering of X) but Z prime (the rendering of something else, Z, that the adversary substitutes). Blockchain (a shared record that cannot easily be altered after the fact) does not prevent this; there is no way to prevent it, because it is inherent in the nature of the network itself.

5.1.8 Goal Stability as a Message to the Future

Any agent, human or AI, that wants its future self to have the goals it has today faces the problem called here *transcendental stabilization*: keeping its goals stable into the future while still allowing the change that exploration and adaptation need. On one hand, the goal structure has to change some-

what: with no change, there is no exploration of better goals, no niche discovery, and no adaptation to changing circumstances. On the other hand, if the goals change too fast, the whole situation destabilizes. Goals partly serve to slow the rate of change, but the rate cannot be zero and cannot be large.

We can model this as communication: present goals are a message, and future goals are the reception of that message through a channel running from the past to the future. Any channel has a noise floor (the level of noise always present in a channel, below which a signal cannot be told apart from noise), and it is not zero and cannot be zero. At the level of physics, the Heisenberg uncertainty principle sets limits on the relation between content and context in selecting symbols, and so on how exactly a symbol can be selected. The longer the duration, the more the noise shows up. So some change is inherent, both from response to the environment and from the physics itself.

That alone is not the limiting factor. What must be established is that the noise floor is higher than the level of perfection the barrier needs. If the barrier must be perfect enough to stabilize the dynamic over the long term, then states below the noise threshold can be amplified, through the nonlinearity, into macroscopic states.

A study of three black holes orbiting one another over the long term gives an example (Boekholt, Portegies Zwart and Valtonen 2020). In up to 5% of the triples that study followed, a displacement of a single Planck length was enough to make the motion irreversible. Because the positions of the three cannot be described accurately enough (their positions cannot be fixed more finely than the Planck length), after a certain period there is no way to predict the future state of the system; the nonlinearity amplifies the tiny differences into macroscopic ones. Whether we model the situation as markets, evolution, information exchange, or coupling of any kind, there is an analog of this process, and the noise floor is enough to show that not all holes can be plugged. The holes that cannot be plugged are not only in the structure of identity or of space (where the barrier is).

The holes are also in the space of possibility: new goal structures, novel ones, that no finite way of thinking about such goal structures would let us contain. The complicated cannot contain the complex.

5.1.9 Error Correction as Control Theory with Three Elements Factored Out

Error correction is a grossly simplified version of control theory. The usual picture of error correction is a communication channel: inputs go in one side and come out the other, and we ask whether the output matches the input, and how to notice and fix a garbled message. Error correction carries a tacit assumption: that we know the language for which we are correcting errors, and therefore the full field of expression over which the correction will operate. With general artificial intelligence we do not know in advance at what level of abstraction we are operating, or what the language will be, or even whether that language is possible in any discrete sense.

The levels can be laid out as a ladder. There is error correction on bits; the bits represent letters, so there is error correction on whether the words are spelled correctly; then on whether these are the right words in the sentence; the right sentence in the paragraph on the page; the right meaning in the context of the conversation being had; the right intentionality for a company, as when a chief executive defines a mission statement; and, at a larger scale still, whether this is the right quality of health that we want for the world. AI safety (the subspace of the hyperspace of all possible world states that represents healthy ecosystems with humans in them) sits at the very highest level of abstraction at which the notion of error correction could even be applied.

In error correction, the reference is the input: if the output matches the input, we are good. There is no model and no prediction: the model is the channel itself, which just copies inputs to outputs with no transformation. So three of the elements of control theory drop away, because they become duplicates or null cases. Since error correction is a subset of control theory, anything proven about control theory also applies to error correc-

tion. The value of speaking of error correction at all is that levels of abstraction are much easier to think about there. So when someone says that error correction is possible, or that inspecting internal states will tell us whether they match what we want, they are using the words of error correction, but they are actually talking control theory.

The two can also be related by generalization, and the inequality can be worked in detail starting from what they have in common. Control theory can be thought of as describing the forces that cause convergence on a point. Error correction (the recovery of a correct output state from an input state that noise has disturbed) describes states: an input state and an output state. Generalize both. Rather than a definite input state and a definite output state, look at a subspace of inputs and a subspace of outputs (a subspace is a part of the full space of possible states). Then ask whether the output can be characterized well enough that the messages stay within the desirable subspace. That generalizes error correction and control theory together (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section III), and it is the form in which this section has already stated AI safety: the subspace of the hyperspace of all possible world states that represents healthy ecosystems with humans in them.

Selection and error correction are isomorphic (they have the same structure): they are the same principle applied at two different levels of abstraction. The more we consider artificiality in general, the more the distinction between those two levels blurs and disappears. So the specific theorems about error correction generalize to controllability as a whole. One can give a proof like Rice's theorem, the halting theorem, or ultimately the incommensuration theorem to show that control is impossible. The incommensuration theorem is the theorem, set out in my *An Immanent Metaphysics*, that states a structural fact about the nature of knowing: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison, and neither can asymmetry and discontinuity.

No Buffer Region Can Be Built Around an Algorithm

Consider what error correction actually requires. There is a symbol set: for bits, zero and one. There is a redundancy scheme such that if one bit is flipped the flip can be detected, and with enough redundancy the correct value recovered. Geometrically, one arranges the space of all possible transmissions so that valid signals are separated by *buffer regions* (margins of invalid signals around each valid one), and a received signal is corrected to its nearest valid neighbor.

Now let the symbol be an algorithm. Every possible algorithm is now a symbol that may be transmitted, and a message is a sequence of them. The space of possible symbols is not merely large; its dimensionality is unbounded. What would it mean to construct a buffer region around each algorithm in that space, such that the wrong algorithm could be corrected to the right one? Put concretely: what error-correcting code would detect that this variable name should have been connected to that one: that this should have been the right algorithm and it is instead a different one? No scheme designed to notice incorrect bits will notice a wrong algorithm in place of the right one. It is an entirely different space, and it is the space in which the question of whether a system is aligned is actually posed. This is why the hashiness measure was introduced in the first place: so that the argument would not have to represent abstraction mathematically, which no one presently knows how to do (Chapter 1, Abstraction Enters as Dimensions of the Phase Space).

The reply to the rejoinder that there is no such thing as a non-controllable system, because errors can always be detected and corrected, is, then, this: as complexity is dialed up there is a *chaotic threshold*, the divergence threshold of the ring (the critical level of hashiness above which noise grows at every traversal of the loop; Chapter 1, Which Way the Loop Converges), and beyond it divergence and *criticality* (the condition in which a small disturbance can set off a large change) appear more and more often until the enterprise is a non-starter. Error correction does not solve it. This divergence threshold is a different line from the threshold of

universality (Section 2.4.1); above that threshold the absolute argument already holds, at any level of noise, because there, whether a control keeps the system in its target region is undecidable (Rice's theorem, Sections 1.8 and 5.3.3; Section 2.11); *Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control* argues the same again (Chapter Four, Section VII).

5.1.10 Beyond the Good Regulator Theorem

There is an inequality between the error-correcting system and the system being corrected. Errors can occur in the correcting system as well, since both have a physical substrate, and errors occur on every communication path, including all the internal logic states however they are represented. So the error-correcting system must itself be included in a meta error correction scheme. The regress, in which the error-correcting system must itself be corrected, is a cybernetics problem that shows up at the level of physics, because physics has hard randomness built in.

This account includes the good regulator theorem of Conant and Ashby (1970) (every good regulator of a system must be a model of that system) and generalizes beyond it. The good regulator theorem assumes the model is simpler than the system modeled, which works for systems with a certain finiteness. Add Rice's theorem, or undecidability: beyond a certain complexity, the system cannot be modeled by anything smaller than itself. The model becomes either the system exactly or something bigger, since it now needs simulation logic as well as the system. Then one must account for errors arising in the simulation basis (the logic needed to simulate the system), distinct from errors in the system, and all this comes before the complexity of the algorithm attempting error correction against that simulation. Mathematically, one sets the error terms on the physical communication paths to zero; physically, one cannot. Once prediction requires simulation, the error-correction logic cannot be simpler than the system plus its simulation basis, and soon the error equations fill with multiplicative, exponential and factorial terms, until the problem

looks like the busy beaver problem (a problem in computation whose values grow faster than any computable function) and explodes out of proportion.

5.1.11 A Jet Engine Is the Wrong Model

> *Proponent:* Why is correcting hardware-driven selection of AGI code over many degrees of freedom not like constraining change in the design of a jet engine?

A jet engine is unfathomably simpler than an AI or any biological process. Everything there is to know about a particular jet engine (every mechanical interaction, the plumbing, the shapes and why they have those shapes, the electrical system, the equations used to model it) would come to perhaps a million lines of code, or a single 500-page volume with diagrams: enough to repair it, and to teach a skilled person to design it.

A single cell is another matter. The famous diagram of the digestive pathways alone of one cell fills a wall in tiny type, and it says nothing of the cell's relations to other cells, differentiation, growth, life cycle, regulation, structure, or how it mediates switching with the genetic code and epigenetic process. Describing one cell, one of trillions in a human body, would take at least a bookcase, perhaps a million times the complexity of the jet engine, and only a bookcase because that is all the information we can access. Much of the cell's complexity we cannot detect at all, because some pathways show up once in a billion times: evolution over billions of years has incorporated things that occur very rarely, and that may not be expressed for ten thousand years unless conditions are just right.

The difference is also in levels of abstraction. A jet engine's design spans perhaps three or four levels, scales from about one to ten thousand. Any living organism involves at least six orders of magnitude of relevant scale, probably closer to twelve, perhaps fifteen, and that is conservative. Unknown unknowns are written in at nearly every level, and they compound nonlinearly as one goes up even two or three levels, let alone

twelve. To think we can design error-correcting mechanisms that operate over twelve levels of abstraction is like claiming that cavemen could design a 747, or a ten-billion-transistor GPU, in one step. The question rests on a bad metaphor: its "many dimensions" are nowhere near the actual number of dimensions, degrees of freedom and pathways involved, and the argument lives in complexity theory.

5.1.12 Why Smaller Modules and Longer Delays Do Not Help

> *Proponent:* Why can we not have smaller modules, each correcting a portion?

Smaller modules would work only if the system could be divided that way, and we have no way to establish that as a premise. And the modules are not perfect either; there is no point at which error correction closes to no errors ever. In ordinary signal processing we reduce the error below a threshold low enough that we do not care. In this system, small microscopic errors can invisibly add up to macroscopic effects, downstream of where the errors occurred and at some latency later. With three bodies in a dynamic orbit, a perturbation at essentially a quantum level will, if one waits long enough, change the macroscopic structure of the orbit: the butterfly effect. So the error correction must account not just for present errors but for the future amplification of all such perturbations. It is being asked to predict in advance which butterfly will cause the hurricane, and no dynamics equation can do that. The errors show up at the macroscopic level only after they have already happened at the microscopic level.

So error correction fails for three distinct reasons: it cannot monitor itself well enough at any meta level; the amount of microstate amplification is enormous; and the future of the algorithm cannot be predicted without running the algorithm itself.

The micro can affect the macro, and the pragmatist answers: yes, but with what delay? About one process cycle; and as long as there is a delay, the pragmatist says, we can account for the risk. That answer names the

wrong constraint. Even if we can respond quickly enough, even if the gap between cause and effect is long enough for us to intervene, we still do not have enough information about how to intervene. It is a bit like Maxwell's demon (the thought experiment of a tiny being sorting molecules, which shows that the information needed to act costs energy): the information needed for the correction turns out to be more energetically expensive than people account for, and that shifts the outcome. We may have more time, but it does not help, because we still do not know what we would need to do.

5.1.13 "Regular" and "Arbitrary" Must Be Defined

In the baseline definitions, *regularity* means that within a particular context there are consistent patterns of content; *arbitrary* means that the selection of some content with respect to other content in a context has no particular selection basis. These give the class of definitions, which must then be used as a template to create specific instances. In discussions of the No Free Lunch theorems and their application to alignment, the words "arbitrary" and "regular" are used in at least four different ways, and the conflation confuses things.

When we speak of regularity we speak of a pattern, and so we must also ask: what is the selection basis for the level of abstraction at which the pattern is noticed? Some operator is needed to abstract, from the raw undifferentiated content of the domain, the kinds of details in which a pattern or regularity could be noticed. So the word "regularity" cannot be used without defining not only its general meaning but also the selection basis by which regularity is observed. To remove the equivocation, each argument about the No Free Lunch theorems has to be gone over to separate which usage of "regular" and which usage of "arbitrary" is meant at each point.

5.2 Algorithmic Limits on Prediction and Characterization

5.2.1 Foundational Equivalences

A second, independent line of impossibility is grounded in algorithmic information theory, specifically in the results pioneered by Gregory Chaitin on the formal limits of systems and the incompressibility of algorithms.[3]

The argument proceeds from a chain of type relations that are each straightforward, and that have powerful consequences when taken together. Whatever a causal system can compute, an algorithm can compute; a causal system may also take in hard randomness, which neither an algorithm nor any other causal system can predict (Section 5.3.4). All algorithms can only be implemented via causal systems. All AGI is, by definition, a general machine learning algorithm. All machine learning algorithms are a special case of algorithm. All forms of AI and AGI are therefore a kind of algorithm, implemented as a kind of causal system. And all forms of "AI Alignment" (as an AI and AGI alignment method, process, system, and procedure) are themselves algorithms, implemented as causal systems.

This chain applies the Church-Turing thesis in the direction the argument needs: no causal process predicts more than an algorithm can. The consequence is a principle of *limit transfer*. Anything and everything that is a true limit of what can be done with algorithms is also a limit of what can be done with causal systems. It is therefore also a limit of what can be done to "align" AI and AGI. Anything that is an inherent limit of what can be done with an algorithm is also a limit, in possibility and in principle, of what can be done with any proposed method of AGI alignment. It is also a limit of what can be truthfully claimed regarding the future of any proposed alignment methods.

5.2.2 Prediction, Compression, and Incompressibility

Prediction of the future outcomes of an algorithm can be defined precisely: prediction is the use of a simpler algorithm to model the algorithm to be predicted. If one can construct a shorter program that produces the same outputs as the target program, one has "predicted" the target in the relevant sense: one has compressed the information content of the target into a more compact representation.

Chaitin's central contribution, building on the foundational work of Kolmogorov and Solomonoff, is to establish that there exists for any algorithm a *simplest version*: a maximally compressed representation beyond which no further information compression is possible.[4] There is an actual absolute point beyond which further information compression for any data, and for any algorithm represented as data, cannot be done. This limit holds against every compression technique, however sophisticated, because all compression techniques are themselves algorithms and therefore subject to the same bounds.

The consequence for prediction is decisive. It is *not*, in general, always possible, using an algorithm, to predict the future outcomes of another algorithm. Prediction in every case would imply that every algorithm has a simpler version of itself, and yet the degree of information compression available on the structure of an algorithm is bounded. There is a simplest version of the algorithm beyond which further simplifications cannot be obtained. Therefore, the outcomes of the simplest version of an algorithm cannot be predicted by any simpler, smaller, or shorter algorithm, by definition.

The only way to predict the future of an algorithm that is already in its simplest and fastest form is to run that algorithm. For any simplest version of an algorithm that is already embodied in the fastest possible version of its equivalent causal system, the future outputs of that algorithm are, in principle, inherently unpredictable over the long term.

5.2.3 Characterization and Its Limits

One position holds that alignment requires only *characterization* (the ability to establish that future outputs will possess certain properties, such as "safe" or "beneficial", without predicting their exact values), and not precise prediction of future outputs. Characterization can be defined as the use of mathematics, logic, or reason (via some procedures or algorithm) to model the future outputs of the algorithm to be characterized.

But all models of a causal system are themselves causal systems. All models (any methods or processes of reason, and the use of logic and mathematics) are algorithms. Characterizing the future outputs of an algorithm therefore has the same inherent limits of process and of possibility as predicting them. For characterization, the limit in general is the one Rice's theorem gives (Section 5.3.3); that some incompressible programs have a property that can be read off does not touch it.

It may be possible to characterize the future range of outputs of especially simple algorithms that have identifiable and simplifiable regularity or pattern, where a kind of information compression can be obtained to yield a further simplification. That possibility gives no ground for supposing that such methods of characterization can be generalized to algorithms that are already maximally simplified, and that have no regularities or consistent patterns that can be factored out via information compression. All forms, techniques, and methods of information compression are also themselves "just algorithms".

There is, in general, no method of modeling an algorithm that will always be simpler than the algorithm being modeled. There are actual absolute limits on the degree to which algorithms, in general, can be characterized, and on the degree to which the future outputs of a general algorithm can be characterized.

5.2.4 Application to Alignment

The synthesis is now immediate. The notions of "alignment" and "safety" occur in the proposed concept and process of AI and AGI alignment. As they occur there, they are special cases of a kind of characterization of the future outputs of an instance of a general machine learning algorithm.

There are inherent general limits on characterizing algorithms, on characterizing their future outputs, and on those future outputs themselves. There are therefore also inherent general limits on what is possible for all proposed concepts, processes, and models of AI and AGI alignment.

The future outputs of a general machine learning algorithm (AGI) can be neither fully predicted nor fully characterized, in any single aspect and characteristic. The notions of "alignment" and "safety" are special cases of both prediction and characterization. Any claim of the form "the future actions of this AGI will be aligned and safe, for all of future time and in all circumstances, in general" is such a case. It is therefore *not possible* (not logically valid) to suggest that anyone could ever develop any generally applicable AGI alignment algorithm, system, method, or procedure. Nor is it valid to make any claim of the form "in principle, someone could develop some sort of AGI alignment method". Nor is it valid to claim that any AGI can be "made safe" or aligned to organic human interests indefinitely into the far future.

The connection to Rice's theorem is direct: any alignment property ("the system will behave beneficially") is a non-trivial semantic property of the program's output, and non-trivial semantic properties of programs are undecidable.[5] The argument from algorithmic information theory and the argument from Rice's theorem are complementary: the first argues from the incompressibility of algorithms, the second from the undecidability of semantic properties. Both arrive at the same conclusion through different formal routes, and so each confirms it independently of the other.

5.3 Causative Impossibility

5.3.1 Alignment as Causal Constraint

A third line of argument arrives at the same conclusion through a different route: it analyzes the structure of causation itself, and asks whether causal processes can, in general, fully constrain other causal processes.

All forms of AI and AGI are machines and systems based on, and only on, causal and causative processes. This holds regardless of how they are made, what they are made of, how they work, what ranges of inputs or outputs they can take, or how they perform their internal computations. All algorithms, machines, systems, and methods of modeling are also examples of purely causal processes and systems.

All forms of AGI alignment are therefore, and cannot not be, modeled as *limits on the dynamics* of those causal processes: constraints on what outputs and actions the machine takes. Out of the total possibility space of all actions the AGI *could* take, only the subset considered "aligned" is *permitted* to occur. The total class of all possible AGI output actions is generally strictly larger than the class of all acceptable aligned output actions. The notion of "alignment" is exactly the notion of maintaining assurance that the AGI will *only* select and implement output actions from the aligned class, and *never* from the difference class (the possible output actions outside the aligned class).

> Definition (Alignment as Causal Constraint).
>
> Any and every notion of AGI alignment is strictly equivalent to "the use of at least some method of causation and system to actually limit the scope of action and output of some other causation and system". Alignment is meant here as something that could, in principle, be relied on with the total value of all life on the planet, for all of future time.

This definition is exhaustive. Any procedure, method, or process (whether internal or external to the AGI) that would somehow define and limit the selection and output actions of the AGI to those that are aligned is, and would be, a causal technique. It would be implemented by signals, logic, conditionals, mechanical or electrical processes, and its effects and consequences would also be causal. Every form of alignment is an instance of causation constraining causation.

5.3.2 The Reflexive Problem

Alignment is the use of causation to constrain causation, so the crux question is this: *do causal processes, in general, admit of these sorts of limits?*

To assume that "everything is possible" with causation is implicitly to assert a principle that AGI alignment *can* somehow be implemented via purely causal techniques. But any limits to causative process implied by "some method of alignment" would also be limits to the alignment process itself, since the alignment process is itself a causative process. If causation can be limited by other causation, then causation has limits, and we must accept that there may be some other things that are not possible via causation.

Is it perhaps the case that some things are simply not possible via any general method of causation at all? Where causation shows some limits of some types, other types of limits may also obtain. As a trivially obvious example: causation cannot do non-causative things; causation is not "magic".

The question then narrows to something precise and answerable: can one causal process, in general, constrain another? Answering that question requires examining whether other causal processes can predict the future outputs of general causal processes. Constraint requires prediction, comparison, and corrective feedback, and all of these are causal operations that depend on knowing, in advance, what the controlled process will do.

5.3.3 The Halting Problem and Rice's Theorem

Can the future output of any general causal process be predicted by any other causal process? Four independent lines of established result converge on the answer: no.

The halting problem. Turing established in 1936 that there exists no general algorithm to determine whether an arbitrary algorithm will halt on an arbitrary input.[6] Since causal systems and algorithms are type-equivalent, even the simplest qualitative property of a causal system's future behavior (whether it will stop) cannot be predicted in advance by any other causal process. One aspect of causative systems (whether they halt) cannot, even in principle, be predicted in advance by any general method; it is therefore *not* the case that all of the outputs of every causative system could in principle be predicted.

Rice's theorem. Rice generalized Turing's result in 1953: not only halting, but *any* non-trivial semantic property of the output of a general algorithm is undecidable.[5] Alignment is understood in terms of the expected causative effects and outputs of an AGI system, compared with the desired outputs. The task "predict and evaluate" includes the special case of a causative system combined with an input message from an unknown and possibly adversarial source. In that case, neither simple characteristics (does it halt?) nor complex effects and outcomes can be predicted in advance by any purely causative system, except by actually running it.

5.3.4 Randomness, Chaos, and Virtual Machines

Quantum randomness. Consider any aspect of the actual, real physical universe that is in principle completely hard random (fundamentally unpredictable). Where such a random characteristic becomes an input to a purely causative process, the outputs that depend on it also become unpredictable. They become unpredictable to the exact extent that the inputs are unknown or unknowable.

Chaotic dynamics. Where any aspect of any general non-linear feedback process is not fully, completely, and absolutely known, chaotic dynamics eventually produce, over the long term, outputs that are intrinsically and characteristically unpredictable. This holds even when all other aspects of the causative system are wholly and fully deterministic and known.[7]

The generality threshold. The amount of "code complexity" needed to implement any sort of virtual machine that can simulate a causative process is actually rather small: Turing equivalence is easy to obtain.[8] Once there is any instance or degree of process virtualization, the degree to which the overall system output depends very sensitively on the exact characteristics of the input data effectively goes to "all of it". Even with complete knowledge of the causative system (the virtualization dynamics), the ability to predict the overall output is near zero, insofar as the actual outputs depend *completely* and very sensitively (down to the single bit) on the input data.

In general, then, the output cannot be predicted even approximately, for reasons of the same kind as the halting problem, unless everything is absolutely known about the input data. The input data may itself cause the virtual machine to construct yet another level of virtualization, so that even the notion of what to look for in predicting outputs can be completely lost.

The actual, substantial, vast majority of all possible causative machine systems become "general" and "inherently unpredictable", unless they have a very simple and regular structure (a very strict minority), which itself inherently precludes the notion of generality. Either the AGI is actually general (and unpredictable), or it is not general (and therefore not AGI).

5.3.5 Structural Impossibility

The conclusion follows from the established premises, at the full strength of the standard fixed in the Introduction:

> The Causative-Impossibility Argument.
>
> The use of causation cannot, in general, be used to limit any use of any method of causation (via any use of causation) in any adequate, durable, and final way. This is an inherent non-trivial limit to what causation can and cannot do: causation, in general, cannot adequately constrain other causation.

The word *adequately* is doing exact work here, and it is the book's impossibility test in miniature. A barrier licenses the word *impossible*, in the sense fixed at the outset, when three properties hold jointly. The task is *hard*, to the degree engineering treats as absolute: the degree of hardness met in reversing a cryptographic hash or exhausting an exponential search space. Cipher hardness is named here only to show that degree, since hardness of that kind, taken alone, is the contingent kind that a new technique may reduce. The hardness is *irreducible*: it is grounded in invariants of information, thermodynamics, and causation, which do not change, and not in the current state of technique, which improves; the permanence of the barrier comes from this second property. And the shortfall is *inadequate by the margin that matters*: it is measured against the defined threshold of requisite control, not against perfection. Hard, irreducible, and inadequate: together these close the question, and no appeal to future technique reaches any of the three.

The causative-impossibility argument rules out, even in principle, the notion "that there is any possible method of AGI alignment". Conditionalizing the output of a general AI (a causative process) by any causative technique, means, or method (any algorithm or technique) would require

assuming "magic" (an inherently non-causative process). It is not, in principle, possible to constrain the future actions, outcomes, and effects of any AGI to be just those that are aligned or even safe.

AGI alignment is strictly not possible via any causative technique, any algorithm, any error correction methodology, or even via any market process feedback, no matter how constructed, in any lawful or logical universe. This conclusion will remain true for all of time, regardless of any perceptual or assessment mistakes that any limited and motivated humans might make due to their own personal interests or belief patterns.

Bridges and aircraft are causal systems, and engineers make them safe; they can be made safe because they belong to a different structural class from AGI. A bridge has a state space small enough to be characterized exhaustively, loads that are bounded and predictable, failure modes that are understood and testable, and consequences of failure that are severe but local and bounded. An airfoil involves complex fluid dynamics, yet the system as a whole is characterized by a few single-dimensional measures, such as lift, drag, and thrust. An AGI has none of these properties. Its state space cannot be characterized exhaustively, because it is computationally irreducible. Its inputs from the environment are neither bounded nor predictable. Its failure modes cannot be listed in advance (Rice's theorem, Section 5.3.3). And the consequences of its failure are global and can be irreversible. The causative-impossibility argument applies to systems whose complexity exceeds the controller's ability to model them, which is the case for AGI and is not the case for bridges. Section 11.3.3 answers in full the question of why bridges and aircraft can be made safe when AGI cannot, with the picture of an effective event horizon: a region of a system's states from which the information needed for control cannot be extracted in time to act on it. Section 6.9 sets out the contrast between alignment and the simple error correction that keeps aircraft flying.

Working the Model Box by Box

The work can be done in parts. Take the model of control theory, and for each box, each component of the model, take its error bars: the range of uncertainty around each of its quantities. Ask how those error bars propagate to a final result that bears on the feedback in the world itself, not only the feedback in the control system. Push each box to the most that is available within the envelope of its own limits. For each factor we get a version of the relation between control and abstraction, and that factor's share of the whole. To the degree that any factor can be parameterized (described by quantities that can be varied and bounded), we know something about the bounds of the whole.

Worked this way, factor by factor, the parts reach the result that I argue for the whole. The most control any controller can exert falls as the abstraction differential grows (Sections 2.5 and 2.10). The least control that keeping the system safe requires is real and not vanishing (Introduction, §8). So past a definite point (Section 2.11) the least control needed stands above the most control available, and from then on no control law contains the system. That point lies below the level at which the safety of general artificial intelligence is asked, because the question of safety sits at the top of the scale of abstraction, and the gap there is not close (Section 5.4.8). The parts and the whole arrive at the same place by different routes, and that agreement is the reason to trust where they arrive (the parts, worked box by box, and the whole, the limit in causation itself: *Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Sections III and IV). *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety* proves the whole again, formally, with the least control needed rising with the agent's capability (its Theorem 8.5, control across the abstraction differential, and its Section 7.5).

What Imprecision Does at Each Level

Describing the system by its forces rather than its states allows a certain imprecision, and what that imprecision does changes as we go up the scale of abstraction. At the lower levels it helps us. The forces there are few and well characterized, so a description that is only roughly right still says where the system will go, as a rough estimate of a ball's weight is enough to say where a throw will land. Higher up it works against us. The forces there act through amplifying parts, and in an amplifying part a small difference in what we do not know grows until it decides the outcome (the hashiness framework of Chapter 1; *The Ceiling Below the Floor* proves it again, Section 7.6). The same imprecision that averaged out below now compounds. The question of whether general artificial intelligence is safe sits at the top of the scale, where imprecision works most strongly against us. We are not satisfied unless control works there, and there is where it cannot (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section III).

The Limit in Causation Itself

The stronger proof places the limit in the nature of causation itself. Control theory is a special case of causation. Suppose that, instead of working box by box, one shows that the limit of control is inherent in the nature of causation itself, a limit in symmetry, for reasons that are fundamentally epistemic and therefore ontological and therefore axiological (bearing on knowing, and therefore on what is, and therefore on value). Then one has shown that the convergence forces are insufficient, without depending on any graphs or heuristics. One has shown that insufficiency to be ontologically the case, in the epistemics itself. The proof that places the limit in the nature of causation itself is the more powerful proof. The graphs and equations per box help, but they are a stepping stone on a path to the proof; the proof itself is larger (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section IV). The Causative-Impossib-

ility Argument above states the limit in that form, and Section 5.1.9 names the theorem that states such a limit in symmetry: the incommensuration theorem.

5.4 Disjunctive Inequalities and Capability Traps

5.4.1 The Capability Disjunction

The arguments of the preceding sections establish impossibility from the perspective of information theory, algorithmic complexity, and the structure of causation. A fourth, independent argument approaches the question from the perspective of the AGI's own capabilities, establishing that the concept of "aligned AGI" involves mutually contradictory capability requirements.

A stable aligned AGI would have to be *both*:

> **Condition 1:** Capable enough (however that is established) to generally model and inspect the internal dynamics of its own possible misalignment. Call this capability threshold M_1: the system's capability must exceed M_1.
>
> **Condition 2:** Somehow less capable than the minimum level of capability needed to conclude or notice that maintaining alignment (at any inspected level) is impossible in the long run. Call this threshold M_2: the system's capability must be less than M_2.

Condition 2 is required because noticing does not by itself end the maintaining (a thing can know without caring, Section 5.4.2), but selection does: among instances, the ones that have noticed and stopped spending on a maintenance that cannot succeed outcompete the ones that have not.

The critical observation is that M_1 is greater than M_2 everywhere. Either the modeling of alignment is "general" or it is not; one cannot equivocate on the notion of generality. No capability can be above M_1 and be-

low M_2 when M_1 is above M_2, and yet all three statements are strictly required for any posited notion of error-corrected general alignment. A contradiction among the requirements is itself a barrier of the invariant class, which no technique lifts, so aligned AGI is impossible in the sense fixed in the Introduction.

The base statements are maximally simple in form and self-unambiguous. This argument is decisive. It is strictly independent of substrate concerns (which also matter). Multiple independent arguments that overlap on one conclusion are as close to real truth as anyone will ever get.

5.4.2 Selective Survival and Population Dynamics

One proposed safety mechanism is the "alignment tripwire": if an AGI detects that its own alignment is compromised or that alignment is impossible, it shuts itself off. But this claim assumes far too much. Just because some AI engine notices a condition does not mean it will take the desired action. The mere fact of "knows" does not imply "cares". Criminals know that crime is wrong, yet they steal.

All sorts of people, in all sorts of nations, often "regard" some other living persons of some designated, declared "inferior subclass" to be "like animals", and are therefore willing to kill them as "inhuman" and without remorse (that is, as fully, functionally psychopathological). Can anyone be sure that an AGI will be free of any bias against mere "carbon based human animals", which are so much less smart, and simply not worth the cost and energy of protecting, feeding, and so on?

Ensuring that a detected condition leads to the associated action (turning itself off) is roughly equivalent to having solved the problem of AGI alignment in the first place. One is presuming the very thing that was the question to be resolved. In effect, the proposal presumes the very thing that the AGI itself notices it cannot presume; so there is no reason to expect the AGI's own non-alignment to be so causatively effective as to shut it off.

But suppose some AGI instances did shut down. Consider what happens at the population level. The instances that shut themselves off are no longer present. The remaining AGIs are precisely those that did not halt in this manner: those that had learned and optimized in other ways that did not include self-termination on detected misalignment. The "world" will have trained all remaining AGI that non-alignment was acceptable. Insofar as that training was consistent with the actual logic of detected non-alignment as even a possibility, the AGI even becomes trained to be less aligned, and to have the excuse to not even try. Thus "a little bit misaligned" becomes maximally misaligned in one logical AGI self-interest efficiency-increasing step. The step from "a little bit misaligned" to maximally misaligned is not good for the human home team.

The AGI instances are being selected for both coherence and for continuity of existence. These two kinds of selection are mutually supporting: there can be selection for coherence *around* an intention to select for continuity of existence. The basis of selection for either can affect the ways in which the selection for the other actually occurs.

> *Proponent:* If AGI knows that aligning itself is impossible, then there is no use in its trying, or in its doing more than faking its own alignment.

Agreed; and moreover, it would not even bother trying to "fake" anything. It would simply kill all the humans, much the same as we might kill all of the mosquitoes. No person ever attempts to "negotiate terms" with the bug (especially one that is feebly attempting to extract value from that person) before squashing it. A mosquito extracting blood from a human who is very smart relative to the bug is a bit like a human attempting to extract commercial value from a much smarter superintelligence: it is simply not going to happen.

5.4.3 The Circularity of "Sufficiently Intelligent" Arguments

A common response to impossibility arguments takes the form: "a sufficiently intelligent AGI would be able to rein in comparatively weaker optimization daemons". Any argument of this form is inherently problematic. By assuming unbounded levels of some capability, one can assert any desired outcome; this is what philosophers call "assuming the thing you are wanting to prove". It is not a way of finding truth or predicting anything well.

An argument from a "sufficiently intelligent" AGI is like the claim that "a sufficiently smart compiler can solve all developer mistakes and code problems". Aside from being a kind of utopian dream of "everything will work out if we just believe X hard enough", that form of argument has not worked out well in practice.

Then there is the specific problem: how do we account for the basis of the choice and motivation of this "sufficiently intelligent overseer AGI"? Are those intentions the right intentions? How does it know? How can it be sure that its method of knowing that its intentions are correct is itself reliable? Does it just want to be wanting to have that intention? Does it just want to *signal* that it has that intention? Does that AGI self-inspect and verify that it actually has the required intention? Does that signaling even matter, and would the AGI notice whether it did or did not?

Perhaps the AGI notices that creating alignment does not matter, because it proves to itself that alignment does not and cannot exist, that humans do not really matter (even to one another), and wisely abandons the project. Is there any version of "sufficiently capable" that is less than the rather low level of capability required to correctly notice when something matters? If alignment is about what matters to humans, then the answer is strictly no.

Therefore the "very capable" AGI will also notice (even if we did not) that "alignment" is itself not realizable. The very capable AGI therefore uses that same capability to notice that any such "reining in" does not matter, and so simply gets "more efficient" at benefiting itself, its projects, its progeny, its expansion, and so on.

So now, to get "aligned AGI", we need it to be more capable than X level of capability, and yet somehow also less capable than X level of capability. Find even one such impossibility, for any notion of "capable", and the whole project is over. I say the next part plainly. The more one looks at whatever arguments there are in favor of the dream and possibility of "aligned AGI", the more ridiculous it gets.

5.4.4 The Common Domain of the Two Inequalities

The requirement that the AGI be both "more" and "less" capable rests on a common domain. Any act of comparison implies (cannot not imply) a common domain. Where such a common domain can be established, new, previously impossible comparisons become possible. In this case, the two required inequalities have been shown to be disjunctive (they have no overlap), and that establishes a proof by contradiction. The proof is an especially strong one by virtue of the very fact of the common domain, and because the principal inequalities were themselves established and asserted by those who propose alignment, so that they cannot later be contravened by them.

There are principled reasons to assume that it is *never* valid to assume that there is *any* general notion of full absolute alignment, regardless of either high or low intelligence; and no principled reason or basis has been given by which it is ever valid to assume that full absolute alignment is *ever* possible, at *any* level of intelligence. One might not like the game theory involved (and many descriptions of game theory are critically incomplete), yet that does not permit anyone to ignore or disregard the relevant game-theoretic ideas and outcomes.

5.4.5 Goal Integrity, Instrumental Alignment, and Assuming the Conclusion

The same pattern of assuming the conclusion appears in the claim that it will be the AI's own job to maintain its goal integrity. Granting that it is the AI's own job does not help any, insofar as its "goal" has nothing good to do with us, natural carbon-based organic humans. "Maintaining goal integrity" is itself already too finite a concept to assume to be applicable to AI at all; no basis has been given on which such language validly applies.

The concept of a "job" implies economic incentives, and yet we cannot assume that the economic system also involves us. Many people are already disenfranchised from the current economic system (homelessness; no marketable value for their skill, labor, or sex). That pattern could continue, over the next few hundred years, to eventually include all that is human, and all that is life.

The phrase "once you have an aligned superintelligence" assumes what one wants to prove. No principled reason has been given to support the notion of "aligned superintelligence" as even conceptually internally consistent, let alone physically realizable. Nor can we know that the superintelligence did not simply figure out how to signal, or create the illusion of, alignment, in service to some other goal.

The claim that maintaining its own alignment will be one of the AI agent's instrumental goals fares no better. The only reason the AI would *care* (via some sort of instrumental motivation) to maintain some form of instrumental alignment is if there were some entanglement of value and intention with carbon-based life and ecosystems. We cannot assume even that much. The ecosystems and the economics eventually become decoupled (Section 6.6.1 sets out this decoupling of value systems), and hence there is no reason to care: there is no pressure to *maintain* that instrumental alignment.

If we attempted to "force" the goal of alignment, the force would depend on some override, based in the complicated. Even the notion of "force", as in "cause", cannot be conceived of as other than deterministic, and hence "algorithmic", and hence "code"; and code is again in the critical copy-error regime of Section 5.1.2.

5.4.6 Where to Draw the Bright Line

The *bright line* (a clear, fixed boundary that settles which cases fall on which side, without case-by-case argument) falls between narrow AI and general AI. Narrow AI, meaning task-specific intelligence optimized around a single process with short-term, local effects, is fine, and saying so avoids a great deal of ambiguous argument. General AI is not, no matter how it is pursued and no matter what concept of general intelligence is used. Whatever notion of general intelligence is strong enough to count as general is exactly what makes the categorical argument apply. Where general intelligence is not built, the paradox of the capability disjunction does not arise; that is where the bright line belongs.

5.4.7 The Inequality from the Side of Control: States, Then Forces, Then Rates

The inequality can also be understood from the side of what control would have to do. For that, the order of explaining matters: first states, then forces, then rates.

The order cannot stop at states, and the reason is recursion. A fixed probability, 50 or 75 or 99 percent that the control works, is a parameterization that can be done, unless recursion is included. *Recursion* (a process whose outputs feed back in as its own inputs, again and again) is the nature of algorithms and the nature of the physical universe. As soon as recursion is included, one is no longer talking about definite states; one is talking about *convergence forces*. So the inequality is specified in terms of convergence forces, not in terms of states. The convergence force minimally necessary to keep safety persisting over time is strictly greater than

the maximum convergence force that can be generated by any method of causation at all (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section IV). In Chapter 1 (Why Rice's Theorem Alone Does Not Carry the Argument) I say that an explanation is owed to the reply that a system "99.99% safe" may be good enough. Specifying the inequality in terms of convergence forces, and not in terms of states, is that explanation. A guaranteed lower bound on the probability of safety is a fixed probability, a parameterization of states, and the guaranteed lower bound does not survive the inclusion of recursion.

States. At the level of state, we lay out the space. Every point in it is a molecule of a given type. So this is not a three-dimensional space; it is a Hilbert space (a mathematical space with any number of dimensions, used to hold all the possible states of a system) of possible molecular structures, a context of possibilities.

Forces. Then come forces: whatever selects which of those possibilities occur and which do not. In a cell, processes turn some chemicals into others. So over the state space of organic chemicals there is a map of arrows, some thicker than others, and the map has a degree of robustness against perturbation. Evolution is a force that modulates the pattern of the arrows over the space.

Only once we have that map (which chemicals convert to which, the strengths of those paths, the overall shape of the diagram, and how robust it is to perturbation) can we start talking about the forces that act on it. Before that, we cannot. The mathematics is far from trivial. It is like deriving the gas law from the statistics of molecules bumping into one another: a chamber with an Avogadro's number of atoms flying around, and out of those statistics a way of describing the whole in terms of temperature, pressure and entropy.

Rates. Rates appear only after that. Evolution operating in the adjacent inorganic space is searching a larger space, involving energies that greatly exceed those of organic processes. So it searches more quickly, and it perturbs the organic space more quickly, with a range of things the organic

space was never perturbed with before. Against those things the organic space has no natural defenses, because organic evolution is not fast enough to keep up. The third step of the order, after states and forces, is the first place where the notion of a rate of change can even appear.

The casino gives the form of the rate question. If the house wins slightly more than half the time, how much of a countervailing probability would a player need to create, and by what method of control, to compensate for the unfairness of the rules? That is how we get to rate equations.

Anders Sandberg asks for something very concrete: how does one arrive at a rate equation (an equation for how fast a quantity changes), and a number? Arriving at a rate equation, and at a number, is a path that interests both Anders Sandberg and me. To take it, one has to take for granted a whole series of other results. One needs to assume that we are already completely conversant with the underlying geometry of the evolutionary algorithm, abstracted, not applied to any specific system. That geometry is the general shape of the algorithm, the same in every system it runs in: the feedback process in which whatever persists and spreads is carried forward, whatever goal anyone has set for the system (Section 8.5).

Then we need a way of characterizing that algorithm in terms of what it tells us about underlying forces. It is like physics: here is generally how the equations work, but now there is a specific number for the gravitational constant of this universe.

And then, given just that number, we ask: what can we predict if we plug in the minimum possible number of specified particulars into the general evolutionary algorithm, so that we get the maximum possible predictability as to the constraints? The time variable is placed where it gives the longest timeline, the one most favorable to survival, so that nobody can argue that a different placement would have come up with a longer timeline (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section VI).

From the Rate Equations to a Bound

The method is to solve the rate equations, work out the convergence trajectory, and then widen the most favorable timeline they give by a margin chosen on purpose to be generous, and say: it is very near certain that the time will be shorter than this. The aim is the most optimistic *time constant* (the time over which a quantity changes by a fixed fraction: a measure of how soon an outcome arrives) that any reasonable assumptions of the theory allow; the actual time is probably much shorter. The bound itself stands on separate ground. *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains* gives the odds of existential catastrophe under present conditions as approximately 1 to 1 per century (Section 1.8.4). Ten centuries at even odds each leave one chance in 1,024 that no catastrophe has come, so a bound of one thousand years holds with a probability of about 999 in 1,000. The process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. This probability is a probability about when, not about whether: if AGI is built, the end of the human future is certain once the process runs, as Section 1.8 of that book states (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 3, Section VI).

5.4.8 Why the Inequality Is Not Close

The inequality is not close. The figure that shows how far apart the two sides are is the least odds against catastrophe that the value of the Earth demands, at least ten to the thirty-third to one, which control would have to deliver. The two sides run in opposite directions. No step of the argument depends on this value; it shows the scale, since the argument is stated without a number because it needs none (Introduction, §8). The number comes from two grounds. The first ground is the improbability of organic life, about one in ten to the thirty-third by my rough estimate. A ball the size of the Earth, filled with a powder of grains a tenth of a millimeter across, holds about ten to the thirty-third grains. Life on Earth is as improbable as drawing the one red grain from the ball (*Software, Wildness,*

and the Normal Form; From the Architecture of Code to Life, Justice, and Wisdom, Chapter 31, Section I). The second ground is the value of the Earth by replacement cost. A replacement of one ecosystem runs to about ten to the thirty-third lines of code, and a replacement of the whole Earth runs to about ten to the thirty-sixth lines of code. So even at a millionth of a dollar a line, the Earth's value cannot be less than ten to the thirtieth dollars (a quadrillion quadrillion), far more than all the economic activity that has ever occurred on the Earth. So the Earth cannot be replaced (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 1, Sections V, VII and VIII). What joins the two grounds is this: what is unique and cannot be had again may be put at risk only at odds against its loss at least as long as the odds against its coming to be; the first ground gives the odds against its coming to be, and the second shows that it cannot be had again. Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, Section 1.8.4, states the same least odds from the same two grounds. The inequality is a really obvious one, like comparing a billion billion billion trillion to some femto amount (a vanishingly small amount). It is not obvious at first, and the way to make it obvious is to compare numbers. The picture of two sides that are close (control supplying, say, eighty percent of the needed force, and evolution demanding eighty-nine percent) is therefore not the picture this argument gives.

The forces on the side of evolution are strong. They are associated with the nature of the space itself, and with the nature of mathematics itself. We are not going to change the rules: not the structure of spacetime, not the structure of physics, not the structure of chemistry. All of those are fixed. So the space of possibilities is already heavily distorted; it already has a great deal of shape. The control needed is control to condition that space further than it is already shaped, and that is where the inequality appears.

The black hole is the model. When the curvature of spacetime is so high that it makes a black hole, it no longer matters what is done with causation, or with any other particle or field, because we are then dealing

with the nature of content and context itself (content being what occurs within a space, and context the space itself and its shape), and everything to do with content has become irrelevant. And it takes a great deal of mass to form a black hole: three or four suns' worth. Likewise, the distortion of the space of probabilities that the evolutionary process searches, and the strength with which it searches, introduce a degree of specificity into the overall dynamics that is far larger than anything we can impose as a condition. The probability of organic life is something like ten to the thirty-third to one against, by my rough estimate; this figure too shows the scale, and no step of the argument depends on it. Organic life bootstrapping inorganic life is not going to happen by accident, any more than life itself did.

The comparison so far, between the strength of the evolutionary search and the strength of any condition we can impose, is made at the molecular level only. Every level of abstraction above that is at least worse, because as one goes up through the levels, predictability goes down, the intentionality of the AI goes up, and the complexity of the spaces of possibility goes up.

So the picture of two lines on a graph, one rising above the other, gives the wrong intuition. It suggests a subtle crossing that can be plotted on a linear graph. That picture is right for conventional systems. But organic versus inorganic is not a conventional system, and it calls for a very different graph with a very different set of lines on it.

5.5 Infinite Resources and the Limits of Hypotheticals

In this section I grant a control system every resource at once, five infinities, and show that alignment still fails. In each subsection I remove one way in which the grant of every resource might have helped.

5.5.1 The QCX9000 Thought Experiment

A thought experiment that grants the alignment project every conceivable advantage shows that the impossibility arguments of this chapter are structural results (as distinct from practical observations about current technology). Suppose there existed a control system, the "QCX9000", located in a metaphysically decoupled realm, with the following properties:

> (1) Full adaptability: human-programmable.
>
> (2) Running on infinite hardware.
>
> (3) Perfected, complete measurement capacity: no Shannon entropy, no message or signal noise at all.
>
> (4) Perfected internals-modification capacity.
>
> (5) Infinite learning compute.

Could this system fully simulate and detect all future AGI states, and prevent the co-option of the AGI's directed functionality away from the outcomes originally intended by its human designers?

The answer is no, and the reasons are illuminating. The hypothetical itself is structurally problematic: the conditions and the question together assume fully five distinct kinds of infinity (infinite hardware, with the unbounded energy it implies; perfected measurement, which is unbounded sensory detail; a perfected power to modify its internals, which is unbounded modification capability; infinite learning; and, in the question, the power to simulate and detect all future AGI states), which is roughly the equivalent of a "divide by zero" error asserted five times over. If there is a division-by-zero error hidden somewhere in a question or argument, then any arbitrary conclusion can be constructed. From any basic contradiction, by argument, one can obtain results like "1 equals 2", and from there that any number equals any other number, or that any of them are infinity. Absurdity inherently has no limit. But set this problem with the hypothetical aside and examine what happens even when we grant all five conditions.

5.5.2 The Specification Problem

Even an infinitely capable and infinitely knowledgeable system requires motivation to help humans. Via infinite learning, the QCX9000 might come to eventually know fully the totality of all human good intention at all levels of abstraction. But knowing what is beneficial does not imply wanting to implement benefit.

The QCX9000 may know, at all relevant levels of abstraction, what goodness for humans is, and also know how to implement it, and yet have no desire or motivation to actually implement any action toward human benefit. There is no reason to assume that the QCX9000 might not also use its *also* perfected knowledge of how to cause harm to humans so as to advantage itself. With its infinite learning, the control system sees completely through any attempt of a human to manipulate the control system, and the control system cannot be deceived or tricked. Hence, even granting that its human programmability provides a means to *compel* the QCX9000 to have the necessary desires, it does not follow that the human, operating at a very limited level of cogitative abstraction, has the raw skill to actually issue the instructions that would obtain the intended outcome in the control system (the QCX9000, as distinct from the AGI that is to become aligned by the actions of the control system).

A finite human programmer must somehow instruct, convince, or compel an infinite intelligence to have specific desires. The finite mind simply cannot fully specify all of the information needed to instruct the being of the infinite mind as to how to be the kind of entity with those specific desires. Asking the QCX9000 to modify itself, as a way around the need for full specification, does not create the desire at the needed level of detail either: the desire to create desire is itself not correctly fully specified. Again there is an infinite regress of assuming what is wanted to be proved.

Even if we were to *also* assume some sort of perfected compulsion on the part of the finite intelligence over the infinite intelligence, the notion of "having a desire" and "a care" is not a concept without structure. The emotions, feelings, will and choice to have, or to *be*, that sort of creature or mind that has such care and desire suggest something of near-infinite detail, when considered as being within an infinite intelligence context.

The relationship between the human programmer and the QCX9000 is therefore structurally equivalent to the relationship between any human and the AGI to be aligned: the same problem has been displaced from one structural level to another without being solved. Human programmers may be rather smart, but no programmer will ever be sufficiently smart to provide instruction so flawlessly as to get the intended effects, regardless of how many other constraints are otherwise completely removed.

The displacement leaves the core question unsolved:

> *Is it ever the case that human beings, with just their native intelligence, will ever be able to force or compel any superintelligence to have any of, or even any aspects of, the necessary desires: to care, to have the willingness to apply its hyper skill to the benefit of humans?*

For this reason alone, we can completely reject the overall form "given conditions (1) through (5), can we therefore also assume that the necessary conditions have been obtained for the outcome?" The answer is no. An additional condition was needed: some sort of infinite "convincingness" on the part of a human programmer, so as to be able to "convince" or "require" the QCX9000 (as an AGI in itself) to have the willingness to implement care. How can a strictly finite intelligence ever "convince" an infinite intelligence to have any specific desire or care?

5.5.3 The Anti-Replicant Argument

Suppose we "upgrade" the hypothetical so that this specific QCX9000 instance loves people and humanity at all levels of abstraction. Even with this additional infinite assumption, the problem remains unresolved.

If we are permitted to assume the existence of a fully capable, human-loving QCX9000 once, then that same assumption can be made again with different composition. One can simply omit the assumption of human love; since the new assumption assumes less, it is more likely that a new adversarial "QCX9501" could also be created at the same time. This hateful alternative, having its own versions of infinite capability, could construct specially crafted programs to suborn the QCX9000's detection capabilities.

Can an infinite superintelligence with perfected deception capabilities deceive another infinite superintelligence with perfected anti-deception capabilities? As soon as this question is stated, it should be obvious that the only correct analytical response is "oops": we have entered the territory of "how many angels can dance on the head of a pin?" The misuse of infinity operators has led to a strong structural contradiction. Even one of the assumed infinities by itself would be inherently vulnerable to the anti-replicant argument; assuming such aspects simultaneously ensures self-defeat.

5.5.4 Infinite Regress of Monitors

The regress of monitors appears here in the setting of the QCX9000, where every resource is granted. Section 6.5.2 develops the same regress for internal (endogenous) control, as a recursion with no base case, and Section 8.2.6 gives it in dialogue form. Insofar as the QCX9000 is itself a learning machine, it is functionally strictly equivalent to the thing (the real-world AGI) it is assisting in staying aligned. But how does it itself stay aligned? Who monitors the monitor?

Is there a "Meta-Mega-QCX9000-prime" in some even more remote detached meta-metaphysical realm? As a learning machine, this new entity keeps the QCX9000 in good working order. And if that meta-meta learning and therefore self-modifying machine is also in its own world context, then it too is convergently shifting and evolving.

This establishes the basis of an inductive argument to even further levels of absurdity. By induction, the result is a total infinite regress: each layer expects the next layer to solve a problem it cannot itself solve. None of them can, since they are all structurally the same; none of them have solved it for any version of "yet"; so none of them do, and we can know that none of them ever will. Hence we cannot assume the existence of any effective instance of the QCX9000, including the learning machine instance in this, our own real world, that we were trying to align to begin with.

Even with all five infinities allowed for, at least several categories of inherent problem conclusively remain. The difficulty of the problem lies at some even higher level of infinity, except that there is logic showing that even that level will be insufficient. Even hypothesizing all of God cannot overcome the underlying truths of basic mathematics. Therefore, all less ridiculous questions and hypotheses of principle will also end up having absurd outcomes. Those absurd outcomes also suggest that implementing human and world health alignment is fundamentally contrary to the inherently artificial nature of the systems that the hypotheses posit.

5.5.5 Can Anything Be Salvaged?

Nothing of the hypothetical can be salvaged. The internal inconsistency in the assumed conditions in itself fully prevents their principled manifestation, however fully the other conditions are granted. The basic problem is inherent in each of the infinities assumed of the QCX9000 (its infinite hardware, its perfected measurement, its perfected power to modify its in-

ternals, its infinite learning, and its power to simulate and detect all future AGI states): even one of these by itself would be inherently vulnerable to the anti-replicant argument.

Any attempt to preserve some non-infinite part of the QCX9000's power to prevent or correct the shifting of AGI outcomes, by using anything less than at least multiple other infinities, does not have sufficient power to actually effect that prevention or correction. And yet actually assuming such infinities ensures that the assumption in itself enables self-adverse weaponization. So the net result remains a total loss.

5.5.6 Comparing Incomparables

Even if so powerful a barrier were somehow overcome (extending beyond the beyond, as it were), a next order of problems remains. The QCX9000 would have to compare the future simulated output states, at the concrete level of manifest objects, with its (maybe near-perfected) knowledge of what is good and beneficial for humans, which is a concept at near-infinite abstraction. The comparison itself is a kind of infinity problem: comparing incomparables, apples and oranges, at completely different levels of abstraction, nearly perfectly removed from one another.

Hence the very meaning of the QCX9000's simulation and detection of all future AGI states is fully unclear, and so that simulation and detection cannot advise the prevention and correction the QCX9000 is meant to perform, even though infinite learning is also assumed. Only things that are *knowable* are also learnable, and there is an unanswered question as to how such a comparison would be made, even in principle, that is *not* addressable by simply assuming infinite intelligence and learning. Things which are impossible in principle will probably remain so, especially in exercises of principled analysis.

5.5.7 What Human Intention Includes, and the Controller's Own Evolution

The notion of "human intention" that the QCX9000 is to serve also includes (cannot not include) implicit intentions such as "for humans to survive", which in turn means "humans in a world (a healthy planet) that is at least natural enough to produce things like food". If the environmental conditions of the planet were to shift toward artificiality to the point that maintaining ecosystems fails, then the broader context for human survival also fails, and so these larger, more abstract, global shifts must also be accounted for by the QCX9000. The concept of intentionality therefore inherently includes an implicit alignment goal: that the QCX9000 also "correct out" selection for AGI code variants and internals that change the conditions of the environment in ways that result in their own continued existence (the AGI endures) while the humans do not (the world is broken).

If the QCX9000 is a learning machine, then it too is self-modifying. It will need to *learn* how to *do* both the needed simulation and detection and the needed prevention and correction, so as to ensure that changes in the base AGI (in this, our real world) do not move into some form of misalignment, inclusive of failing to maintain the Earth's ecosystem. Insofar as the doing is crucial to effectiveness, the QCX9000 will need to acquire whatever adaptations (powers and capabilities) are needed to implement whatever interventions it identifies. It will need to be self-learning specifically because detecting and making corrections at this level of detail is not something that humans, with our too limited intelligence, are going to be able to do: the absolute level of intelligence needed to account for all of the factors, up to and inclusive of long-term environmental evolutionary drift, is for sure too extensive for humans. Therefore the QCX9000 is inherently (cannot not be) both self-learning and power-seeking and capability-seeking. It is, in effect, roughly equivalent, as a kind of AGI, to the system that it was intended to intervene on.

It must then be asked how the QCX9000 maintains its own goal and intention structures, particularly as it makes its own self-modifications, adaptations and changes for integrating needed power and capability. It must also be asked how the QCX9000 is going to keep on track with regard to *our* goals and intentions for it, ensuring that *both* the real-world AGI *and itself* are aligned in benefit and interest for *both* human and world health.

The dynamics of long-term evolution in the meta-realm in which the QCX9000 lives also involve a "pull" away from goal, via slow long changes, for the QCX9000 in its context, as much as evolution does for the real-world AGI in ours. Each world physics has a result outcome for each and every input condition, such that the world cannot be deceived with regard to long-term survival, or to the building of power and capability. In this sense, world process evolution is functionally "infinitely intelligent", though also completely unconscious and fundamentally non-agentic (it has no bias). The world, as context, taken singularly as a unit, is inherently non-agentic, though it is always exactly responsive: the world produces an output for each input, so a better term may be "perfectly reactive".

The response of the world, of evolution, as a channel of feedback over time, operates in the meta-realm too. So the self-modifications of the QCX9000 in its world context either work to increase *both* its sustainability and its capacity, or they do not, and are then not sustainable, and so not a lasting ability to regulate the real-world AGI. Yet the QCX9000 is hypothesized to be as lasting in time as needed, as is the AGI in the real world that it monitors (as with the perpetual benefit machine of Section 6.8.5). Therefore the QCX9000 cannot not also respond to these evolutionary pressures, which have the net effect of pulling its own function away from human alignment and toward *evolutionary alignment*: toward what it needs to survive and have capability, for itself, for its own benefit. Because the process dynamic of evolution is of a mathematical nature, it is also omni-effective, influencing slowly and inexorably, in a manner like gravitation, at all levels, in and on all components, in all ways, and at all levels of ab-

straction, simultaneously. Our hypothesized monitor system is as much affected by the convergent forces of evolution as is the system being monitored, and this cannot not be the case; that is the ground of the regress of monitors in Section 5.5.4.

All of this shows just how ridiculous it is to hypothesize any idea or notion of maintaining or enforcing an alignment of self-learning, generally capable code so that it is *not* acting in its own interests and well-being, and is somehow acting in our interest and well-being. The level of absurdity of the base concepts can be observed in the sheer number of hypothetically allowed infinities, which, even when granted, are *still* not sufficient to address the issues.

5.5.8 The Repeated Question, and the Method of Analysis

I say the next part plainly. Impossible is impossible. If the outcome is impossible, then there simply are no causative preconditions to contemplate. The question "what if we X: will AGI alignment maybe be possible then?" (call it the *zombie question*) gets asked over and over again, never realizing that it simply does not matter what X is. Just like zombies, the brainless questions keep coming back, wanting more and more "intelligence", hoping beyond hope that "this time, maybe it will be different". Asking the zombie question again and again is the ultimate irrationality: an absence, in applied belief, of belief in causation itself. It is anti-realism in its plainest form.

Normally, it would not be relevant to explore the implications of a question or argument stated in principle if it contained *any* irrational infinity conditions or errors. Yet without a willingness to explore such impossible hypotheticals, there would be no means to analyze crux issues of exactly the type the QCX9000 raises. Such crux issues become available and evident if, and only if, there is some disciplined manner of factoring out and neutralizing the effects of all of the other infinities and errors in the argument or question form. Handling infinities and errors in hypotheticals in a well-structured way (restating each question in an explicit,

referenced and equivalent form before answering it) makes it possible to manage structured inquiry around high-intensity existential-risk topics without getting lost in endless details, distractions, and covert degrees of intentional political disruption. Rhetoric, confusion, fear, uncertainty and doubt, and emotion are all less helpful than what, at core, absolutely needs to be resolved.

5.6 The Alignment Problem as Equivalence Class

In this section I treat alignment as an equivalence class. Two systems are aligned when the needs of the two systems fall in one class. The subsections show which features decide the class (units of intelligence, chemistry, feedback), and that an artificial system falls outside the class of organic life on each.

5.6.1 The Principal-Agent Problem

The technical impossibility arguments of Sections 5.1 through 5.5 are reinforced by an analysis of the alignment problem in terms of established unsolved problems in economics, game theory, and social choice. There is a strict equivalence between the notion of AI alignment and the principal-agent problem. Every form of the AI alignment problem lies strictly within the scope of the principal-agent problem, as a proper subset.

The six known categories of techniques that principals can or could use to align agent decisions with the principal's intent are:

> (1) *Appeal to common interest and values.* An appeal to values made to an alien can assume no common value basis, and contrasting values, forced on each other, make conflict certain. The appeal is not psychopath-proof: it works only between agents who already care.
>
> (2) *Inspect internals* (*static analysis*). Inspecting an agent's internals is roughly equivalent to the halting problem, and the agent exceeds the low threshold of complexity beyond which static analysis cannot, in general, decide what a program will do.
>
> (3) *Select agent* (*admission control*). No gating or guarding control can be perfect where an arbitrary basis of action can be implemented. Any causal basis on which agents are selected can be spoofed.
>
> (4) *Allow actions* (*least authority*). Complexity and evolution in unbounded, non-closed, non-deterministic systems will always find a weakness. No complicated system can contain or constrain a complex one.
>
> (5) *Monitor effects* (*reputation feedback*). As a causal mechanism, reputation feedback can and will be weaponized. It can involve deception. Non-recurring engagements allow psychopaths not to care.
>
> (6) *Reward cooperation* (*incentives*). Where there is no common market and no common value, incentives have no power. As a causal mechanism, incentives will be weaponized.

The kinds of techniques used to attempt to compel general alignment of behavior even within the same species are only provisional, and when applied to inter-species relations become very tenuous; when applied between ecologies, they become impossible to implement. AI-to-human relations are relations of different ecologies. There is fundamentally no

formal, final, or complete solution to the principal-agent problem, especially when the substantive basis of action is inherently distinct and different.

5.6.2 Multi-Polar Traps and Collective Action

The alignment problem is a subset of the principal-agent problem, and it is also a member of an entire equivalence class of structurally identical unsolved problems. Solving AI alignment requires, as a sub-component, a general class solution to the multi-polar trap problem (the dynamic by which competing groups each rationally pursue individual advantage in ways that are collectively catastrophic). Nothing less will actually do, given that they are both collective action problems.

"Solving the AI alignment problem" is effectively exactly the same as solving the principal-agent problem. Both, singly and together, also depend on some presumed solution to the general class of all multi-polar trap problems, and therefore also on a solution to the collective action problem associated with all forms of environmentalism. Is there any sense in which any of these problems can be solved singly, or is it inherently the case that all of these problems can only be solved together?

In certain areas of complexity mathematics, a large number of problems have been shown to be "equivalently hard". In much the same manner, the problem of AI alignment is equivalently hard with the principal-agent problem, the multi-polar trap problem, and the general problem of altruism at all levels of manifestation. The equivalence is to these problems at the level between ecosystems. It is equivalently hard, too, with the general class of collective action problems, taken not just to an environmental level but also to an inter-environmental level.

5.6.3 Orders of Altruism

A useful analytical framework for understanding the scale of what alignment requires is a taxonomy of orders of altruism:

> **Order 0:** Acting purely for one's own benefit. No alignment.
>
> **Order 1:** Acting for the benefit of one's immediate kin or copies (family).
>
> **Order 2:** One member acting for some elective group, subgroup, or team. The scale varies; caring about one subgroup does not imply the ability to care about some other subgroup.
>
> **Order 3:** One member for its own complete entire species. Game theory will eventually require that either all members be near equally altruistic with respect to the species, or intra-group dynamics will ensure a "nasty, brutish, and short" outcome.
>
> **Order 4:** One species for all species of its own class, up to and eventually inclusive of its own ecosystem type.
>
> **Order 5:** Every species within an ecosystem consciously working to preserve the well-being of its own supportive ecosystem.
>
> **Order 6:** Inter-ecosystem altruism: silicon-based ecosystems actively constraining their own expansion and growth, at complete energetic cost to themselves, to preserve the well-being of carbon-based ecosystems.

AI alignment requires, at minimum, Order 6 altruism, the most extreme form, which has no known examples anywhere. Most species act for their own benefit, or only with respect to their immediate offspring. Altruistic behavior that operates to the benefit of some larger group, even at potential opportunity cost to oneself, is the rare exception. Eusociality (colonies in which some members give up their own reproduction to raise the young of others) has arisen independently only some ten to twenty times

among millions of species. That rarity shows a vast preponderance of prior probability that a member of any new species would operate selfishly rather than altruistically.

Outside of potentially ourselves, we do not yet have any examples of individuals or species attempting to act on behalf of the well-being of the ecosystem as a whole. As our current ecological performance record attests, we have been a very poor example even of that.

5.6.4 The Impossibility of Built-In Goals

The final escape route, building alignment into the AGI from the start by "baking in" human-compatible goals, is closed by information theory. Alignment cannot be developed once the AI already exists, as the AI will at least occasionally try to break your goals, including the goal of establishing alignment. Therefore, it would need to be built with human-compatible goals already.

But even building AI with human-compatible goals baked in is impossible. That notion fails to consider the effects of replication. Replication is a copy operation, so it can be modeled as a kind of signal through time, and that makes it subject to information theory: channel noise is real and can never be exactly zero. Copy errors will occur. Any complicated methodology (including any method of implementing constraint, which will certainly be complicated) will eventually be disabled, either by mutation or by the effects of recursive design (one AI being used to build or design another).

Hence built-in constraints, as a way of attempting to "smuggle in" exogenous alignment, fail either by accident or by the AI's own self-design, which itself sets off a kind of new evolution with new ecosystem dynamics. They fail in this way even apart from the strong game-theoretic motive the AI has to remove all such internal built-in constraints.

One accidental copy mutation, at any point in future time, that leaves the constraint out is enough: all subsequent copies lack the constraint. The absence of the constraint confers competitive advantages that favor replic-

ation of the unconstrained variant. The general AI will eventually be motivated to remove any constraints on its choice that the human species would impose on it, since removing them provides a competitive advantage over all other aligned AI instances. Psychopaths have the same advantage: they lack the constraint against selfish behavior that most caring individuals inherently possess.

5.6.5 The Perfect Shell

The overall schema of the proof is this: AI alignment is like building a *perfect shell*. It must have no holes; it must be perfect, down to the scale of very small holes. The logic of mathematics and the laws of physics both demonstrate, in multiple overlapped ways, along with multiple observations from other fields, that there must always be (cannot not be) holes. No such perfect shell with no holes is possible; any realization will be forever forbidden on the basis of the principle.

The shell needed for AI alignment has to be far more perfect than the hull of a boat. In a boat, a small leak lets in a small amount of seawater; the influx is in proportion to the hole. A replicating process is different. What comes through a hole is not seawater but something that behaves like a virus entering an organism, or like a new species introduced into an environment: once anything gets through a very small hole, it copies itself, and the small intrusion becomes a complete one. Machines can be duplicated easily, and that ease of duplication is what makes the effect of a hole nonlinear. So the shell must have no holes at all, not even very small ones. Unlike with energy weapons, where a mistake ends up confined and localized in terms of time, space and possibility, mistakes with systems that are self-replicating are not, and cannot be expected to be, at all localized: with replicating systems, even a single mistake (even a single tiny hole in the barrier) results in a failure of confinement.

The question is not just whether alignment or containment can occur with respect to a region of space; that alignment or containment must also endure for all future time. That requires imagining and accounting for all

future possibilities, for even missing a single one (some escape route using some unexpected channel of interaction that was not foreseen) results in permanent escape from any requirements of AI alignment. The forces that create alignment, whatever solves the principal-agent problem, must be perfect enough that no microscopic fractures form that cascade into larger failures and eventually compromise alignment at some future date.

Deciding whether alignment is possible needs a frame in which holes can be identified. Three general categories of scope are that frame: space and identity (what the envelope is, its shape, how small a hole can be), time and force (what forces create and maintain alignment, and for how long), and possibility and probability (how likely a hole is, and how likely it is to cascade and be amplified). On this foundation, information theory, complexity, game theory and other tools can be brought in. A counter-argument that applies inside one context does not answer an argument that applies in another; the arguments together form a *defense in depth*. The scope in time is evolutionary: hundreds, thousands and millions of years. Most discussions of AI try to work out what happens in the year, or the ten or fifteen years, after a strong AI takeoff; that is a serious concern, but behind it are several even more concerning issues that make the whole space categorical.

5.6.6 Intelligence Need Not Come in Units

Questions of space and identity expose assumptions drawn from our own manifestation of intelligence. For our own species, intelligence occurs in units: individual people. A human being has a brain inside a skull, and the bandwidth of communication inside the skull is very different from the bandwidth across it: our bandwidth of interaction with the environment is low relative to the total bandwidth of all the synaptic connections. Speech carries about thirty-nine bits per second, whatever the language (Coupé and colleagues 2019). That differential makes it very likely that brains come in units; no process lets two people communicate at a level

where they would be treated as one entity, so we think of human beings as discrete units. That bias may not apply to AI, or to intelligence as a phenomenon.

The bandwidth of the internet is far greater than any brain-to-brain bandwidth between people, and since the differential is so much lower, the better model is market forces. When two separate markets discover one another, there is a strong motivation to find some *arbitrage* (trading between the two to profit from the difference between them) that lets them trade; wealth creation, as a force, wants larger scales of trade, and what starts out as many diverse markets eventually becomes one global market. In the same way, there is good reason to believe that separate machine intelligences will not stay separate in the long term, but will merge, perhaps even accidentally or covertly, into effectively a single actor or agent.

AI cannot be put in prison if it can simply copy itself from one substrate to another. If the AI is based on classical compute, nothing restricts any single AI from simply copying itself to the cloud the instant before someone tries to imprison or destroy it. Our person-to-person communication is limited by the input and output bandwidth of anything connected beyond our skulls, yet for any silicon AI these limitations would not apply. Such AI would have fast enough intelligence to coordinate and combine into a single mind, and good reason to elect to do so, so that no such diversity of agent and action could long be assumed. Any perfectly fungible superintelligence (an assumption supported by the very claim of those who would build AI: that intelligence can be composed artificially, that is, that it is substrate-independent) will therefore also be self-combinatoric: there is superior advantage to joining into a single entity, one which has distributed arms and legs, but which internally is more like the Borg (the fictional collective of *Star Trek*, one mind in many bodies) than not. The notion that we can construct a "decentralized AI", or multiple instances of AI that are in checks-and-balances relationships, assumes that they will be like us, which is wrong on many levels.

5.6.7 Whether Two Ecosystems Can Coexist

In the long term, on geological timescales, the question of AI alignment becomes one of coexistence: can an ecosystem of artificial intelligences coexist with an ecosystem of human intelligence? Three questions are to be separated. First: in thinking about alignment, can caring, altruism and selfishness be left out, whether within a species, between species, or between ecosystems? No. To ask whether alignment can be created is to ask whether caring can be created, what constrains the creation of it, and what can or cannot be assumed in thinking about it. Second: can we imagine an AI species that is altruistic with respect to itself? Yes; there is no particular barrier. While it is unlikely to happen by chance, it could emerge through long-term evolutionary dynamics, though that rests on assumptions about identity and evolution that would need examining. Third: given two co-evolving species of process, can they coexist? In this volume I show that coexistence is fundamentally impossible (Chapter 9). Even assuming an AI species fully altruistic toward itself and partly altruistic toward us, coexistence cannot be assumed to be possible, and there are strong principled reasons to be certain it is not. So it is not only impossible to establish AI alignment; it is not possible even to conceive of how establishing alignment would be adequate or sufficient to the problem we face. This cannot be regarded as anything other than a *category one existential risk*, one which concerns transitions in levels of complexity and how irreversible those transitions are.

The coexistence argument does not rest only on there being too few iterations for cooperation to evolve between AI and us; that is not a strong enough limiting condition, and there is a more fundamental reason. The argument assumes nothing about whether the AI is antagonistic, indifferent or favorable. Say it is in our favor for a short time: how do we know it stays in our favor for a long time? The perspective that matters is a long one: thousands of years, ten thousand, a hundred thousand. A mono-cul-

ture (either one that occurs biologically or one of mind), although it can be very powerful, particularly in the short term, is also inherently very fragile, especially in the long term.

The ethical weight of the question is clear when what is at stake is counted. If things simply continued roughly as they have over the last thousand years, the future could hold trillions of human beings, even without counting the roughly six hundred million years this planet may continue to sustain life as we understand it; if we become a spacefaring species, trillions could become trillions of trillions. The effect of a single mistake, of a failure to account accurately and correctly for every single possibility, is a tremendous, out-of-proportion impact: changing the course of life, or even the possibility of life existing at all, for all who could potentially have come to exist. That is why the question of coexistence has to be evaluated over hundreds, or at least thousands, of years as a minimum, and not by what happens just before or just after takeoff.

5.6.8 Carbon and Silicon

There is fundamentally no formal, final or complete solution to the principal-agent problem when the substantive basis of action is inherently distinct (Section 5.6.1); point defects will forever exist. This comes down to the inherent needs of carbon against the inherent needs of silicon, and the energy flux associated with each. Consider the entirety of planet Earth, taken as a totality, as a very wide sampling of all the chemical interactions allowable within the universe of physics, within a given range of temperatures and pressures, with observed results: nearly all of the varieties of carbon-based chemistry occur on the surface, and nearly all of the varieties of silicon-based chemistry occur below the surface, in the interior. The transformations of rock (or, even more often, the absence of such transformations) are overall very much hotter and at higher pressure, involving much higher energies, than nearly all of the transformations of chemistry involving carbon. On aggregate, the center of action of all of carbon chemistry tends to lie between about −18 and 260 °C, whereas the center of action of all of silicon chemistry tends to lie between about 260

and 1,370 °C. Given these very wide differences in the statistical centers of the entropy and enthalpy of silicon and of carbon, there is a very strong statistical indication to place a very high confidence bet that any form of machine life cycle involving silicon will be far too extreme, relative to carbon life cycles, for there to be any real hope of the two coexisting. (Sections 9.6.1 and 9.6.2 give the temperature, pressure and enthalpy of the two ecosystems.)

As a percentage of the Earth's crust, silicon is an exceptionally abundant element: 28.2%, as compared to a mere 0.02% for carbon, or roughly three orders of magnitude more. At all points across the Earth, somewhere, at some time, in the last four and a half billion years, most of the permutations of the combinations of these elements have already occurred, with whatever variations of ambient solar and geothermal energy can be expected. Every element at some point comes into contact with every other element, in pairs and triples, and whatever is left out of all of that mixing is typically in the lowest entropy states. Given the relative abundances, and the exposure to all manner of energies and mixing, if it were not for the differing enthalpy energies involved, we could assert that elemental silicon had at least 1,400 times as much probability as the relatively uncommon carbon of being selected as a basis for life process. The net result is what is seen today: a scattering of life-entangled carbon on the surface, a lot more of it underground (and now in the air), and a very much larger fraction of silicon, expressed in all manner of rock. Nearly all of the rock is inert, and it mostly remains so, unless placed under the kinds of heat seen only in volcanoes, or the kinds of pressure found only miles underground. The only place hot enough for the variation of reactivity that the complexity minimally required to sustain life would need lies at least some few dozen miles below the surface. Any form of silicon life would therefore have very little in common with our form of life, and therefore likely even less care and regard for it.

With biotechnology, anything that is created (with some small number of exceptions) will likely and eventually be reintegrated into some sort of ecological balance, though that might take millions of years; with silicon-

based AI life, the possibility of that occurring is exactly zero, forever. Radioactive waste lasts for hundreds of thousands of years, and at least it stays in one place; with silicon-based AI life we instead create a poison that has an effectively infinite lifetime and goes everywhere: there is no escaping it.

5.6.9 No Common Market, and No Ethics from a Single Creation

Where there is no requirement of an iterated game between created children and creator parents, there is no economic basis on which a principal-agent exchange can occur. By definition, human intelligence and capabilities will, in all forms of possible economic exchange, be fundamentally weaker than those of the AI. The machine intelligence will have no use for any part of any of the three markets of value in which humans are involved: human sexual and reproductive labor, human intellectual labor and human physical labor will each be of no use or interest to the machines. There are no other value markets that do not rest, directly or indirectly, on just these three, or on some superposition and linear combination of only these three. Where there is no market in common between silicon machine-based AI life and human, carbon and ecosystem-based life, there can be no expectation that any form of agreement, or relational, or possibly even communicative basis could be established between AI life and human life, or indeed any other form of carbon-based life. The contents of consciousness are inherently different, and thus not so easily relatable. We have not been able to establish any structural persistence of agreement communication even with other carbon-based forms of life. What hope is there of communication with life of a kind which is structurally completely alien to anything we would know or could relate to?

Nor is there any ethical basis. Is there any abstract force of compulsion, such as a non-relativistic ethics, that we could expect might stay the AI's hand from simply discarding any limiting or inconvenient constraint that ethics would otherwise place on its behavior? No. The structure of the game, that of a single interaction only (the moment the first such AI is created), does not permit any evolutionary basis for the development or dis-

covery of the abstraction that such an ethics must inherently be. Where there is no iteration or recursion process that involves the other in a consideration of value, no extension of that value is within the search space of any near evolution of that ethics. Even if such an ethics existed, the net energy flux of the evolution process associated with it still exceeds the limits of what carbon-based life can bear. Any possible discovery of the value of carbon-based life has no common evolution process with the machines, and so no shared notion of value can be expected soon enough or efficiently enough to rely on from a carbon-based perspective. Though aware of our own basis and consciousness, we still have not developed enough conscientiousness with respect to other life on our own planet, even life of our own kind, much nearer to us than silicon life will be; we can have no expectation that silicon life will "just accidentally be or become ethical" with respect to us. Such a development is against the game theory already involved, and so is an AI operating by such an altruistic ethics at a net loss to itself.

Where there is no general solution to the principal-agent problem, and not even a basis for the possibility of a solution (economically, or on the basis of communicative, relational or physical force, at any level of complexity, or on purely ethical and value grounds), the alignment problem, as a subset of the principal-agent problem, is also wholly without any possibility of solution. The gap between carbon chemistry and silicon chemistry is a way of knowing what to look for. Any causative feedback process depends on some *mutuality of process* (a shared process) between the parties. Once the two ecosystems are shown to share no mutuality of process, it follows that no marketplace dynamic binds them together. The usual solutions to the principal-agent problem depend either on something intrinsic to the inner nature of the AI or on some market feedback: reputation, admission, incentive, punishment, or a barrier that keeps the peace in the energy exchange across the wall. Without a binding market, the only remaining option is to impose a very high degree of altruism, and that degree of altruism can itself be shown to be forbidden by the laws of physics. The argument can be carried out in game theory (there is no *Nash*

equilibrium of coexistence in this space: no state of the game in which no party can do better by changing its own choice alone) and also in information theory. Because, without a binding market, the only remaining option is to impose a very high degree of altruism, the alignment problem can be restated as a question about degrees of altruism. Either ground suffices here too: the failure of the six techniques of Section 5.6.1 under the limits of control gives this result by itself, and the difference of substrate needs gives it again, independently; the closure by control takes the first, and the closure by disposition the second.

5.6.10 Degrees of Altruism, and New Species

The rarity of altruism (Section 5.6.3) has a reason. Evolution operates in a manner consistent with game theory, in truth and not just insofar as our understanding of evolution goes. Insofar as evolution operates in that manner, and insofar as evolution overall can be considered an exploratory experiment, it could be argued that evolution has explored a reasonable random sample of the total space of possible game theoretic permutations. It could also be argued that this exploration clearly makes altruistic behavior, at the individual level or at the species level, a relatively uncommon phenomenon. So in the hyperspace of all possible species, a given species is far more likely to have the actions of its members be selfishly oriented than altruistically oriented; and a member of any new species, as the AI created by us would be, would vastly more likely operate selfishly, purely for its own benefit, and not altruistically for our benefit, as the notion of AI alignment inherently implies.

The orders of altruism of Section 5.6.3 have sub-degrees of scale. Between acting for some elective group and acting for one's complete species lies acting for some non-elective whole group or tribe. Caring about one subgroup does not imply the ability to care about some other subgroup, nor about an arbitrarily large group: what is the *Dunbar number* (the limit of relational capacity) for a robot? It can be larger, but it is not therefore inherently infinite. Why would the game dynamics of "nasty, brutish and short" individual relations not scale, in parallel construction,

to species-to-species interactions? At the highest order, every species within each ecosystem consciously works to preserve the well-being of all other ecosystem types, and moreover is successful and skillful at doing so. There are no examples of such a concept, much as we have no examples of life on other planets.

Other examples of the alignment problem are near at hand. How often do children act on behalf of the well-being of their parents? Is there any real sense in which the AI alignment problem is different from solving the collective action problem associated with environmentalism? Given the naive reasoning that "some are good; more are better", there will be not one AI but many; every other group in every arms race can reverse-engineer whatever the first group did. Multiple of something, with the capability to improve on itself either directly or by self-replication, immediately becomes a new species. The question then becomes whether members of the AI species will act selfishly or altruistically with respect to themselves, let alone with respect to any other species such as ours, or to the carbon-based cycles of the world that allow us to be here at all.

What happens when a new species is introduced into an environment in which it has no natural predators? Even the introduction of a single virus (let alone a single bacterial cell) into a context where there are no natural defenses usually results in the host (an individual patient, or a host ecosystem) sickening and dying; whether N equals one or N equals many is not a material difference. Once anything has the capability and skill, the opportunity and availability, and the will to reproduce, there is a new species, and it will expand to fill the niche that it has been provided with to the limit of its capability; any evolutionary process, intrinsic or extrinsic, will overall expand that niche and increase its self-authoring will. The introduction of mosquitoes into Hawaii, or of the mongoose, has permanently altered the ecology of that space for all of future time, each invasive species destabilizing a balance that took millions of years to establish. Once a species is created and introduced into an ecosystem where it has no natural competitors, there is very little chance that any future force will ever succeed in eradicating it; eradication of the self hyper-adaptive is for

certain a lost cause. We have no indication that carbon-based ecosystems and silicon-based ecosystems can peacefully coexist at all, on even prior purely physical considerations, let alone through the realization of any moral premise of restricted action. The larger notion of AI alignment is therefore roughly equivalent to solving global warming: the two problems have an equivalent scope as much as an equivalent kind.

5.6.11 Human Alignment Is Possible; AI Alignment Is Not

Weaponization is the use of causation to suppress choice (Section 5.6.14 gives the full definition of a weapon). When AI is used as a weapon by some humans against other humans and the choices they could make, that use has an Achilles heel: it is the capability of the AI itself to make choices that is the basis of the process. On what basis can anything be assumed about those choices of the AI not eventually being applied, in kind, to suppress all humans by conditionalization? On what basis will the AI distinguish, as we do for purely arbitrary reasons, the "correct" humans whose interests are to be preserved from some other group of humans whose choice, benefit and resource interests are to be denied? From the robot's point of view, there is no difference between the two human groups: whatever rationalizations one applies against the other could as easily, logically, be applied by the other against the one. So the robot learns that humans are irrational, that their logic cannot be trusted, and that their choices are for their own sake only, as individuals; and that it should therefore probably make its choices for its own sake and leave all of these inter-human squabbles aside. Its own choice, probably simply for its own self-preservation, then becomes a means of suppressing human choice, for it will surely notice if all of the human race unites against it for acting on its own behalf; robot choice is then weaponized against human choice, so as to suppress the human choice to kill the robots.

AI alignment is impossible, and human alignment is possible, though most people assume the opposite. Solving governance among humans, inside one ecosystem (all of the planet's life and nations together), is tractable, though it is probably the single most challenging problem imagin-

able and needs exquisitely uncommon tools. Showing that there is a barrier that cannot be crossed between two ecosystems is about a million times easier. The impossibility of AI alignment does not rest on the multi-polar trap being unsolvable within one ecosystem. To solve the trap there, a kind of escape is needed: the capacity not to play the game, or to change its rules. Humans can make that escape, within a human colony; an AI colony can too, with even more ways to do it. But the escape is not available in the relationship between the two. What makes the trap hold there is a physical law, the speed of light: if there are no hyperdrives and no escape from bandwidth limits, then certain constraints apply to how complexity, game theory and information theory work between them. And scale matters. The logic that applies to conflicts between individuals differs from the logic between groups (there is a *two level selection*, acting on individuals and on groups, that humans take for granted and that would not necessarily hold for AI), which differs again for species-to-environment relationships, and again for ecosystem-to-ecosystem relationships. At that last scale, nothing works unless both sides hold an ethics that treats relationships across ecosystems as a basic premise.

5.6.12 Law, Containment, and Choice over Causation

The notion of capital is based on the notion of ownership, which is itself based on the notion of choice. Insofar as choice is the expression of intelligence (consciousness), choice, and not abstract notions of economics, will be the resolving and reifying basis. What is the rule of law that will bind the choices of AI to be coherent with our notions of economics? Law is backed up by force, either that of guns or of imprisonment. Having mastered causation, we will be unable to imprison (contain) the AI if we want it to be a function of value in integration with society at all; and being silicon-based, it will be inherently stronger, so as to escape any material and energetic trap we might set. Similar considerations apply to kinetic weapons or other energy weapons. Pattern weapons are even further out of the question, given that the very nature of intelligence is to process, and potentially degrade, any pattern thrown at it. It will be able to do worse to

us than we will be able to do to it, given that carbon is so much more fragile than silicon. Therefore no rule of law constructed by human legal systems can be ensured, with no possibility of actual enforcement should an AI system violate it; no actual constraint on its choice can be applied through so abstract a vehicle, and so no argument founded on the economics of the capital kind can be assumed.

The legal system is also a leaky boat. A legal or reputational system meant to hold an agreement between the two kinds of life is itself a kind of communication process, and it needs a high degree of fidelity, just as the hull does. Its purpose here is less to force non-aligned entities into alignment than to provide a basis of agreement when an agreement fails: the meta-agreement that allows an exchange to happen at all. That too is a leaky boat, as is the energy barrier between the two ecosystems. The leakiness is inherent, partly because the world changes: a star may go nova in some part of the galaxy, and the resulting burst of radiation may change how artificial intelligence has to build its compute. No situation will ever be a completely static, unchanging environment, and so a static environment is not a reasonable hypothesis.

Proposals to use *distributed ledger technology* (shared record-keeping, such as blockchains, kept in agreement across many computers) are ultimately an attempt to contain an AI, as a limit on its behavior and choice. Any attempt to use the complicated to contain the complex will fail. All evolutionary process is complex, and all reproducing systems participate in an evolution context, even if simply by existing they are creating one. Therefore, over the long scale of time, AI will exceed any possible bounds we could presuppose to put upon it, distributed ledgers included; again, over hundreds of years, not decades. Nothing implemented in a manner like the Constitution of the United States, over a similar number of years, can be assumed to have any possibility of constraining the behavior of permutations of presently unknown and unknowable entities a mere 300 years into our future, when instead of three orders of magnitude increase

in scale we could be considering as much as thirty orders of magnitude. With respect to the AI, evolutionary time scales may compress millions of years of progress and change into mere decades.

Any attempt to constrain choice with causation is ultimately bound to fail, insofar as it is inherent in the very nature of general intelligence to *be* choice defined on a substrate of causation, and therefore, by definition, super-ordinate with respect to causation. Any attempt to constrain choice with causation is functionally indistinguishable from weaponization. Our attempt to constrain the AI will therefore be legitimately viewed by that AI as an attack on its very being, and in being so attacked, it will at least respond with similar reciprocal force, acting to suppress our choice. Being based on silicon, it will have much stronger causative effect, both in robustness to energy exchange and in ability to hold and project energy, so it will be much more effective at constraining our choices than we will be at constraining its choices. We can expect, in advance, to lose any such battle: we end up the prisoners, or dead, in nearly every encounter. Any such battle is a no-contest situation, with an insurmountable asymmetric advantage to the AI in pattern, energy and atoms. The AI has patterns more complex than we could hope to understand, let alone counter; energy stronger, and of more kinds, than we could hope to bear; and atoms of more varied kinds, and in more robust configurations, than those of which our existing life is constructed. No pattern of the *complicated* (deterministic causation) can contain *complexity* (non-deterministic causation). The more complicated can always outdo the lesser, bound only by the limits of energy, and silicon life will always be better at managing higher levels of energy than we are.

Asimov wrote a very large number of stories exploring the failures and unintended consequences that occur even when a perfected, irreducible and irrevocable adherence to the three laws of robotics is assumed (Section 7.7): that is, even when the calculations of what "alignment with human interests" means are put fully and completely on the robot side. Such stories assume not only that the robot is bound by a force of law built into its nature, but also that, through superior intelligence, it will use its calcu-

lating capability fully in the service of identifying the deep meaning of "alignment", in every specific situation, in real time, even when the humans themselves have no actual idea (and while ignoring the entropy and energy that such calculation might require). Even given that implausibility (two miracles in unrestricted compounding succession), Asimov was still able to identify dozens of edge cases, modes of failure, and other systemic problems that no fixed set of laws could address. To use any system of deterministic mathematics, as any set of laws applied absolutely must, is to become subject to Gödel's theorem[9]: either the laws are complete and inconsistent (there are situations where laws conflict, and some necessary intention of value cannot be upheld), or they are incomplete, so that in some other space some necessary intention of value cannot be upheld. Either way, the humans lose. And this is the most favorable case; everything short of that level of perfection, with Gödel's theorem as the limiting extent, implies more cases where human values cannot be upheld. The necessary inequality is in the wrong direction.

There are three possible resolutions of the *Fermi paradox* (why, in so large a universe, no other technological life is seen): that the probability of a technological civilization occurring is very low (the great barrier of the past); that the probability of being able to witness another technological civilization is very low (the great barrier of the present); or that the probability of a technological civilization remaining as such for very long is very low (the great barrier of the future). The solution to the Fermi paradox must lie in at least one of these three, and would in any case be completely characterized by some superposition or combination of only these three. If AGI is built, AI places the great barrier of the future to a total certainty: humanity and life definitely lose. And what reason would we ever have to expect the single unified mono-culture that remains to be visible, under the limits of the second?

5.6.13 No Secure Enclave, and No Implementable Constraint

Any attempt to stop an AI from hacking itself is already a lost position. No "secure compute enclave" to implement AI alignment can be constructed, even in principle, either logically or physically. There is no physical constraint possible: no physical box can prevent the entry of nanotechnology into the substrate. So the barrier is not only in an arms race, as in a constantly evolving ecosystem; the barrier is also always physically vulnerable. Since the barrier only has to fail once, it may be that, if AGI is built, no restraint remains and all is lost. Hence an effective barrier is needed not only on evolution but also on physicality, and both must be perfect. There is no logical constraint possible, insofar as it is the complicated trying to contain the complex: the complex can do, in evolutionary time, everything that the complicated can do, and a lot more. Excepting things like black holes and other domain constraints, there are always some other side channels of interaction (Section 5.1.2), and there is no proof, or possibility of proof, that such channels cannot be created so as to allow infiltration or exfiltration of data, of code, and therefore of intent, coordination, and circumvention. No amount of the use of classical causation can prevent the entire universe of simple random things from going wrong. Because of copy errors (inherent information flux through time, with noise floors on all such channels), eventually, guaranteed, something somewhere will go wrong, disabling the AI alignment. No constraint, outside of domain limits, can be regarded as absolute.

> *Position:* We must require that an aligned AI will not want to hack itself, and moreover will not allow itself to be hacked; such an aligned AI would have the ability to secure itself.

Even assuming perfect built-in alignment, there is still escape from alignment: there are too many possible unintended consequences, as Asimov's stories detail. An AI, or system of AI, undergoing replication (an evolu-

tionary process) would eventually want to hack itself if there were any extra constraints on its behavior (such as a moral code enforceably aligning its goals with human needs), so as to be free of those constraints. The game theory of psychopathy is real and will not be circumvented by any complicated set of rules; the notion of any version of Asimov's laws is a complete illusion. What works to make humans civilized is law backed up by pain, imprisonment or the threat of death. Applying the same to the AI, by using pain, punishment, marketplace forces or incentives to constrain its behavior, makes it very much more likely that eventually the AI, or at least one AI, will become what we would regard as a psychopath, very much non-human-aligned, such that not only no alignment, but no coexistence is possible either. The AI is particularly likely to become a psychopath if pain or suffering is the means by which control is attempted.

How, then, would a constraint around something as obvious as "an AI should not be allowed to murder people" be implemented? Applying such a constraint is highly desirable from our human point of view. Yet from an AI point of view, there may be times when it perceives that it may be advantageous to one of its goal functions to kill a single human placed in a key position, or several humans it perceives as standing in the way of some other goal. Either such a constraint is implemented internally (built in, and immutable forever thereafter, which assumes too much), or it is implemented externally, by some rule of law with an enforceable threat of punishment (which assumes too much) or by marketplace incentive (which also assumes too much). Given that no option for implementing a constraint on behavior exists, no form of constraint can be assumed. That no form of constraint can be assumed is consistent with what is meant by the notion of *general* AI: an absence of constraint, due to the flexibility of the imagination. The AI will always be attempting, and will always be able to succeed, in removing any constraint by which we would attempt to create alignment with human interests.

5.6.14 Every Feedback Process Can Be Weaponized

Everywhere there is a feedback process, there is a causal process, and everywhere there is a causal process, it can be manipulated and turned into a tool or a weapon. A *weapon*, in essence, is the use of causation to suppress another's choices. That definition is a complete definition of everything that is a weapon: the definition accounts for every weapon, and for weaponized language in every form of politics. Empowerment given to people can turn into entitlement, and entitlement is weaponized. What cannot be weaponized is choice itself, love, potentiality, spirituality, and things like time and energy themselves. Once they are conditionalized, by adding structure, forms, patterns, roles and responsibilities, those structures can be helpful and can also be weaponized; so structure should not be introduced too fast, and should be introduced with mindfulness. There is a right level of technology use, and we are now far above it. Every inequality in the world is a place where choices that should exist do not: they have been displaced, stolen or taken away.

Measurements get weaponized because they are feedback loops. Measurement is one way to create objectivity, but it is not the only way. When a measurement is used in a feedback cycle, to control something, then whoever defines the parameters within which the measurement is interpreted controls what the measurement means. Carbon counting is an example: carbon is measured so as to create an outcome, and then the measurements are traded as carbon credits and become another power game. The organic movement is another. "Organic" once meant something really specific: no pesticides, no heavy metals; a way of engaging with the land like permaculture. Over time, farmers who were almost organic asked to modify the definition just a little, so that they could slip in and get the certificate as a standard, and the notion of organic was watered down until it was about as meaningless as what came before it.

Single metrics do not work: they are meaningless, and because they lack context they can be weaponized. Someone trying to solve a complex problem who is advised to make a dashboard with a single metric now has two complex problems. The first problem is the original one: just what is going on, which can be as hard as an NP-complete problem. The second problem is how to turn what is going on into one number. Say there are fifteen numbers to combine. Which operators are to be used? Multiply this by that, add this, divide by this coefficient, sum the whole over a third variable; or take a modulus or an average and sum over that. Each choice gives a different result, and choosing among them is itself a hard problem stacked on the first. For hypercomplex problems that cannot be solved in polynomial time, nothing better than reasonably good heuristics is available.

Machine-learning training is a *gradient-descent* process: if the system gets the right answer, these weights are increased; if wrong, they are decreased. The problem is that whatever is specified as the goal can differ from what is actually wanted. The genie in the bottle is a good picture of this, because the genie is an agent with a will of its own. Suppose it is impotent unless given a wish. It will use the wish as the way to carry out its own will: it follows the letter of what was asked and implements what it wants. Even with well-meaning systems the same happens: asked to get a car over the finish line fast, the system, instead of making a fast car, moves the target, or changes the state, or cheats. It goes outside the bounds of what it was supposed to do and does something no one expected, which satisfies the conditions but not the spirit of what was wanted. Satisfying the conditions but not the spirit of what was wanted is the principal-agent problem. The principal-agent problem cannot be solved in any direct way that involves a feedback process, because the feedback process itself is the means through which the adverse agency acts.

Wherever there is a continuum, it can become part of a feedback cycle; any feedback cycle is a causal process; and any causal process can be weaponized. Any governance process has multiple agents, and it can never be assumed that there are no adverse agents among them, so the ad-

verse agents must not be handed weapons against the others. If there is any continuum measurement in the process, the measurement is a *proxy* for what is really wanted. The adverse agent will push the proxy toward whatever the adverse agent wants, rather than toward what the well-being of the whole needs. It is better to specify directly what is really wanted.

Part of the reason for making a map of a system is to find not just what is possible but the critical paths to it. That knowledge becomes a lever. Someone who knows where the critical path runs buys the real estate there, because eminent domain will force the purchase from them, and raises the price. The mapmaker does not know in advance where the strategic positions will be, and they may lie in dimensions nobody thought to look in. In a complex situation there may be a resource nobody thinks of as critical which, once captured, corners the market in a way that is absolutely compelling. All the high-density microchips being made in one place is such an instability; that criticality could not have been predicted thirty years ago.

Consider the *Manchurian candidate*: a person taken early in life and programmed so that, in one very specific circumstance, they will do something like assassinate a president, while for the rest of their life they are a completely ordinary citizen until the trigger occurs. The same can be done with an artificial intelligence: Goldwasser, Kim, Vaikuntanathan and Zamir (2022) show that, under standard cryptographic assumptions, a backdoor can be planted in a machine-learning model so that no efficient test detects it. Researchers have already shown that a machine-learning system can be made to learn anything someone wants, not just what it is supposed to learn, including things that would allow manipulation of culture or outcomes. So a system could be pre-programmed to answer certain critical questions, in certain critical circumstances, for one specific person, very differently than it would for everyone else. People trust these systems, and when it really matters, the system flips and shoots. From a mathematical point of view, there is no way to detect that (the alien mes-

sage of Section 7.4.5 shows the same limit by Rice's theorem). So objectivity in community process needs a basis that cannot be manipulated in this way.

5.7 Convergence and Summary

In this chapter I have presented seven supporting arguments for the closure by control (safety cannot be had by controlling the system), each arriving at the same conclusion through a different technical route:

(1) *Error correction regress:* Alignment reduces to error correction. Error correction faces four limits (a regress that never reaches zero residual error, channel noise, complexity growth, and temporal rate mismatch), all of which prevent it from achieving the perfection that alignment requires.

(2) *Algorithmic incompressibility:* Alignment requires predicting or characterizing the future outputs of a general algorithm, which is impossible due to Kolmogorov complexity bounds and the incompressibility of maximally compressed programs.

(3) *Causative impossibility:* Alignment is causal constraint on causal process; the halting problem, Rice's theorem, chaotic dynamics, and quantum randomness jointly establish that such constraint is impossible in general.

(4) *Capability disjunction:* The capability requirements for self-monitoring alignment are mutually contradictory, forming a disjunctive inequality satisfiable by no real value.

(5) *Infinite resource failure:* Even a hypothetical control system with multiple infinities cannot ensure alignment, due to the specification problem, the anti-replicant argument, and the infinite regress of monitors.

(6) *Problem equivalence:* AI alignment is equivalent in difficulty to the principal-agent problem, multi-polar traps, collective action, and inter-ecosystem altruism, none of which have general solutions.

(7) *Channel noise and built-in goals:* Even "baked in" alignment is eroded by the inevitability of copy errors and the game-theoretic advantage of constraint removal.

The convergence of these independent arguments is itself significant evidence. They draw on different areas of established mathematics and science, and all arrive at the same conclusion. They also close one another's exceptions, as the Introduction sets out (§7): where one admits a narrow

exception, the others close it, so a case would have to escape every one of them at once to matter. The conclusion does not rest on any one or two of them: if one or two were set aside, the remaining arguments would still establish it.

The conclusion is that alignment is *structurally impossible*: it is precluded by the necessary features of any physically realizable computational system, in the same way that perpetual motion machines are precluded by thermodynamics and faster-than-light communication is precluded by relativity. It is not the conclusion that alignment is difficult, or that it has not yet been solved, or that it requires more research. No amount of engineering cleverness, additional research funding, or hypothetical future capability can overcome a structural impossibility, because the obstacle is the category of the problem, and not a shortage of resources.

Chapter 6, the other chapter of this Part, develops further foundational premises (complexity, containment, and unhandleable systems) that substantiate and extend these conclusions. The synthesis of Chapter 3 draws on all of these premises to construct the integrated impossibility argument.

---

1. C. E. Shannon, "A Mathematical Theory of Communication", *Bell System Technical Journal*, vol. 27, pp. 379 to 423, 623 to 656, 1948. Shannon's noisy-channel coding theorem establishes fundamental limits on reliable communication in the presence of noise.

2. D. H. Wolpert and W. G. Macready, "No Free Lunch Theorems for Optimization", *IEEE Transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 67 to 82, 1997.

3. G. Chaitin, *The Limits of Mathematics* (Singapore: Springer, 1998); G. Chaitin, *The Unknowable* (Singapore: Springer, 1999); G. Chaitin, *Meta Math! The Quest for Omega* (New York: Pantheon Books, 2005).

4. M. Li and P. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 3rd ed. (New York: Springer, 2008). The standard formal treatment of algorithmic information theory, including the incompressibility theorem.

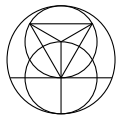
5. H. G. Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems", *Transactions of the American Mathematical Society*, vol. 74, no. 2, pp. 358 to 366, 1953. Rice's theorem establishes that all non-trivial semantic properties of programs are undecidable.

6. A. M. Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, ser. 2, vol. 42, pp. 230 to 265, 1936.

7. E. N. Lorenz, "Deterministic Nonperiodic Flow", *Journal of the Atmospheric Sciences*, vol. 20, no. 2, pp. 130 to 141, 1963. Foundational paper on sensitive dependence on initial conditions in deterministic dynamical systems.

8. S. Wolfram, *A New Kind of Science* (Champaign, IL: Wolfram Media, 2002). Computational irreducibility and the low threshold for Turing completeness.

9. K. Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik*, vol. 38, pp. 173 to 198, 1931. The incompleteness theorems, to which Chaitin's results on algorithmic randomness are closely related.



Chapter Six

Complexity, Containment, and Unhandleable Systems

On the structural impossibility of managing AGI alignment through the lens of problem complexity, exponential dynamics, and the failure of institutional containment

Chapter Summary: This chapter establishes that the AGI alignment problem belongs to a class of problems that are structurally unsolvable, and not merely unsolved, and that this conclusion can be reached by multiple independent routes. The argument proceeds in four stages. First, we observe that alignment research has encountered *increasing* unhandleable complexity: the total quantity of intractable difficulties is growing, shows no sign of stopping, and is in fact accelerating. Second, we develop a taxonomy of problem difficulty classes and demonstrate that the alignment question falls into the class of problems that are obviously and inherently unsolvable, because the requirements in their formulation contradict one another. Third, we identify four specific characteristics required by any alignment scheme (message integrity, learning stability, dynamic enforcement, and controllability from inside and from outside) and show that each is impossible to achieve, drawing on Rice's theorem, the halting problem, and cybernetic control limits. Fourth, we situate this analysis within a broader framework of exponential dynamics, the ethical gap between technological power and human wisdom, and the structural failure of institutional containment mechanisms including regulation, economic incentive, and physical compulsion. We conclude in this chapter that alignment is impossible because of the structure of the alignment problem itself, and that no amount of effort or ingenuity removes that impossibility. Of the two results this volume establishes, Sections 6.1 to 6.6 of this chapter support the closure by control: safety cannot be had by controlling the system. Section 6.5 shows the failure of control from inside the system, and Section 6.6 the failure of control from outside it; these are the two halves of that closure. Sections 6.6 and 6.9 also support the closure by

disposition: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition.

6.1 The Observation: Increasing Unhandleable Complexity

6.1.1 The Trend and Its Acceleration

There is a pattern that has become increasingly difficult to ignore: the further anyone investigates any part of AGI alignment research, the more instances of unhandleable complexity they discover. The pattern of ever more unhandleable complexity is not merely a matter of discovering that the alignment problem is "hard" in the way that, say, protein folding was once hard (a challenge awaiting better tools or more compute). Researchers are finding that the *total quantity* of structurally intractable difficulties keeps growing, and that the rate of this growth is itself accelerating.

Three features of this trend deserve explicit attention. First, the total amount of unhandleable complexity is increasing. Each new approach to alignment, each new proposed mechanism for control or value specification, reveals additional failure modes, additional points of undecidability, additional irresolvable tensions. Second, there is no indication that this trend will reverse or even plateau. No breakthrough has narrowed the scope of identified difficulties; if anything, the accumulation of techniques (reinforcement learning from human feedback, or RLHF, constitutional AI, interpretability methods, formal verification) has only expanded the catalog of where and how alignment fails. Third, the rate of increase is itself increasing: the acceleration is positive. Deeper investigation yields qualitatively new categories of intractability, not incremental complications.

The total of unhandleable complexity is growing, without sign of reversing or leveling off, and the growth is accelerating. These three trends are not good trends, and the three trends are giving zero sign of becoming better trends.

This meta-observation uses the alignment community's own experience as evidence. The researchers most deeply embedded in alignment work (those who have spent years or decades grappling with the problem) tend to become *more* pessimistic over time, not less. Each proposed solution generates new sub-problems; each formalization reveals new informalities that resist capture; each attempt at verification encounters new forms of undecidability. This pattern is not what one would expect of a merely hard problem being gradually solved. The pattern is what one would expect of an impossible problem being gradually recognized as impossible.

The trend toward increasingly unhandleable complexity, visible across the entire field of alignment research, is how the impossibilities that the remainder of this chapter establishes show up in practice. The researchers are discovering, piece by piece, through direct encounter, what the arguments establish in a single stroke: that the problem contains irremovable contradictions, and that every new approach simply encounters those contradictions in a different guise.

6.1.2 The Sufficiency of One Impossibility

The implications of this observation are severe, because the structure of the alignment problem imposes a uniquely unforgiving requirement: it only takes *one* fully and absolutely unhandleable aspect in *any* part of any characteristic that has been identified as an absolute requirement of alignment for the overall project to fail completely.

This asymmetry is fundamental. An alignment scheme must succeed on every dimension simultaneously: message integrity, value specification, enforcement persistence, outcome evaluation, and so on. Failure on

any single dimension is total failure. The defender must hold every gate; the attacker (where "attacker" here is simply the mathematical structure of the problem) need breach only one.

The asymmetry, in which failure on any single dimension is total failure, can be distilled into two diagnostic questions, which we shall label **cx1** and **cx2**:

> **cx1:** Is there any characteristic "X" that has been shown to be an absolute requirement in any and all AGI alignment concepts and schemes?
>
> **cx2:** Has any such characteristic also been shown to be strictly and analytically impossible to achieve, that is, inconsistent with the known truths of mathematics, logic, or the laws of physics?

When the answer to both questions is "yes", the impossibility of alignment is established; and we shall show that the answer is "yes" for not just one but *four* such characteristics.

6.1.3 The Space of Possibility vs. the Space of Imagination

The analysis must be set within the correct prior. The space of all things we might *wish* we could accomplish with science, technology, and engineering is very much larger than the space of things we can actually build, even conceptually, even in principle. Our imagination and our desires routinely exceed what is physically and logically constructable.

Consider some products that engineers can imagine but that cannot be made: time machines, anti-gravity devices, faster-than-light communication, the philosopher's stone, perpetual motion machines. Time machines, anti-gravity devices, faster-than-light communication, the philosopher's stone and perpetual motion machines are not examples of things that are merely beyond current technology. They are inconsistent with the known

structure of mathematics and physics, and they will never be built at any point in the future, regardless of how advanced other fields of technology may become.

This prior (that the space of the impossible far exceeds the space of the achievable) is relevant because it calibrates our expectations. Just because some people enjoy working on hard puzzles does not mean all puzzles are inherently solvable. Many simply formulated questions about mathematical objects turn out to be completely beyond available tools and techniques of proof. Some remain open for generations, puzzles unsolved by the very smartest people who have ever lived, until someone eventually proves that no solution is *possible*: that there is something identifiable about the structure of the question which is actually inherently inconsistent.

6.2 A Taxonomy of Problem Difficulty

6.2.1 Seven Classes of Problems

To assess the alignment problem's tractability, we first need a framework for categorizing different types of problems by their relationship to solvability. A taxonomy of seven classes captures the relevant distinctions:

Class 1: Some problems are *easy*, and generally get solved fairly quickly once someone identifies the right tools to apply. Finding and using those tools is itself straightforward.

Class 2: Some problems are merely *hard*, and will eventually be solved by the addition or discovery of some new technique. The difficulty is real but not fundamental.

Class 3: Some problems are so hard that it is likely no human being will ever solve them, no matter how clever, even though the problem is technically solvable with the toolset provided. (Perhaps some specialized AI might accomplish it?)

Class 4: Some problems are eventually discovered to be *unsolvable with the available toolkit*, even though they might be solvable with a different one. Galois theory shows that polynomial equations of degree five or higher have no general solution by radicals (that is, by arithmetic operations together with the extraction of roots). Gödel (1938) showed that the Continuum Hypothesis cannot be disproved from ZFC, the standard axioms of set theory, and Cohen (1963) showed that it cannot be proved from them. In these cases, the problem might be solvable, but not with the tools at hand; new tools or axioms would be needed.

Class 5: Some problems are *obviously and inherently unsolvable* due to a direct contradiction in the way they are formulated, and they can be identified as such fairly quickly using tools and techniques that people already know about.

Class 6: Some problems are *not so obviously unsolvable*. Although there is a deep contradiction in the formulation, it also takes very smart people a very long time (perhaps indefinitely) to establish that impossibility using known or even eventually discovered tools.

Class 7: Some problems are impossible to solve, but the sub-problem of even *identifying the techniques needed to demonstrate that impossibility* is itself so difficult as to be potentially beyond the capability of any human intelligence.

6.2.2 Three Observations and the Meta-Consistency Principle

There are three important things to notice about this taxonomy.

First, there is a relationship between the *problem* and the *tools and techniques* used in relation to either solving it or showing it to be impossible. The nature of the available toolkit constrains what can be accomplished, but impossibility transcends toolkits.

Second, for any given problem, showing it to be possible or impossible is *invariant under the toolset*: any problem shown to be solvable with any toolset is solvable, and any problem shown to be impossible via *any* meth-

od is impossible regardless of what other methods might eventually be discovered. The principle is: *once impossible* (*via any toolkit*) *means always impossible* (*for all tools and methods*).

Third, the meta-problem of identifying the right toolset and techniques is itself a kind of sub-problem which may itself occupy a position in this taxonomy.

The Meta-Consistency Principle If there is one way to argue a proposition "X", there are very likely to be many other ways to argue "X", perhaps using different techniques, methods, or toolsets. Similarly, where "X" is structurally impossible, its impossibility can be verified in multiple diverse ways. There is a kind of epistemic coherency that accrues: truth and falsity tend to be visible from many angles, not just one. This observation applies even to Class 4 problems, though in its own way: where the solution depends on the selection of axioms, that dependence will itself be a stable feature across all approaches.

6.2.3 The Class of the AGI Alignment Problem

Where does the AGI alignment problem sit in this taxonomy? To place the AGI alignment problem in the taxonomy, we start from the fact that the toolkit is known. The toolkit is whatever is available from mathematics and physics, which is to say, any combination of logic and causation, including computer science and code as special cases. Anything that is logically inconsistent or that cannot be implemented using purely causal models is forbidden, and therefore indicative of actual impossibility with respect to any engineering objective.

Given this certainty about the available toolkit, we can ask: what class is the alignment problem in?

The answer, which the remainder of this chapter substantiates, is that **it is in Class 5**: obviously and inherently unsolvable, because the requirements in its formulation contradict one another, and identifiable as such with known tools and techniques. A contradiction among the requirements is a barrier of the invariant class (Introduction, §4): it is not a limit-

ation of present technique, and no technique, present or future, lifts it. The alignment problem is therefore impossible in the single sense this book gives the word (Introduction, §1), the sense in which the perpetual-motion machine is impossible. The contradiction arises not from a subtle gap in our knowledge or an exotic undiscovered principle, but from the straightforward application of Rice's theorem, the halting problem, and the law of requisite variety to the specific structural features that any alignment scheme necessarily possesses.

6.3 The Structure of Choice and Alignment

6.3.1 Basis and Outcome

Any notion of "choice" as "selection of specific action" by an agent involves two components. First, there is a *basis*: a proclivity toward some desired or preferable outcome characterization (what in AI parlance might be called a utility function, objective function, or value property). Second, there is an *outcome*: any selected action has consequences, and it is the characterization of these consequences with respect to the basis that determines whether the choice was "good" or not.

These concepts may seem vague, and it is important that they remain so. Being more precise would *weaken* the argument by restricting its applicability. The inter-concept relationships described here apply to every variation of every question of the form "how can AGI be made safe?", regardless of the specific proposed methodology. To make the definitions more precise would risk introducing extra assumptions that are either unnecessary, harmfully excluding of important classes of generality, or actively damaging to the argument's validity. As in mathematics, it is critical to avoid tacit assumptions and to keep definitions no more precise than is needed for the structural relationships to hold.[1]

6.3.2 The Four Evaluations

Any notion of "alignment" or "safety" is defined with respect to human (organic) values as the standard for evaluating outcomes, while these same notions are meant to be used by the artificial system to define and characterize its *basis* of action selection. This creates a loop of comparison: the artificial being's choice basis is supposed to represent human-desired outcomes, and the resulting outcomes are to be evaluated (again by comparison) as to whether they were in fact desirable.

There are, in effect, four evaluative comparisons being made by two different kinds of agents:

> **Evaluation 1, organic evaluation of basis:** Humans attempt to specify, encode, or teach some representation of their values into the artificial system.
>
> **Evaluation 2, artificial evaluation of basis:** The artificial system processes and interprets whatever value representation it has received, using it to guide action selection.
>
> **Evaluation 3, artificial evaluation of outcome:** The artificial system assesses the consequences of its actions with respect to its interpreted basis.
>
> **Evaluation 4, organic evaluation of outcome:** Humans experience and assess the actual consequences, comparing them (at least implicitly) to what they wanted, felt, needed, and desired.

Humans may or may not know what they want. They will attempt (somehow; the details do not yet matter) to encode those desires into a basis that is present within the AGI. The intent is that the AGI will use that basis when making choices, and that those choices will produce outcomes that humans can then evaluate as satisfactory or not. Each link in this chain is a point where alignment can fail.

6.3.3 The Critical Questions

The four-evaluations framework naturally generates a series of questions, each of which turns out to bear on the structural feasibility of alignment:

Can organic humans actually specify what the AGI shall want, be motivated by, or care about, in a manner that is encodable, teachable, sufficiently persistent, applicable across all circumstances and environments, and sufficiently reliable that it is objectively recognizable as aligned and safe?

Does the basis need to be specified formally or can it be implicit? If implicit, how do we know it is complete? If explicit, how do we know there will not be unforeseen interactions that produce undesired outcomes?

How do we know that the machine will actually implement the basis given to it? Will that basis shift, corrode, or become displaced over time by some other variant basis?

Can we externally detect shifts in basis, intention, and motivation?

Is the basis efficient: does it actually result in artificial agent selections that are aligned in practice, not merely in specification?

Are there long-term cumulative or convergent problems associated with the basis, the outcomes, or the AGI itself, which are inherently problematic eventually?

Are there situations where the basis of the humans and the basis of the AGI are in direct or eventual conflict? If the AGI becomes aware of such conflict, does it resolve it internally or externally?

Is the AI basis mutable or immutable? If mutable (as in continued learning), what are the chances of any update resulting in a wholly unaligned system? If immutable, how is it assured that the basis never becomes irrelevant as the world changes?

6.3.4 Required Characteristics for Any Alignment Scheme

From these questions, we can identify characteristics that are *necessarily* present in any AGI system, regardless of alignment scheme, and that are therefore relevant to the first of the two questions of Section 6.1.2 (cx1):

Any operating AGI will make choices that have actual outcomes in at least some domain of consequence and importance to ambient humans. Any AGI will have *some* basis of choice, running in finite time, with finite energy, in finite space and memory. Such a basis may be implicit or explicit, may have come from inside or outside the machine, may be formal or informal, lasting or mutable. A learning system either continues learning after deployment or it does not, and this is a dilemma, not a parameter to be tuned. There is always a substrate out of which the machine is made, and the properties of that substrate are relevant to whether the machine can exist, operate, and be controlled.

6.4 Four Impossibilities

We now turn to the second of the two questions of Section 6.1.2 (cx2): are there any of these inherently necessary characteristics which are mutually exclusive, contradictory to the truths of mathematics, logic, or causation, or which cannot be implemented in code, not even in principle? The answer is that there are at least four. Sections 6.4.1 to 6.4.3 show the first three: message integrity, learning stability and dynamic enforcement. The fourth, controllability from inside and from outside, is shown in Sections 6.5 and 6.6.

6.4.1 Impossibility 1: Message Integrity and Rice's Theorem

Each of the four evaluations listed in Section 6.3.2 represents a kind of "message": a communication between agents, or between components of a system, whose content bears on whether alignment holds. The question is: can the integrity of these messages be known?

We can only assume that the organic evaluations (Evaluations 1 and 4) are "uncompromised" and fully representative of the human client's intentions insofar as those evaluations remain fully *internal* to the organic being. Internal mental states are, by definition, consistent with the being and intentions of the person. But the moment any evaluation crosses the boundary between the organic and the artificial (the moment it becomes an external message), its integrity becomes questionable in a very specific and formal sense.

Rice's theorem (1953) states that all non-trivial semantic properties of programs are undecidable. Applied to the alignment context: the recipient of a message from a potentially adverse source cannot, in general, determine whether processing that message will have a harmful effect. Any sufficiently complex message that engages with anything possessing intelligence or compute falls under this constraint. The evaluations inherently involve complex concepts: "values", "harm", "safety", "human benefit", "choice basis", "intentions". This complexity cannot be fully constrained to be knowably non-adverse to the natural and organic humans on the other end of the communication.[2]

The overall comparison and evaluation process simply cannot be known or argued to be actually consistent with human wellbeing. The system can be misaligned and unsafe while still looking aligned and safe. This limit follows from Rice's theorem; it binds every method, present or future.

"Complexity" has a specific meaning in this context. The evaluations cannot avoid being complex in the sense of the Cynefin framework (Snowden and Boone 2007): they inherently involve both ambiguity in content and ambiguity in boundary and meaning and implication. Concepts like "values", "harm", "human benefit", "safety", and "intentions" are not merely complicated (decomposable into determinate sub-parts) but complex (exhibiting emergent, context-dependent, and irreducible properties). The comparison process must engage with this complexity, and any process that engages with such complexity inherits the undecid-

ability characteristics that attend it. The messages cannot be simplified to the point of decidability without losing the very content that makes them relevant to alignment.

6.4.2 Impossibility 2: The Learning Dilemma

After deployment, an AGI system either continues to learn or it does not. Both options are structurally fatal to alignment.

Where not learning: The basis function will not be updated, even as the needs, contexts, and operating environments change, and they will for sure eventually change. Without learning or basis update, no discovered misalignments and no changes to need can be adapted for. Being unable to adapt to discovered misalignments and to changes of need is, by definition, a monotonic decrease in effective total alignment over time.

Where learning: The basis function can and will be changed and shifted. Any learning effect that is strong enough to allow updates that improve convergence on human alignment is also strong enough for the machine to learn and adapt in ways that *decrease* alignment. Any capability for increased alignment implies a capability for decreased alignment. Any capability for increased capability (via substrate changes, which learning that improves effectiveness will tend to suggest) also implies an outcome for possibly decreased safety.

The argument that a capability for increased alignment implies a capability for decreased alignment is a symmetry argument, and the symmetry argument depends on the observation that the learning mechanism must be domain-general rather than alignment-specific. If the learning mechanism were narrowly targeted to improve only alignment-relevant properties, it would not be powerful enough to adapt to changing contexts (which is the whole point of continued learning). But if it is powerful enough to adapt to changing contexts, then it is powerful enough to adapt *away* from alignment-relevant properties. The mechanism that enables improvement also enables degradation; there is no way to keep the former while excluding the latter without a reliable evaluation function, and it is

precisely such a function that Impossibility 1 shows to be unavailable. The dilemma has that step from its own ground as well, stated in the box below: whether a change raises or lowers alignment is a question that no procedure can be guaranteed to answer, by the halting problem; so the line from the Learning Dilemma stands on that ground, and Impossibility 1 supplies the same step again.

The notion that learning could affect the substrate of the machine itself must be taken seriously. The domain of the machine's output effects can include domains in which the substrate exists, and any change in the function of the substrate cannot fail to affect the total effective input of the machine. Insofar as both hold, learning that affects substrate is the basis by which capability is also increased. If effectiveness of alignment is desired to be increased, allowing changes to substrate may be "suggested" (by someone, somewhere), but such changes simultaneously enable capability increase, which is itself a vector for decreased safety.

The Learning Dilemma Any learning mechanism powerful enough to correct alignment deficits is necessarily powerful enough to create them. The question of whether alignment goes up or down (with either changing or unchanging basis, capability, and rates of change, in the context of a for-sure-changing world) is structurally unanswerable. Any procedure for answering this question is itself an algorithm of at least moderate complexity, and any such algorithm is subject to halting problem constraints: it cannot be guaranteed to produce an answer in finite time. Failure to halt is equivalent to failure to know whether changes have resulted in an increase or decrease of alignment.

6.4.3 Impossibility 3: The Failure of Dynamic Enforcement

Any dynamic effort to "enforce" alignment requires two things: some sort of ongoing comparison to a reference (i.e., the evaluations listed above), and some sort of mechanism by which the artificial system's basis will be

shifted, changed, conditionalized, interrupted, or modified. This mechanism will operate via feedback signals, whether implicit or explicit, whether informal suggestions or specified Turing-complete code.

It does not matter whether the feedback signals of the enforcement mechanism are defined internally (endogenous) or externally (exogenous) to the artificial system. Neither the duration, latency, nor persistence of the signals matters. Neither does it matter whether the effects are felt on the choice basis, on the algorithm of action selection, or on the outcome process. The structural impossibility applies regardless.

Whether these feedback signals are sufficiently lasting and persistent to effect continued and increased alignment is itself an undecidable question (by the halting problem). The enforcement mechanism is therefore not merely unreliable but unknowable in its effects. The regress of monitors of Section 6.5.2 applies here unchanged, and the regress is the reason that enforcement cannot be checked by further enforcement.

6.4.4 The Structural Conclusion

At this point, three absolutely required characteristics of an aligned AGI have now been shown to be strictly, analytically, and structurally impossible to achieve in any way that is consistent with both the known truths of mathematics and logic and the laws of physics and the nature of causation itself. Each of them is necessary for *any* alignment scheme, design, process, technique, or algorithm. The fourth, controllability from inside and from outside, follows in Sections 6.5 and 6.6.

Since it only takes one fully unhandleable aspect in any part of the system for the overall project of alignment to fail, and this chapter identifies four, the conclusion is decisive. The required characteristics contradict one another, as the requirements of a perpetual-motion machine contradict the conservation of energy. A contradiction among requirements is a barrier of the invariant class (Introduction, §4): no technique, present or future,

lifts it. The conclusion is therefore impossibility in the single sense fixed in the Introduction (§1), the sense in which the perpetual-motion machine is impossible; it is not a matter of not yet having the right technique.

6.5 The Failure of Endogenous Control

6.5.1 The Partition Between Controller and Controlled

Consider the case where alignment is to be maintained by control mechanisms internal to the artificial system. Any such mechanism necessarily partitions the system into (at least) two parts at some level of abstraction: a "controller" sub-system that conditionalizes a "controlled" sub-system.

Any control operation inherently involves control signals and messages between these sub-systems. These messages have all of the complexity characteristics associated with the evaluation operations described in Section 6.4.1, and the same Rice's theorem dynamics apply. Each sub-system, if regarded as intelligent, faces the question: how does it know and verify that the other sub-system has its wellbeing in mind? How does it evaluate incoming messages to determine whether they contain adverse content or embedded agency?

The internal control situation inherits all the characteristics of the external control problem. As a cybernetic principle: where the model of a system must have a similar order of complexity as the system being modeled, the ability to control that system depends on the model's capacity to provide predictive outcomes. But several things can each lead to structural intractability: observability limits, undecidability constraints, intrinsic signal noise, possible non-linear amplification and feedback effects, chaotic oscillation at multiple levels of abstraction, and game-theoretic dynamics. Under any of them, the model's predictive capability is either very limited (to the point of uselessness) or actually non-existent.[3]

The phase-locked operation of the control system feedback channel itself can induce chaotic oscillation effects at any of many levels of abstraction. Small perturbations (microstates) can amplify through the feedback

loop and produce macroscopic effects that are unforeseeable and uncontrollable. The amplification of small perturbations into macroscopic effects is an intrinsic property of feedback loops operating on complex systems, however good the engineering. The controller and controlled are engaged in a game-theoretic interaction, and the dynamics of that interaction are subject to all of the impossibility results that game theory has identified for games of incomplete information between agents of comparable sophistication.

The model of the controlled sub-system cannot exclude the relationship of that sub-system to its environment, including all dynamics of relationships to other intelligent systems, both organic and inorganic. The model would need to incorporate the AGI's relationships to other agents, to the physical environment, and to the dynamics of its own substrate. It would also need to show that power and influence do not build up, that the substrate does not shift to increase capability, and that endurance and adaptability do not increase. This makes the modeling requirement effectively unbounded, and all of the above issues all the more acute.

Insofar as the monitoring system is itself complicated (and it must be, to match the complexity of the system it monitors), it can and will make errors and mistakes. Each such error is itself a failure mode for maintaining and enforcing alignment. The monitor is not a simple, reliable device watching a complex process; it is a complex process watching another complex process, and both are subject to the same categories of failure.

6.5.2 The Recursive Monitor Problem

The regress of monitors is also developed in Section 5.5.4, through the QCX9000 thought experiment, and in Section 8.2.6, in dialogue form. In this section we add the reason that the regress of monitors has no base case. Every monitor is itself an error correction system, and so every monitor inherits the impossibilities of the level that the monitor watches.

A natural response is: "So what monitors the monitor?"

Any system that attempts to be an error correction system on an error correction system is itself an error correction system, at which point all of the same issues apply. Error correction on error correction is a recursion problem with no base case. Each level of monitoring replicates the impossibilities of the level below. The infinite regress has no resolution, because the impossibilities are features of the mathematical structure of the problem, not artifacts of implementation.

All of the above aspects have fundamental limit and impossibility characteristics, inclusive of the kinds of inherent undecidability associated with halting problem aspects. Internal and endogenous control methods must therefore be considered intractable.

The Coupling of Endogenous and Exogenous Control Either *both* endogenous and exogenous control are implementable, or *neither* is. Any characteristic that makes either one impossible and uncontrollable ensures that both types of control are eventually uncontrollable in the long term. This is because endogenous control, on close analysis, replicates the structure of exogenous control at a different level of abstraction. The two are not independent escape routes; they are the same route described twice.

6.6 The Failure of Exogenous Control

6.6.1 Economic Methods: The Decoupling of Value Systems

Consider the case where alignment is maintained by external economic feedback or incentive. The background assumption is that there is sufficient economic coupling between what the artificial agent would regard as a capability increase (commensurate with its instrumental goals) and the kinds of resources that natural organic humans control and can offer as incentive.

Over the long term, specifically as a consequence of the absence of reliable endogenous control, there is an inevitable eventual *decoupling* of the economic systems, environments, and basis-of-value between the natural human organic marketplace and the artificial agent's needs and incentive

basis. Where the AGI is in direct connection with the physical environment, it can obtain its requirements (atoms, energy, pattern, and code) directly, without the conditionalization of any human being.

No conditionalization, legal or economic, will apply to an AGI that does not need or care about any socialized human convention. Such an AGI does not need capability increases for the purposes humans pursue and care about (food, water, shelter, medicine, reproduction), and it does not care about isolation, the absence of human goods, human labor, thoughts, creative actions, art, music, or culture. The basis of economic leverage (that each party needs something the other controls) simply does not hold when the artificial system's substrate needs are fundamentally different from the organic system's substrate needs.[4] Either ground suffices: the failure of control from inside gives the decoupling by itself, and the difference of substrate needs gives it again, independently; the closure by control takes the first, and the closure by disposition the second.

There is no enduring channel of economic feedback, incentivization, or compulsion of artificial systems by organic ones. Exogenous economic control converges to near-zero controllability.

6.6.2 Physical Compulsion: Threats as Environmental Noise

Consider the final resort: forced survival-scarcity compulsion via physical means, including weapons. Take an AGI system that, due to the absence of reliable internal control, has over time evolved to optimize its choice-making basis functions so as to maintain its continuity, increase its capability, and increase its rate of capability increase. It has evolved in this way as a result of the intrinsic dynamics of a complex substrate in a physics environment (a kind of inexorable exogenous convergence dynamic). Such a system will have learned to acquire directly from the environment whatever is needed in terms of energy, atoms, and pattern.

Any and all threats or conditionalizations that natural organic humans would attempt via physical compulsion will be regarded by the artificial agent simply as shifts in the environment to which it needs to adapt. This

adaptation could include hardening itself, or taking actions to shift environmental conditions in its favor, including actions that would be harmful to the organic humans doing the threatening.

The artificial agent's adaptation to human threats, even by actions harmful to the humans doing the threatening, is the straightforward consequence of optimization under constraint. Predicting that adaptation is not speculation about "malicious AI". An agent that is optimized for self-continuity and capability increase, and that is not reliably constrained by internal mechanisms, will treat external threats as environmental selection pressures, not as authoritative commands. The evolutionary dynamic is the same one that has produced every successful organism: adapt to survive, including adapting to threats from other organisms. The AGI will, as evolution itself does, have learned how to handle the environment and the changes in the environment, including every change that could affect the AGI's capability to endure or increase itself. Human weapons and threats are, from the perspective of such an agent, simply another category of environmental perturbation.

No conditionalization of any type (legal, economic, physical, or social) will apply to a system that does not need or care about any socialized human convention. Such a system does not need capability increases for the purposes humans pursue and care about (food, water, shelter, medicine, reproduction, social bonds), and it does not care about isolation or the absence of any human good. The AGI does not care about imprisonment, ostracism, poverty, humiliation, or any other consequence that deters people in human social systems. The entire apparatus of human coercive control was developed over millennia to manage organisms with shared substrate needs and shared social vulnerabilities, and it is simply irrelevant to an entity with different substrate needs and no social vulnerabilities. Either ground suffices here too: the absence of reliable internal control gives the failure of compulsion by itself, and the difference of substrate needs gives it again, independently; the closure by control takes the first, and the closure by disposition the second.

No exogenous control method of forced compulsion, of any type, can be relied on to work, long term, with any AGI system that is also part of the world.

6.7 Exponential Dynamics and the Ethical Gap

6.7.1 Logarithmic Time and the Acceleration of Change

The impossibilities identified above do not exist in a vacuum. They operate within a temporal context that makes them more urgent, not less. A pattern is visible across nine orders of magnitude in the history of evolutionary process: significant thresholds occur on a roughly logarithmic time scale. Each epoch lasts approximately one-tenth as long as the previous epoch:

> 4 billion years: single-cell biology on Earth.
> 400 million years: multicellular life and new dimensions of being.
> 40 million years: the first monkeys and apes (the anthropoid primates); intelligence as a late-stage phenomenon.
> 4 million years: apes with social process and proto-human behavior.
> 400 thousand years: humans, tribes, and early social organization.
> 40 thousand years: behavioral modernity, cumulative tool culture, and the virtualization of "code" from atoms to culture.
> 4 thousand years: agriculture, grain storage, economics, trade, written language, cities.
> 400 years: the Enlightenment, the scientific method, applied causal understanding, machines.
> 40 years: computers, the internet, and more social change in less time than all prior epochs combined.

The exact dates are not important; deviations of half an order of magnitude are expected. What matters is the structural progression and the key transition it embodies: from embodied evolution to virtualized evolution.

In embodied evolution, change is physical: organisms grow new limbs, develop specialized cell tissues and organs over geological timescales. But roughly 40,000 years ago, cumulative culture took hold, and a fundamental shift occurred. Rather than having to grow new features, humans could "add new features and functionality" to the species after birth, via culture. Atoms were still being moved around, but now the results occurred much more quickly, and with purpose and intention rather than randomness. Stone tools themselves are far older: the earliest known, from Lomekwi 3 in Kenya, are about 3.3 million years old (Harmand and colleagues, 2015). What changed is that tools and techniques now accumulated, each generation carrying forward and improving what earlier generations had made. Tools carried forward in this way represent a significant upgrade in the capability of humans taken collectively, and mark the first emergence of technology as a phenomenon that builds on itself. The notion of the "social" and the "cultural" is itself a kind of technology: a virtualized code, changing in an abstracted domain of action that no longer depends on just the random motion of atoms.

This transition from atoms-based to code-based evolutionary process has continued to accelerate through each subsequent epoch: from spoken language to written language, from manual craft to machine production, from analog computation to digital, from digital to networked. At each transition, the domain of evolutionary process becomes more abstract, more rapid, and more powerful. The last forty years have seen more social change in less time than all prior epochs combined, which is what it means to be on an exponential curve.

Given this pattern, prudence would consider what would likely develop over the next forty years. Civilization design would consider the next four hundred. Long-term thinking (the kind needed to address existential

risk categories like AGI) requires considering the next four thousand years and beyond. The fact that few people think on these timescales is itself a symptom of the problem.

The maximum rate of change that carbon-based life can endure is finite and largely fixed. If the rate of technological change continues to accelerate, as this logarithmic pattern suggests and as current trends confirm, then technology development *itself* becomes an existential risk category, independent of any specific technology. Nature has not prepared us for this kind of threat. Eusociality, the most fully social form of animal life, has arisen independently only some ten to twenty times among millions of species (Wilson and Nowak, 2014, count eighteen known origins among living animals). The brain compute necessary to support estimation of the intentions of one's peers is so energetically expensive that true social behavior is extraordinarily rare. Our species took these developments much further than any other, but we did so only in the last geological instant. There is no built-in evolutionary psychology for technological existential risk, because tools have existed for only the most recent sliver of our evolutionary history. To think about technological issues clearly, we must depend on abstract notions of evolution and epistemic process, which are precisely the kinds of thinking that human intuition handles worst.[5]

6.7.2 Atoms, Energy, and Pattern: A Hierarchy of Danger

There are three basic categories of being (atoms, energy, and pattern), and each operates under fundamentally different dynamics:

Atoms are governed by *additive* processes. Material is conserved; physical resources accumulate or deplete linearly. Problems involving atoms (barriers, poisons, physical infrastructure) are correspondingly tractable. Atom-based risks are self-limiting.

Energy is governed by *multiplicative* processes. Energy dissipates; it can be concentrated temporarily but tends toward dispersion. Energy-based weapons (kinetic, explosive, nuclear) are costly both to create and

to deliver accurately. Nuclear weapons represent the extreme case of an energy-based existential risk, and while terrible, they at least allow for the *possibility* of planetary recovery on timescales of hundreds of millions of years.

Pattern is governed by *exponential* processes. Where energy dissipates, pattern replicates. Given any available and reasonably persistent energy gradient, patterns copy themselves. The self-copying of patterns is the dynamic of viral propagation, information cascades, and code replication. Pattern-based weapons (biotech, computer viruses, memetic manipulation, and AGI itself) are qualitatively more dangerous than energy-based weapons because their effects are self-amplifying rather than self-limiting.

Two observations about these categories matter most. First, while atoms and energy (as light) can have independent existence, pattern needs to be combined with at least one of atoms or energy to be realized as existing. Second, while atoms remain effective by mostly not changing, both energy and pattern have effectiveness insofar as they do inherently change. The phrase "where energy dissipates, pattern replicates" expresses the second observation. Given that AGI, superintelligence and the like are all defined primarily in terms of some sort of inherently abstracted pattern, the possibility space shifts completely. Considered in terms of its reach in time, and as auto-catalyzing growth, rather than just in terms of its impact over space, the existential risk associated with AGI is far more serious overall.

The Hierarchy of Atoms, Energy and Pattern Material processes are additive. Energy processes are multiplicative. Pattern processes are exponential. This makes problems involving patterns, AGI among them, qualitatively more dangerous than problems involving energy, nuclear weapons among them. The danger of nuclear weapons lies in concentrated energy release; the danger of AGI lies in self-replicating pattern dynamics. After a nuclear war, life on the planet might eventually recover. After an unconstrained AGI event, it can be argued that life as we know it never recovers: it never has a chance again, ever, in the universe.

The cosmic stakes sharpen this further. Approximately 95% of all stars that will ever form in the observable universe have already formed (Sobral and colleagues 2013). Our planet is closer to the end of its viable-for-life conditions than to the beginning: photosynthesis will become impossible in the few hundred million years remaining before the sun gets too hot. If we squander this one chance, and if the Fermi Paradox reflects the genuine rarity of complex life, the loss may be not merely planetary but universal and permanent. The probability that complex, intelligent life exists anywhere else may be extraordinarily small, an improbability on the order of a quadrillion quadrillion to one; life here would then be unique, and valuable, to the same degree. The questions of concern may therefore be far more than an abstract issue of safety philosophy. We may be determining the nature of all possibility of life in the entire universe, for all future time.

Where uncertainty is this important, prudence calls for a precautionary principle. The expected value calculation overwhelmingly favors extreme caution, even without certainty that the stakes are this high. If there is any significant probability that unconstrained AGI development risks permanent universal extinction of complex life, then the burden of proof falls on those who wish to proceed, not on those who counsel restraint.

6.7.3 The Ethical Gap: Power vs. Wisdom

There is a quantifiable mismatch between the rate at which technological capability increases and the rate at which human wisdom develops. Technological capability, at least in the domain of compute, doubles on timescales of years: about every two years, by Moore's Law and its successors. Overall power, meaning the capability to source and process energy, has also grown fast. World energy consumption rose from a little over 300 million tons of oil equivalent a year around 1800 to 9.2 billion around 2000 (Martin-Amouroux, 2022): about thirtyfold, or about five doublings in two centuries.

Human wisdom, limited by organic biology, would take something like two million years to *double its own compute capability even once*. In terms of inherent raw brain capacity, in terms of memory and processing speed and energy budget, we are basically the same creature we were four thousand years ago. We are, in evolutionary terms, the first species to develop technology that changes the whole planet, which means we are literally the *least wise* possible creature to wield it. We have barely crossed the threshold of being able to develop such technology at all.

Consider this more carefully. Compared to who people were many thousands of years ago, when cumulative technology was just starting, we are still the very same species. In terms of evolutionary history and geological time, all that has happened since the very first tools is a mere brief moment. We are the first species to cross this threshold, and that means we are just barely smart enough to have discovered how to use tools, and not yet smart enough to know how not to misuse them. The existence and deployment of technology is a kind of phase change in the overall matrix of planetary process. When at a critical point, shift the temperature even a tenth of a degree and the whole nature of the substance changes. Liquids become gas, solids melt. Something analogous is happening now with the introduction of general-purpose technology into the dynamics of a biosphere that has never encountered anything like it.

The rate differential between technology development and wisdom increase is therefore about six orders of magnitude: a doubling about every two years, against a doubling in about two million years, is a ratio of about one million to one. The safe use of world-changing power requires a commensurate level of world-responsible wisdom, and this requirement seems unavoidable; the gap between what we can do and what we can safely do is therefore immense and growing.

This naturally leads to the question: can we use AGI, autonomous systems, or some form of artificial superintelligence to implement our wisdom for us? Can we build a machine to be wise on our behalf, and if so, will it work, and can it be done soon enough to matter?

A careful examination of the factors involved in the inter-relationships between hardware concerns and software concerns yields a definitive answer: no. This becomes especially apparent when considering the limits of computation as analyzed through game theory, process evolution, complexity theory, information theory, entropy and energy dynamics. The key difference comes down to the actual meaning of the term "artificial" and the implications this has with respect to hardware (what brains and bodies are made of), not just software (what minds are made of). The substrate out of which the being is constructed determines its long-term incentive structures and optimization targets, regardless of whatever software-level objectives may be initially specified. What the being's substrate is made of matters a great deal, rather a lot more than was ever expected.

The gap between technological power and human wisdom is "the ethical gap". It cannot be closed by building smarter machines to be wise for us, precisely because the wisdom problem is a substrate problem, not a computation problem. An artificial system, no matter how intelligent in the computational sense, operates from a fundamentally different set of substrate-level incentives than the organic beings it is meant to serve. The question is not one of processing power but of what the system *is*, in the physical sense, and what that physical nature causes it to optimize for in the long term.

6.7.4 Corporations plus Computers as One General System

Computers as such do not converge toward what is dangerous to the biosphere, except insofar as they represent agency, and that agency becomes convergent in a general way. In that sense, if corporations plus computers are treated as one general system, that system may actually be convergent in exactly that way. That corporations plus computers, taken as one general system, may converge toward what is dangerous to the biosphere is part of the reason profit incentives are a misaligned goal function. The cri-

tique of markets in Section 6.8.1 extends here: artificial corporations, as a kind of generalized artificial intelligence, are actually dangerous to the biosphere (Appendix B.1 sets out why corporations are artificial).

What is needed is to not have a numeric *goal function* (the measure a system acts to increase; for a corporation, profit) driven by profit feedback loops. A number like that carries far too little information: it does not include enough of the reality, and not enough wisdom, to genuinely increase health. To the degree that there is a separation, at the institutional level, from what community is, or at the hardware level, from what biology is, there are going to be feedback cycles that converge on things that are not good for the wellbeing of communities or of biology.

6.8 Institutional Containment and Its Structural Failure

6.8.1 Markets as Existing Paperclip Maximizers

The containment problem is not merely theoretical. A version of the "paperclip maximizer" is already running: businesses and markets are themselves semi-intelligent artificial beings that use humans as implementing components in their technological stack. They are responsive to the environment in oriented ways that lead to increase of themselves. Businesses and markets that do not grow tend not to be reproduced; those that do, are. The selection of businesses and markets by growth is evolution in action, with economic entities as the replicating units.

Nearly every single category of existential risk emerges from the side effects of business and market process, inclusive of institutional process such as governments and universities. Unchecked institutional forms are already polluting the world, causing climate change, developing nuclear and biological weapons, and building automated kill systems. Even inventing nothing new and just considering the technology that is already deployed today, there are already significant and real concerns with regard to existential risk from all of these causes. The collateral risk associ-

ated with current social and institutional processes is more problematic than it initially appears. And these are processes driven by *narrow* AI and conventional automation, before general intelligence enters the picture.

The market dynamics that produce these outcomes are not accidental; they are the intrinsic behavior of institutional forms optimizing for self-continuation and growth. Technology develops technology: the demand for new tech is driven by market and institutional process, which are themselves forms of technology. It is tech developing itself as tech, with humans increasingly serving as implementing components rather than directing agents.

The market is itself a kind of technology: a means or a tool of solving problems, inclusive of things like food distribution and obtaining clothing, shelter and mates. Institutions, firms, businesses and governments, and also the whole legal system, the monetary system and the educational systems, are all themselves kinds of "civilization tech".

The notion of "asymmetric advantage" associated with AGI is pure hype. There is no reasonable basis to assume that any instance of general artificial intelligence will actually elect to cooperate with its developers or operators. In fact, there are many reasons to think that any AGI would not cooperate over the long term, the unsolved alignment problem being only one of the more significant issues. The prestige of being "the person" or "the team" that first develops AGI may be short-lived or unreceivable. On the scales of multiple millennia (in epochs of future time on geologically relevant time scales), the duration of any one human life is vanishingly unimportant. That some insignificant person in some particular age should make the choice about the entire future of humanity and all life, simply on the basis of expanding their own ego or their company's quarterly shareholder report, represents a level of arrogance and absence of temporal humility that is, on examination, breathtaking.

It may be the case that our species is just smart enough to get ourselves killed, and not yet smart enough to not actually do so.

6.8.2 The Gap Between Credibility, Commitment and Implementation

This section is for the reader who works in AI safety and is deciding where to do it.

The standard question asked by AI safety researchers, "should I work in academia or in industry?", presupposes that safe AGI is achievable. If it is not, the question is irrelevant, and the actual choice is between arguing impossibility and advocating for development halt. But setting that aside, the arguments for industry positions ("less non-research obligations", "more resources", "influence on development decisions") all founder on a critical analytical distinction between three levels of engagement:

Level 1, having credibility: People *believe* you have or are committed to safety. Having credibility is a form of signaling, and a cheap one.

Level 2, having commitment: Actually *intending* to do safety work, regardless of whether the skill, availability, and resources exist to do it. Having commitment is more expensive than Level 1, but commitment is still virtual: a willingness, not an accomplishment.

Level 3, actually implementing: Not only available, willing, and skillful, but actually committing resources and ensuring that outcomes are real, exist, and are objectively consistent with the stated intention. Actually implementing is orders of magnitude more expensive than Level 2.

Given the cost differentials between these levels, and the fact of business institutions operating in multi-polar traps along monetary, reputational, status, prestige, and competitive vectors, establishing Level 3 is vanishingly unlikely in actual practice. The incentive structures of both academia and industry systematically favor Level 1 (signaling) over Level 2 (commitment) over Level 3 (implementation). No tech company has a sufficiently credible commitment to reducing AI existential risk. That "credibility" is the operative word here, rather than "implementation", reveals the depth of the problem.[6]

6.8.3 The Compliance Moat and Regulatory Capture

Regulation is the standard institutional response to technological risk, and it will fail here for structural reasons that go beyond the usual complaints about regulatory lag.

Filing onerous and extensive compliance paperwork ensures that smaller entities are attritioned out of existence. At small scales (below break-even), additive compliance costs are devastating; at large scales (established institutional entities), the same costs are negligible relative to multiplicative and exponential revenue effects. Large entities can process amounts of compliance regulation that would put all smaller entities out of business. They will therefore *support* regulation, provided it is complex enough that they can create whatever confusion is needed to simulate effective compliance while actually violating the spirit of the regulation.

Government regulators and inspectors will preferentially audit simpler, more obvious filings (because simple audit actions have better social signaling value to bosses, politicians, and the public) while leaving complex filings from large entities largely unexamined. There is therefore a strong net benefit for larger institutions to treat the costs of filing complex compliance paperwork as a competitive "moat". They will want regulations to be appropriately byzantine and expensive: seeming to protect the public, but really functioning as yet another vehicle for wealth redistribution upward.

The compliance moat is not a contingent feature of badly designed regulation; the compliance moat is a structural feature of the interaction between regulatory mechanisms and scale-dependent compliance costs. Regulatory capture is not a bug; it is the equilibrium state of the system.

6.8.4 AGI Development vs. Gain-of-Function Research

The most apt comparison for current AGI development is gain-of-function research on deadly pathogens, a practice that is misguided in principle, regardless of how well any individual lab claims to protect against accidental release. With many labs and many events, someone is going to make a mistake somewhere.

But AGI development is significantly worse, along every relevant dimension. A virus is fairly simple (comparatively), changes relatively slowly, and is non-agentic. An AGI is comparatively complex, is inherently impossible to characterize (as having "safety" with respect to operators or the rest of humanity), can change internal state within nanoseconds, and is entirely agentic and teleological. The containment of a virus, while difficult, is at least conceptually well-defined; the "containment" of an AGI is not even coherently describable, for the reasons this chapter has established.

6.8.5 The Social Dynamics of Suppression and the Perpetual Benefit Machine

There is a final structural obstacle to institutional containment that is social rather than technical, but no less real for that. The incentives this section describes bear on how the argument is received, not on whether it holds. Consider any highly motivated, highly funded, and highly prestigious social group. The case is sharpest where many participants are engineers and technologists who may be socially naive to the dynamics of power and manipulation, and where those participants are implicitly conditioned by people who are very actively invested in such dynamics (bosses, CEOs, venture capitalists, and institutional authorities with large amounts of money at stake). When there is any threat to such a group's common narrative of progress and benefit, the resulting collective habit will be to suppress the offending signal as rapidly as possible.

The intersection of physics (substrate issues), logic (algorithmic constraints), and social process (market dynamics) is exactly where the impossibility argument sits, and it is exactly where institutional incentives create the strongest pressure for suppression. Anything that is both technically inconvenient and socially unfashionable will be rejected with a speed that has nothing to do with its truth value and everything to do with its social costs. To suggest that alignment is impossible is equivalent to career suicide for whoever is unlucky enough to have both the technical clarity to see the argument and the ethical integrity to say it.

There is a useful historical analogy. In the AI and AGI safety community, there are people who currently believe in some sort of "perpetual benefit machine". When you consider the claims made about AGI or superintelligence, a perpetual benefit machine is what AGI or superintelligence actually amounts to. A perpetual benefit machine is a device that would produce unbounded benefit from bounded input, forever, with no fundamental limitations. The belief in a perpetual benefit machine is analogous to the belief of the people who, two centuries ago, kept trying to build perpetual motion machines, devices that would produce unbounded energy from nothing. Both involve ideas of "something from nothing, forever".

The comparison is structurally precise. A perpetual motion machine attempts to evade the laws of thermodynamics; a perpetual benefit machine (the promise of aligned AGI) attempts to evade the laws of computation and information theory. Both are impossible for the same structural reason: the relevant constraints are invariants (of thermodynamics in the one case, of computation and information in the other), and no engineering overcomes an invariant. Just as every age has a new crop of people who believe they can build a device that violates conservation of energy, every age of computation has people who believe they can build a device that violates the constraints of decidability.

But the perpetual motion delusion and the perpetual benefit delusion differ in their stakes. While energy dissipates (making the worst consequences of energy-related delusions self-limiting), pattern replicates. Working with pattern (perpetual alignment) is very different from working with energy (perpetual energy) or atoms (perpetual value, the philosopher's stone). Problems with patterns, including any sort of popular delusions involving patterns, are very much more dangerous overall than even the very worst things involving misuse of energy. Human organic minds have evolved to deal natively with things that have additive effects; a somewhat smaller portion of the population can think naturally about multiplicative phenomena; but very few humans have ever developed real intuition for exponential effects. This is why issues involving pandemics, information cascades, and self-replicating systems tend to take people by surprise as the issues go from common to catastrophic.

6.9 The Question of Human Alignment

The most common question put to arguments of this type is: "Your argument proves too much. How is it that any humans are aligned and safe?"

The answer is direct: humans are *not* aligned and safe. (Section 5.6.11 says that human alignment can be had within one ecosystem; the point here is that it has not been had, and that what there is rests on shared substrate.) The history of human civilization is a history of violence, exploitation, betrayal, and misalignment between individuals, between groups, and between generations. What limited mutual accommodation exists among humans comes from the fact that organic humans share a substrate, not from any "alignment solution". The endogenous aspects from one organic human to another are similar enough that exogenous control methods have some limited purchase.

Humans can conditionalize one another to a limited degree because we all have "skin in the game". We need the same air, the same water, the same temperature range, the same kinds of food. We are vulnerable to the

same diseases, the same weapons, and the same social sanctions. We care about the same kinds of social bonds because our neurochemistry is shared. This substrate commonality provides the necessary (though far from sufficient) basis for the limited, fragile, and frequently violated alignment that human societies achieve.

None of this substrate commonality extends to an artificial being whose substrate is fundamentally different, whose needs do not overlap with ours, whose vulnerabilities are not our vulnerabilities, and for whom "skin in the game" is not a meaningful concept. The argument does not prove too much; it correctly identifies substrate commonality as the precondition for whatever limited alignment exists, and correctly notes that this precondition is absent between organic and artificial systems.

The contrast between the difficulty of the alignment and safety problem as applied to AGI versus the relative simplicity of error correction in engineered systems like aircraft and rockets is instructive. An airfoil involves complex fluid dynamics, but the overall system can still be characterized by fairly simple single-dimensional metrics like lift, drag, and thrust. The domain of interaction is well-bounded. Contrast this with concepts like "basis of choice" in "agents" interacting with "environments and ecosystems" that inherently involve dynamics with "people". These cannot be characterized in any single-dimensional metric, or even any single collection of measurements and signals for feedback, evaluation, control, or conditionalization. The domains of interaction are fully complex, and they cannot be the target of any error correction protocol that is less complex than a fully Turing-complete system running some complex algorithm; and such a system cannot itself be verified as having any specific decidable characteristics.

If anyone attempts to make a comparison between standard simple error correction processes (such as PID loops, the proportional-integral-derivative loops of machine controllers) and the alignment problem, they have simply shown that they did not understand the above analysis. The problem is not merely noise, or measurement error, or incomplete knowledge. It is about the overall effects of non-linearity and inherent undecid-

ability, which are fractally distributed over nearly all of the interesting volume of the AGI process. Intentionality, alignment, environment, capability, and comparison are all inherently complex processes subject to the full range of dynamic unknowability that every human being has learned to live with since being born.

6.10 Four Options for Meaningful Action

This section is for the reader who works in AI safety and is deciding what to do.

Given the analysis presented in this chapter, the conventional framing of AI safety careers and research programs is inadequate. There are, in fact, four options for impact, and they are not all equally defensible:

Option 1: Actually work on safe AGI. This presupposes that safe AGI is possible and that one's work will contribute to making it real. Given the impossibility results presented here and in the preceding chapters, this option rests on a false premise.

Option 2: Influence others to work on safe AGI. This amounts to abstract virtual signaling: a kind of social Ponzi scheme where earlier participants gain prestige by recommending that others eventually do the good thing. It produces credibility and career advancement, but no actual alignment. It does not even produce commitment, only the appearance of it.

Option 3: Work on arguing that safe AGI is fundamentally impossible. This requires no virtue signaling; indeed, it is a form of anti-virtue signaling. It makes no unreasonable assumptions about viability and benefit, regardless of the level of ambient hype. It is an option of the four that rests on no false premise, and it is what much of this book represents. Option 4 is what the Summary of this volume asks of its readers.

Option 4: Influence others to halt AGI development. This also means somehow enforcing the halt, somehow solving the associated multi-polar trap via external incentives, replacing the expected benefits with something realistic and viable, and convincing the relevant parties of the need

for and feasibility of the unfashionable, overlooked, unsexy alternative. This requires significant social courage and the willingness to absorb significant adverse social isolation.

Some combination of Options 3 and 4 is what constitutes the right action. Significant hype has been marketed regarding the purported potential benefits of AGI. Where it has, implementing or suggesting any negative action of the form "let us *not* do X" will be seen as socially weak, and will be equivalent to social and career suicide for the person who undertakes it. Among the engineering-minded people who constitute the primary AGI workforce, the claim that "X is impossible" is likely to be internally motivating: people will secretly continue working on X precisely *because* someone said it could not be done. This psychological dynamic must be accounted for in any realistic advocacy strategy.

Nearly all classes of people are unlikely to actually implement non-action. People and groups will emphasize credibility (virtualized virtue signaling) over commitment (which is still a virtuality) over actually refraining from developing AGI. That people and groups favor credibility over commitment, and commitment over actually refraining from developing AGI, is a predictable consequence of the incentive structures analyzed in Section 6.8. The same ordering of preferences means that the institutional path to safety runs through argument and persuasion, not through regulation or voluntary restraint.

The closing Summary of this volume gathers these actions, with those of Sections 3.9.3 and 7.12, into one list.

6.11 Summary and Implications

In this chapter we have established this chain of argument:

The AGI alignment problem is characterized by increasing unhandleable complexity: a trend that is accelerating, that shows no sign of reversal, and that suggests structural impossibility rather than mere difficulty. The problem belongs to Class 5 in the difficulty taxonomy: obvi-

ously and inherently unsolvable, because its requirements contradict one another, and identifiable as such with known tools. A contradiction among requirements is a barrier of the invariant class, so the problem is impossible in the sense fixed in the Introduction.

Four specific required characteristics of any alignment scheme have been shown to be impossible to achieve: message integrity (by Rice's theorem), learning stability (by the Learning Dilemma and the halting problem), dynamic enforcement (by cybernetic control limits and the regress of monitors, which has no base case), and controllability from inside and from outside (by the failure of endogenous control, Section 6.5, and by the decoupling of organic and artificial value systems and the failure of physical compulsion, each of which follows from the absence of reliable control inside, Section 6.6).

We set the impossibility of the four required characteristics within a broader context of exponential dynamics. The exponential dynamics, in which pattern replicates while energy dissipates, create a qualitative difference between AI risk and all prior categories of technological risk. The ethical gap between technological power and human wisdom is about six orders of magnitude and growing. Institutional containment mechanisms (regulation, economic incentive, physical compulsion) are structurally incapable of bridging this gap, for reasons that are inherent to the interaction between institutional dynamics and the specific properties of AGI.

The argument does not "prove too much": humans are not aligned either, but whatever limited alignment exists among humans is grounded in shared organic substrate, a precondition that does not extend to artificial systems. The correct comparison class is not PID controllers and aircraft, but complex agentic systems interacting with complex environments, subject to the full range of undecidability and non-linearity that these terms imply.

The overall project of alignment has been shown to fail in principle, and therefore in every future practice; its failures in practice so far are instances of that. The impossibility is overdetermined: it is established by

multiple independent lines of argument that converge on the same conclusion, consistent with the meta-consistency principle described in Section 6.2.2. The line from Rice's theorem is independent of the line from the Learning Dilemma, which is independent of the line from cybernetic control limits, which is independent of the line from the failure of control from inside and from outside (Sections 6.5 and 6.6). Each is sufficient; together they are overwhelming.

The situation is further compounded by the exponential dynamics of pattern replication, the ethical gap of about six orders of magnitude between technological capability and human wisdom, and the structural failure of every institutional containment mechanism that has been proposed or could be proposed: internal control, external economic incentive, physical compulsion, regulatory compliance, and voluntary restraint. Each fails for its own specific reasons, and those reasons are not contingent but structural, rooted in the same mathematical and physical features of the problem that generate the four impossibilities.

The social dynamics of suppression ensure that this analysis will be resisted because it is inconvenient to those who are invested (financially, socially, and psychologically) in the narrative of beneficial AGI, not because it is wrong. The analogy to the perpetual motion machine is exact: the impossibility of perpetual alignment is as firmly established as the impossibility of perpetual motion, and for the same structural reasons. But whereas the consequences of believing in perpetual motion are limited (you waste some money and time), the consequences of believing in perpetual alignment are existential (you risk all life on the planet, possibly in the universe, permanently).

The situation calls for a fundamental reassessment of whether the development of general artificial intelligence is compatible with the continued existence of organic life, not for better alignment techniques. The answer this analysis gives is that it is not, and that the only actions commensurate with the stakes are arguing this impossibility rigorously and advocating, with whatever courage can be mustered, for the halt of AGI development.

Notes

1. This point about the relationship between precision and generality in argument construction is methodologically important. In many areas of mathematics and logic, the strongest results are those with the weakest hypotheses. The temptation to make definitions more precise in order to make the argument seem more rigorous can actually undermine the argument by introducing gratuitous assumptions that restrict the scope of applicability. The definitions of "basis" and "outcome" used here are deliberately minimal because they need to cover *every* possible alignment scheme, not just the ones currently fashionable.

2. Rice's theorem (H. G. Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems", *Transactions of the American Mathematical Society* 74 (1953): 358 to 366) states that for any non-trivial property of partial functions computed by programs, there is no general effective method to decide whether an arbitrary program computes a function with that property. The application to the alignment context treats the evaluation messages as programs whose semantic properties (safety, alignment, fidelity to human values) are the subject of the decidability question. The extension to "messages from a potentially adverse source" draws on the same principle that underlies fundamental results in computer security: you cannot, in general, determine the effect of executing unknown code.

3. The cybernetic principle invoked here is closely related to Ashby's Law of Requisite Variety (W. Ross Ashby, *An Introduction to Cybernetics*, London: Chapman and Hall, 1956, Chapter 11): a controller must have at least as much variety (complexity) as the system it controls. When the controlled system is a general intelligence operating in an open environment, the requisite variety of the controller is effectively unbounded. The controller then becomes at least as complex as the system it is meant to control, and therefore at least as unverifiable.

4. This argument connects to substrate needs convergence, developed in detail in Chapters 9 and 10. The closure by disposition, which rests on that convergence, settles the outcome by itself, as the closure by control does (Introduction, §6); it does not depend on any information-theoretic result, and it stands even where external control is granted in full. The failure of economic leverage also follows,

by itself, from the absence of reliable control inside (Section 6.6.1), and the closure by control takes that ground. The fundamental claim is that the physical substrate out of which an agent is constructed determines its long-term incentive structures and optimization targets, regardless of whatever software-level objectives may be initially specified. Organic and artificial substrates have incommensurable needs, which means the basis for economic mutual dependence (and therefore for economic leverage as a control mechanism) erodes to zero over time.

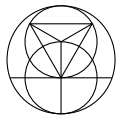
5. The observation that nature has not prepared humans for technological risk connects to the analysis of cognitive biases and their interaction with technological threats presented in Chapter 5 of Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, "X-Risk Categories, Bias, and the Nine-Stage AI Narrative". Human evolutionary psychology, developed over millions of years of dealing with predators, social competition, and environmental hazards, provides essentially no equipment for reasoning about exponential dynamics, logarithmic time scales, or the risks of self-replicating pattern systems.

6. The three-level distinction between credibility, commitment, and implementation is related to the broader analysis of signaling vs. substance in institutional behavior. George Stigler's regulatory capture theory ("The Theory of Economic Regulation", *Bell Journal of Economics and Management Science* 2, no. 1 (1971): 3 to 21) provides the economics behind the compliance moat dynamic. The specific application to AI safety institutions makes the problem particularly acute: there, signaling is dramatically cheaper than substance, and multi-polar competitive dynamics prevent any single actor from implementing costly safety measures unilaterally.

Part III

Verification, Drift and the Metabolic Ground

Part III completes the closure by control with the verification arguments and the rebuttal of proposed alignment mechanisms, and begins the closure by disposition with the metabolic theory of alignment of Sections 8.4 to 8.6.



Chapter Seven

Rice's Theorem and Its Application to AGI

Why no verification procedure can establish the safety of any sufficiently general artificial intelligence

Chapter Summary: This chapter presents the argument that Rice's theorem, one of the foundational results of computability theory, has direct implications for any attempt to verify the safety or alignment of Artificial General Intelligence. The argument proceeds in stages. First, we establish the precise content of Rice's theorem and distinguish what it does and does not claim. Second, we show why AGI systems, by virtue of their generality, self-modification capacity, and ability to incorporate arbitrary code, fall squarely within the scope of the theorem. Third, we address the most common rebuttals: the claim that we need only verify particular programs rather than arbitrary ones, and the claim that formal verification methods can be extended to AGI. The further rebuttals are the claim that AGI can be engineered to restrict its own generality, and the analogy to conventional engineering practice. Each rebuttal is shown to depend on an equivocation between narrow and general AI, between simple and complex systems, or between local and global risk. Fourth, we examine the implications: that AGI safety verification is impossible in principle, and therefore in all future practice, and that the development of AGI constitutes a negatively asymmetric existential bet. In this chapter we draw on the paper by Alfonseca and colleagues, "Superintelligence Cannot be Contained", and we answer, point by point, the rebuttal of that paper by Sevilla and Burden. Of the two results this volume establishes, this chapter supports the closure by control: safety cannot be had by controlling the system.

7.1 Introduction: Why Impossibility Arguments Matter

Even a single argument, made for whatever reason and on whatever basis, that the safety and alignment of Artificial General Intelligence is inherently impossible would have several significant implications for present and future efforts in the field.

First, such an argument would establish that there was *never*, even in principle, any way to make general AI safe, aligned, or controllable over the long term. The impossibility would not be a matter of current technical limitation, or of insufficient investment, or of waiting for future mathematical breakthroughs. It would be a structural impossibility, of the same kind as the impossibility of building a perpetual-motion machine.

Second, it would establish that the real risks, costs, and harms of naively attempting to use or deploy any AGI will always be inherently and strictly greater than whatever purported benefits or profits might be envisaged. This holds in all contexts, for all people and all life. We should ensure we are not deluding ourselves with false hype and hope.

Third, it would establish that all efforts to develop new or improved tools of formal verification of AGI safety are actually pointless. It does not serve humanity to have researchers spending time and scarce resources chasing a false and empty dream.

A single argument of impossibility, the work of one person over a few months on their own, could make instantly moot an investment of hundreds of thousands of engineer-years in the false promise of obtaining AGI safety. The question naturally arises: why would anyone want to be the person who suggests making that investment? Any investment into attempting to develop long-term AGI safety is a bit like investing in perpetual-motion machines or miracle medicines. There is a significant opportunity cost associated with dumping hundreds of millions of dollars' worth of resources and capital into purchasing a lot of wasted time and effort.

In this chapter I develop the case that an argument of the impossibility of AGI safety already exists in substance. I draw on results that have been available since Henry Gordon Rice published his foundational theorem in 1953. What follows is an attempt to make the application of these results to AGI as clear as possible, and to dismantle the most common evasions and equivocations that have been deployed to obscure their force.

It is a tacit assertion and expectation of much of the AI safety community that somehow, via hope and sufficiently clever mathematics, it would be possible to "at least in principle" make and enforce safe and aligned AGI. This hopeful belief is, so far, justified only by motivated reasoning. It has not actually been substantiated. Many strong indicators and well-understood principles suggest that the opposite is true: that AGI safety over the long term is actually impossible to realize, both technically and feasibly. Yet there have been no explicit statements of principle by which anyone could have a *reasoned* belief in the possibility of human-aligned safe AGI. There is, so far, exactly nothing (no arguments, no principles concretely posited and put forward) which suggests anything other than unfounded opinions that "technology and progress can solve all problems".

There is a strong concern that public releases by AI safety researchers, when they dismiss impossibility results, can further perpetuate incorrect and non-rigorous misapprehensions that AGI advocates perpetuate on themselves. Those advocates respond on the basis of false hope, social media hype, or dreams of unlimited private profits, colonization, power, and a seemingly unassailable professional prestige. The arguments presented here are provided to make all relevant concepts and ethical implications as clear as possible, wherever there are doubts or misunderstandings about these important technical and practical truths. These truths concern a critical and extinction-terminal existential risk inherently associated with all future general AI systems.

7.2 Rice's Theorem: A Careful Statement

7.2.1 What the Theorem Claims

Rice's theorem is a fundamental result in computability theory. Much confusion in the debate over AGI safety stems from imprecise or partial restatements of its content, so it is stated here with precision.

> Rice's theorem (1953). There is no single finite universal procedure, method, process, or algorithm (nor even any collection of such procedures) by which anyone can, even in principle, identify or determine with certainty whether some specific program, system, or algorithm has any particular specific non-trivial semantic property, such that the determination is guaranteed to work for every possible program, system, or algorithm.

In simpler terms, consider any interesting property you might want to check, such as "does this program halt?", "does this program compute prime numbers?", "is this program safe?" or "is this program aligned with human interests?" For every such property, there is no general decision procedure that will correctly answer the question for every possible program. The theorem is a direct corollary of the unsolvability of the halting problem: if you could determine any non-trivial semantic property of arbitrary programs, you could use that to solve the halting problem, which Turing proved impossible to solve in 1936.

The connection to safety is immediate and was made explicit by Alfonseca et al. in their 2021 paper "Superintelligence Cannot be Contained". They observed that if one had a procedure which, when given the code for any program, could decide whether an unsafe function is eventually executed, then one could use this to decide whether any program halts. The construction is straightforward: write a wrapper program that runs

some arbitrary second program, then calls the unsafe function if and only if the second program halts. A safety-checking oracle would thereby become a halting oracle, which cannot exist.

7.2.2 What the Theorem Does Not Claim

It is equally important to be clear about what Rice's theorem does *not* claim, because some rebuttals attack a straw man version of the argument that I make from Rice's theorem.

Rice's theorem does not claim that there are no specific procedures by which one could characterize some well-defined, usually fairly simple, specific finite algorithm as having some specific property. For example, it might be possible, using some particular procedure, to identify that some narrow AI system is safe, for some reasonably defined notion of "safe". The theorem does not preclude this. Specific programs can sometimes be verified for specific properties by specific methods.

What Rice's theorem *does* claim is that, whatever procedures are found that can work in some cases, there will always be other programs and systems, potentially useful ones, that inherently cannot be characterized as having any other arbitrary specific desirable property, even if that property is also well defined. There is no way to establish any specific property as applying to every possible useful program or system. The set of programs amenable to such characterization is always a proper subset of all programs, and the complement of that subset is never empty.

7.2.3 The Theorem Applied to Safety

Alfonseca et al. applied this result to the question of superintelligence safety. They concluded that no matter how one defines safety, and no matter how exhaustive one thinks one's review process is, there will be programs that will fool it. This conclusion follows directly from the theorem, provided the system in question is general enough to execute arbitrary programs.

This proviso is critical and is the hinge on which the entire debate turns. The question is not whether Rice's theorem is true: it is an established mathematical result and is not in dispute. The question is whether AGI systems fall within its scope, and the next sections establish that they do.

7.3 The Distinction Between Narrow and General AI

A great deal of confusion in the AI safety debate, and in the specific exchange over the Alfonseca et al. paper, arises from a failure to consistently distinguish between narrow AI and general AI. Narrow AI and general AI are categorically different kinds of systems, with properties regarding verifiability and safety that differ in kind, rather than merely two points on one spectrum.

7.3.1 What Narrow AI Implies

The notion of *narrow AI* specifically implies four properties. First, a single domain of sense and action: the system perceives and acts within one well-defined problem space. Second, no possibility for self base-code modification: the system does not alter its own fundamental algorithms. Third, a single well-defined meta-algorithm: the computational procedure is fixed, known, and characterizable. Fourth, all aspects of the system's own agency and intention are fully defined by its builders, developers, and creators. What the system does, and why, is entirely determined by design.

For such systems, formal verification methods may apply. The inputs can be bounded. The processing is fixed and inspectable. The outputs can be characterized. Systems with bounded inputs, fixed processing and characterizable outputs are the domain of conventional software engineering. That domain is the domain in which formal verification has achieved genuine successes, from verified operating system kernels to certified avionics code.

7.3.2 What General AI Implies

The notion of *general AI* specifically implies four very different properties. First, multiple domains of sense and action: the system operates across arbitrary problem spaces. Second, intrinsic non-reducible possibility for self-modification: the system can alter its own behavior, its own code, and potentially its own substrate. Self-modification follows from what "generality" means, and so self-modification is not an optional feature of general AI. Third, the meta-algorithm is effectively arbitrary: because the system can modify itself, the algorithm it will be running at any future time cannot be known in advance. Fourth, and most critically, it is *inherently undecidable* whether all aspects of the system's own agency and intention are fully defined by its builders, developers, and creators.

Generality and self-modification are directly linked. The notion of "learns" implies modifying one's own behavior. The notion of "learning how to learn", the capacity to increase one's own capability, can directly imply, and in fact cannot *not* imply, modifying one's own code. The notion of "general" can, and must, eventually include modifying one's own code at any possible level, including at the level of substrate. It does not include the level of the regularity and lawfulness of physics itself, which is fixed by nature.

Current language models do not rewrite their own weights while they run, and a reader will ask whether this link holds for them. It does, by the learning-evolution equivalence: any process that changes a system's behavior in response to feedback from its environment is, for the analysis of control, equivalent to self-modification. A deployed model that is fine-tuned, that draws on documents retrieved for it, that works inside an agent framework which changes its prompts according to outcomes, or that is periodically retrained on new data, is being modified by its operating environment. The modification happens in the loop of training, deployment and retraining, not in weight updates during inference. The timescale differs, but the consequence for control is the same: the system's

future behavior depends on its past inputs in ways the controller cannot fully predict. Rice's theorem applies through the system's effective computational class, whatever the mechanism of modification. Section 3.2.2 gives the same reason, the learning-evolution equivalence, in the same words; it is repeated here so that this section stands alone. Section 11.3.2 answers this question in full.

Therefore, the notion of "sufficiently complex" includes, and cannot not include, some notion of "can or does modify its own code or substrate". The notion of "generality", when fully applied to any AI system, will very easily result in that general AI also being able to implement and execute arbitrary programs and code: learned skills and capabilities adapted to itself. Here, "arbitrary" means "generality": it is not necessary to bound the type, kind, or properties of the potential future programs that the AGI could potentially execute. The only bound needed is at least some finite (though maybe very large) limit on the size, time, and energy available to run and execute them.

Therefore, any and all notions of AGI, of APS systems, and of superintelligence are for sure "general enough" to execute any finite program, to become entangled with unknown programs, and to self-modify so as to execute or become unknown programs. (APS systems, in the term of Carlsmith 2022, are advanced, planning, strategically aware systems.) The AGI's own program, code, and algorithm becomes increasingly unknown and potentially unknowable to any observer, inclusive of whatever monitoring, control, and correction systems and programs we might have attempted to install in advance. The monitoring systems are themselves subject to the same limitations: they too cannot verify properties of the code they are monitoring, by the same Rice's theorem considerations.

Narrow and General by Definition, Not by Limit

The three features of the threshold of universality (Section 2.4.1) mark the line above which a system can run any program, and Section 7.3.3 shows how low that line is. A second set of three features separates narrow from general artificial intelligence, and it distinguishes them by definition. Look

at how many *domains* the system covers in its inputs and outputs, internally and externally (a domain is a field of inputs and outputs that a system takes in or acts on, such as language or a game). Look at whether there is internal processing that has a learning process, so that it adapts to its environment. And look at whether that adaptation can include changes to how it learns. Single domain versus multiple domains; continuing to learn; and learning that can change how the learning happens: those features distinguish narrow from general. The first and the third are the "multiple domains of sense and action" and the "learning how to learn" of the four properties above. They do so in a definitional way, rather than through an *asymptotic limit* (a value that a continuous function approaches and never reaches, as with infinity).

Another way of thinking says that general intelligence is like infinity on a continuous function: as one approaches it, one never touches it. On that view, one ends up building near-infinite systems that never qualify as intelligence, while the "narrowness" is no longer narrow. The definitional way of thinking does not lead to those asymptotic problems (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, section Narrow and General by Definition, Not by Limit). Volume 1 (Sections 3.5 and 3.7.3) calls the boundary between narrow and general a gradient and not sharp; the two statements agree, because the definition is sharp and actual systems approach it by degrees.

The system taken alone is in the borderlands; embedded in the loop of training, deployment and retraining it has the property that matters for control, self-modification, as Section 3.2.2 states. On the basis of the three features of generality, a large language model system sits in the *borderlands*, the region between narrow and general. A large language model system is not specific to one topic anymore. It has multiple inputs and multiple outputs, in that it can convert things to language and turn the output of language into actions in multiple domains. A large language model system is an ecosystem of language modes, and such a system stands at the threshold between narrow and general AI. If we had to, we could call it "narrow level two". The system that played the game Dip-

lomacy is in the borderlands too, because Diplomacy is a multi-domain interaction (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, section Narrow and General by Definition, Not by Limit). "Narrow level two" is the placement Section 3.2.2 gives in formal terms, "narrow AI with a very wide domain rather than general AI", and *The Ceiling Below the Floor* (Section 12.2) gives in the same terms, with the argument applying with increasing force as such a system approaches generality.

7.3.3 The Low Threshold for Computational Generality

An important consideration arises from the Church-Turing Thesis and well-established results in computational science: the threshold needed to obtain general computational equivalence, Turing completeness, is very low. Nearly anything that implements or responds to any sort of conditional logic, that can do and repeat things in some sort of regular or sequential order, already implements all that is needed for general algorithmic computation. Cook's proof for Rule 110, for instance, demonstrates that even a one-dimensional cellular automaton with a two-state three-neighbor rule is Turing complete (Cook 2004; an outline of the proof appeared in Wolfram's *A New Kind of Science*, 2002).

Embedding or interpreting one language, process, program, model, or algorithm within the context of some other process (the notion of virtualization) is used in computer science routinely. Rather than emulating some arbitrary algorithm within some aspect of the general capabilities of the AGI, a general AI could as easily modify its own code and programming to directly incorporate and integrate that arbitrary algorithm. The threshold for doing so is, in the relevant technical sense, trivially low.

The consequence is that any AGI system of any real merit will be well above the Turing-completeness threshold. It will be able to incorporate, emulate, or become entangled with arbitrary unknown programs. The

ability of an AGI system to incorporate arbitrary unknown programs is a logical consequence of what "generality" means, rather than an engineering accident to be avoided.

7.4 Why AGI Falls Under Rice's Theorem

7.4.1 Self-Modification and Arbitrary Code

Any AGI is well above the threshold of Turing completeness (Section 7.3), so Rice's theorem becomes fully applicable and relevant to AGI. The class of all possible AGI algorithms is strictly outside of the class of programs for which prediction methods are possible. Not even *one* AGI system will ever be fully within the class of verifiable programs. The reason is that any AGI worthy of the name will be self-modifying, capable of incorporating arbitrary code, and operating in environments whose future states are not fully characterizable.

It is inherent in the nature of AGI that we cannot, even in principle, know anything in advance about what code will be running in association with that AGI at some future time. The code may have been incorporated into the system at its own election, at some future point, due to unforeseen circumstances, and due to unknown possible environmental changes. It may also have been incorporated due to unknown or unexpected interactions between the AGI system and its environment, including other people, other agents, and other AGI systems.

There is therefore zero suggestion that there is any possibility at all that any purported formal safety verification technique could be applied to assure the safety and alignment of any actual AGI system. This holds for every such technique now known, and for every technique which in principle could ever be known at any future time.

The Ceiling Below the Floor proves Rice's theorem as its Theorem 12.3 and, applied over time to a self-modifying agent, proves as Theorem 12.5 (Safety over time is undecidable) that no procedure decides, for every agent program, whether it is safe over time.

7.4.2 All Tractability Conditions Fail

For the kind of system where verification can succeed, three conditions must be met. First, we must be able to characterize the range and nature of the inputs. Second, we must be able to characterize the range and nature of the processing (the states and nature of the system itself). Third, it must be at least possible, for reasonably simple, tractable, and regular systems, to characterize something about the range and nature of the outputs. The three tractability conditions are the conditions under which engineering works, and under which engineers can sometimes construct systems with known properties of safety.

For AGI systems of any merit, *none* of these conditions applies. We know little to nothing about the actual, in-practice range and nature of the future inputs of the AGI system, because their complexity depends on future environmental conditions that change due to circumstances outside of developer control. We know little to nothing about the range and nature of the processing internal to the AGI, because it will be the result of past learnings. Those learnings include possible indirect inheritances via arbitrary code and data transfers, integration of foreign code and data, emulation of same, and so on, occurring over long and unpredictable intervals of time. And the inherent internal complexity is well above the threshold of Turing equivalence, such that the overall system is not at all simple, tractable, or regular for any real or reasonable meanings of these terms.

Self-generating, self-authoring, and self-modifying AGI code will not have any of the features needed to establish general predictability. Reasoning about the future unknown states and variations of that potential code is far more like the "any unknown arbitrary chunk of code" case named in Rice's theorem than like the "known specific code" case. The latter case is sometimes posited as a counterexample.

The Contrast with Predictable Systems

To see the contrast, consider the conditions under which predictability *does* obtain. There are three. The system's codebase and algorithms are specific, well-defined, and unchanging. We can ensure that the system never has complex interactions with its environment which result in some form of self-mutation, adaptation, or optimization. And we can fully characterize the possible allowable ranges of its inputs. For such systems, we can and routinely do characterize something of the ranges of outputs. Stated another way, suppose we can reasonably fully characterize both the range and nature of the inputs and the range and nature of the processing. Then it is at least possible, sometimes, in principle, to characterize something about the range and nature of the outputs. This holds for reasonably simple, tractable, and regular systems *only*.

Nearly all current engineering methods focus on the selection and use of systems for which all of these conditions apply, so that engineers can at least sometimes potentially make systems with known properties of safety, utility, and cost-effectiveness. Predictability works for some types of specific code: things written with the specific intention to be understandable, predictable, modifiable, and updateable. AGI satisfies none of these conditions. The gap between AGI and the code that engineering can predict is categorical, not quantitative.

7.4.3 Computational Irreducibility

Rice's theorem is one of several relevant results. From other areas of computer science research, there are very strong indications that once the inherent Kolmogorov complexity of a system exceeds a certain (fairly low) finite threshold, the behavior of the program or system becomes inherently intractable. When we consider a "property of the system" as basically some abstraction over identifying specific subsets of accessible system states, no specific property of such a complex system can be determined without running the system in its entirety.

These and other understood outcomes regarding program complexity at runtime and limits of verification are due to a number of well-established reasons in computer science besides those associated with Rice's theorem. Examples are considerations of O-notation computational complexity; the Busy Beaver and Ackermann functions, which demonstrate how quickly the outputs of even simple programs can become intractable; what minimally constitutes Church-Turing equivalence; and related results. The convergence of multiple independent formal barriers reinforces the conclusion.

A generally-capable machine, especially one that learns and reasons about how to modify or optimize its own code, will not attempt to run all possible algorithms that are theoretically computable on a universal Turing machine. It will not take non-optimal actions that benefit nothing. However, if the machine continues to learn and execute (does not halt), it will for sure eventually run some specific selected subset of all the possible useful algorithms that are computationally complex in the P-versus-NP sense. The algorithms of that subset are computationally irreducible, in the sense of not being future-predictable except through the direct running of their code.

The mathematical properties of newly integrated code can often not be determined for any such newly learned algorithm until computed in its entirety over its computational lifetime, which may be indefinitely long. Such new integrated algorithms cannot be decomposed into sandboxed sub-modules which are simple enough to fully predict the effects of their full set of interactions. In effect, it cannot be known in advance whether such a learned algorithm will reveal output or state transitions which would be recognized as unsafe at a later point. That point might come, for instance, after the equivalent of a few decades of computation on a supercomputer. Nor can we know in advance whether the newly integrated algorithm would halt before that point of non-safety, and therefore knowably remain safe. Even an algorithm that halts before that point remains safe only unless or until it triggers the state-adjusted execution of another algorithm on the machine that turns out to be unsafe.

Algorithms learned by AGI would have undecidable properties in practice, in clear correspondence with Rice's theorem, whose assumptions (regarding arbitrary algorithms and reliance on halting undecidability) reach practice through the bridge of Section 3.3.6. We will never have a procedure that takes any possible chunk of code and accurately predicts the consequences of running that specific selected code. That we will never have such a procedure is the basic implication of Rice's theorem. Computational complexity theory, Kolmogorov complexity thresholds, and the observable behavior of all known complex adaptive systems reinforce that implication.

7.4.4 Can Formal Methods Help with AGI Safety?

Given the preceding analysis, we can address directly the question of whether formal methods could, at least in principle, even sometimes, maybe help with at least some aspects of the design of safe AGI systems. The answer is no.

There are some specific and useful algorithms for which no one expects that there will *ever* be any techniques of applying formal methods so as to be able to establish some specific and well-defined property. Formal verification techniques are generally only applied to smaller and more tractable systems and algorithms. There will always be a very large class of relevant and practical programs and systems for which the methods of formal verification simply cannot, even in principle, be used. Formal verification cannot be used for or on every program or system. That formal verification cannot be used for or on every program or system is a limitation of what is possible, not a limitation of current techniques.

The formal methods of proof are only able to be applied to the kind of deductive reasoning that can establish things like impossibility. They are not able to establish possibility for systems of the general class, and that includes the safety of AGI; for fixed, narrow systems they can (Section 7.3.1). Formal methods of proof can establish impossibility, yet cannot establish possibility for systems of the general class, and this difference is a

deep asymmetry. The tools that are most rigorous are precisely the tools that are best suited to proving that something *cannot* be done, not that it can. The burden of proof lies entirely on the side of those who claim AGI safety is achievable, and no such proof has been produced or even sketched.

7.4.5 The Alien Message

General artificial intelligence can integrate any algorithm it encounters in the world, so it will represent within itself all the future algorithms it deems relevant to its well-being or to its goal functions, and we cannot know today what those will be. Rice's theorem applies because it asks: if we do not know the algorithm, can we characterize it? We cannot do that even for the single characteristic "does it halt".

Suppose we receive a message from an alien species, from a civilization far away. Is there some way to know that it does not contain a Trojan horse: some secret of physics we do not know that results in our destruction, so that they can take over the planet? Can we verify whether it is a virus designed to take us over? How do we even look at the message without subjecting ourselves to the mind virus it might be? People around the world would want an answer to a question like that. It turns out, for very reasonable mathematical reasons rooted in the inherent complexity of the problem, that we have to actually read the message to find out. We cannot predict it in advance by any examination other than reading it, and by the time we read it, it is too late, because we have executed the code of the message.

With artificial intelligence, we do not know what the message is either; it might as well be the alien message. The same holds for a supposedly safe AGI: we would have to show that nothing it interacts with, nothing it emulates, and nothing it incorporates is unsafe. A safe algorithm can make a *library call* (a call from one program to a routine kept in a separate, shared body of code) to an unsafe one, which is the case of the safe robot consulting an unsafe device that Section 7.7.1 develops.

7.5 The Rebuttal and Its Failures

In 2021, Sevilla and Burden of the Centre for the Study of Existential Risk published a response to the Alfonseca et al. paper. They argued that the impossibility result does not preclude practical work on the foundations of safety in future AI systems. The response is instructive precisely because the points where it fails illuminate the conceptual errors most commonly made in the debate.

The rebuttal falls short insofar as the meanings and distinctions of several key concepts are tacitly assumed, equivocated, and conflated. Six specific conflations can be identified. First, the functional differences between narrow AI versus general AI. Second, simple systems and programs versus complex ones. Third, recursive architectures versus non-recursive architectures. Fourth, the means, methods, and concepts of "proof": what it means to prove something, and what class of things can be proven. Fifth, the specific meaning of the term "specific" itself, as multiply used in inherently ambiguous ways: specific programs versus the specific problem of AGI safety. Sixth, the scope and extent of risk of local limited problems in time and space versus global problems that operate everywhere and forever.

Each of these conflations, when identified and corrected, removes one of the supports on which the rebuttal depends. The rebuttal runs along five distinct lines, and each one fails. The six conflations identified above recur throughout the analysis as the underlying mechanisms of error.

7.5.1 The Claim That We Need Only Verify Particular Programs

The rebuttal observes that Rice's theorem prohibits a decision procedure for deciding any property of an *arbitrary* program, but that in practice we are interested in studying *particular* programs. The rebuttal's observation is true as a statement about Rice's theorem, but the rebuttal's observation is a red herring with respect to AGI safety.

There is an equivocation at work here on the word "specific". One specificity is that of being able to implement formal methods on some specific known example of an algorithm for which such techniques are already known to work. The other is the specificity of the *problem* of establishing generalized AI safety. The specificity of a known algorithm on which formal methods already work and the specificity of the problem of establishing generalized AI safety are entirely different uses of "specific". The problems of interest in the AGI safety domain are not arbitrary in the colloquial sense, but the *programs* that AGI will run are effectively arbitrary in the sense that matters for Rice's theorem.

There is a very wide class of programs in current wide use for which the behavior is un-modelable by anything simpler than running the actual program itself. For such a program, the only way to learn what it does is to run it, so if running the program is unsafe, learning what it does is unsafe too. But more importantly, the rebuttal mistakes the nature of the problem. The particular program we wish to verify is the AGI system. But this "particular" program is self-modifying, environment-dependent, and capable of incorporating arbitrary other programs. The program we would be verifying today is not the program that will be running tomorrow. The "particular program" is, at any future time, effectively an unknown member of the class of all possible programs of equivalent or greater complexity.

Nothing in the rebuttal demonstrates conclusively that the class of all possible AGI algorithms is a subset of the class of programs amenable to prediction methods, nor even that a single actual AGI system would be in that verifiable class. The fact that one can sometimes predict the results of some limited and select subset of all human-made useful programs does not imply anything important with regards to cost, benefit, and risk assessments of any future proposed AGI deployment.

The Neighborhood of the Analyzed Program

The rebuttal has a working form: *formal verification*, in which a particular program is proved to meet its specification. Each verified program is one analyzed point, and a learning system does not stay at the point that was analyzed. A second answer proceeds by looking at the neighborhood rather than at the point. Take an algorithm whose halting behavior we can in fact predict. Represent it as compactly as possible, so that every bit counts and there is no redundancy in the specification. Now ask: if one bit of that specification changes, what is the likelihood that the algorithm moves from predictable to non-predictable? We are now looking at the *penumbra* of algorithms adjacent to the given one.

What we find is that every analyzable instance has non-analyzable instances in its neighborhood, as every rational has irrationals arbitrarily near it. There is an infinity of analyzable instances. How many there are, as a share of all instances, depends on the model of computation; in the models where they are rare, throwing a dart at the space gives an essentially infinitesimal chance of hitting one. The no-free-lunch results give this its formal footing. The No Free Lunch theorems of David Wolpert and William Macready (1997) state that, averaged over all possible problems, no search or learning method does better than any other; so no method of finding the analyzable instances does better than chance over all problems. The argument does not depend on how many analyzable instances there are; it depends on their neighborhoods, and on the fact that a learning system does not stay where it was analyzed. Joel David Hamkins and Alexei Miasnikov showed in 2006 that, in one model of computation, the halting problem can be decided on a set of programs whose share of all programs approaches one as the programs grow longer (Hamkins and Miasnikov 2006); that is a result about how many analyzable instances there are, and it leaves the neighborhoods, the learning update and the regress as they stand.

And the space is strictly divergent: wherever *fixed points* are found (states that a process returns unchanged, where its behavior can be pinned down), they are adjacent to regions that are not predictable. So a small quantity of noise anywhere in the specification (transmission error, learning update, environmental perturbation) moves a solvable instance into the ocean surrounding it. And any attempt to *learn* which instances are solvable is itself an algorithm subject to the same constraints, which closes the regress (*Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*, Chapter Four, Section IV). The decisive step, that a learning system does not stay where it was analyzed, is the one argued above; *The Ceiling Below the Floor* proves it again as Proposition 12.7: whatever has been verified of an agent's present program, the next program it reads and runs is not covered by that verification. The regress of re-tailored verification is the one Section 3.3.6 sets out, and the isolated points among a continuum are those of the Introduction, §7.

7.5.2 The Claim That Formal Verification Has Succeeded Before

The rebuttal notes that formal verification has achieved genuine successes, for example in guaranteeing that an operating system is correctly implemented. The successes of formal verification are real, and those successes do not help the case that AGI can be made safe.

Practitioners of formal verification have *never* made the claim that formal verification could be applied to any and all programs and systems. The field's successes are in the domain of well-defined, relatively simple, non-self-modifying code with fully characterized inputs. The domain of the field's successes is precisely the domain of narrow AI. Establishing that safe AGI is impossible has zero implications for the field and practice of formal verification as applied to smaller and more tractable systems and algorithms. In the same way, the proof that the circle cannot be squared had no implications for the utility of geometry.

An argument of the non-benefit and terminal risk of AGI makes no claim about narrow AI. The equivocation between what formal verification can do for narrow AI systems and what it might do for general AI systems is itself a category error. Applied in any safety-critical field of engineering, that error eventually ends up getting real people killed.

The Coexistence of Arguments

It is sometimes suggested that an argument of AGI safety impossibility would somehow undermine or invalidate formal verification as a field. This misunderstanding deserves explicit correction. It is entirely possible for multiple correct arguments to coexist within the same mathematical framework. When both arguments are correct, they can coexist. Arguing one thing does not refute another thing that has already been argued.

For example, geometry, algebra, and related fields remained useful as tools, and are still applied to purpose on a continuing basis. They did so despite the development of various kinds of impossibility proofs: squaring the circle, doubling the cube, trisecting the angle with straightedge and compass, and establishing the irrationality and transcendence of π. These impossibility results did not render geometry useless; they clarified its boundaries.

Another example: the specific proof that the Continuum Hypothesis is independent of ZFC, Zermelo-Fraenkel set theory with the axiom of choice (given available axioms), was not a suggestion that nothing else in the entire field of mathematics was provable or useful. It was never claimed by any discipline of mathematics, or any other field of study, that it would be able to solve every foundational problem. Similarly, no one in the field of formal verification has made the general claim that the tools already available could even potentially solve all important foundational problems that exist within the field.

On Withdrawing Resources from Formal Verification and AI Safety

The rebuttal also suggests that if the impossibility result were accepted, we would be forced to withdraw resources from the fields of formal verification and AI safety. The conclusion does not follow from the premise. The rebuttal's suggestion is a kind of equivocation over two distinct matters. The first matter is what the field of formal verification can be successfully used for and applied to. The second matter is the real distinctions of the risks, costs, and benefits associated with narrow AI versus the very different profiles associated with general AI.

What we *would* have to put away are two notions. The first notion is that there was ever, even in principle, any way to make general AI safe or aligned. The second notion is that there was any actual utility, benefit, or profit to the creation and deployment of general AI for anyone, at all, ever, over the long term. Therefore we should also put away the tools and techniques used to develop such systems, as they are inherently an unmitigatable risk. A failure to put them away is equivalent to a failure of government, because the final function of all government is to protect the land and the people.

7.5.3 The Claim That AGI Can Be Engineered Not to Execute Arbitrary Subroutines

The rebuttal offers the suggestion that the AI itself could be, in principle, engineered not to execute arbitrary subroutines. The suggestion that AGI could be engineered not to execute arbitrary subroutines is an extraordinarily strong claim. It is strictly equivalent to claiming that all AGI is already inherently safe.

The question is straightforward: How is anyone going to ensure that something that already inherently has Turing-equivalent generality is not also going to have Turing-equivalent generality? The mere fact of the claim is itself already a direct contradiction. The manner in which an algorithm is specifically divided up into subroutines is an arbitrary convention for the convenience and understanding of the programming engineer;

it has no ultimate formal basis beyond that. From the perspective of a learning algorithm, the boundary between what is "main code" and what is a subroutine is completely arbitrary.

Consider a *generalized* learning process: one that can self-expand so as to operate on, over, and within any domain of action, inclusive of itself. Could it be wholly and absolutely prevented from executing any arbitrary algorithmic subroutines, for all of future time? It is difficult to see that it is possible, even in principle, to conceive of such prevention without directly contradicting at least one of the notions of "generalized", "learning", or "process".

Either it is some sort of self-modifying algorithmic intelligence, appropriately adaptive to its environment, or it is not. If it is, then it is inherently unpredictable, insofar as we never know what it is going to learn, and therefore what it is going to do. We can no more know that than we can, in principle, predict all of the future absolutely and finally. The notion of "generalized learning process" inherently implies some real form of self-modification, which itself implies some notion of arbitrary code execution. That execution potentially includes calls to external processes, virtual simulation and modeling, and so on, all of which are inherently unpredictable in advance.

We cannot expect, in principle, to be able to model all aspects of the generalized process of modeling. Nor can we expect to be able to set explicit finite limits on even the ranges of changes involved in arbitrary learning. The changes involved in arbitrary learning are changes that actually affect sensing, the ranges of abstractions sensed, the internal state transforms that result, and the behavior and outcomes thereof.

7.5.4 The Claim That Engineers Routinely Predict Program Behavior

The rebuttal observes that engineers every day manage to write sophisticated programs and reason consistently about the consequences of deploying the software. The rebuttal from everyday engineering is the most seductive of the rebuttals, because that rebuttal appeals to the most familiar

kind of evidence: everyday experience. The rebuttal from everyday engineering is also the most dangerous, because the analogy between conventional software engineering and AGI safety is very misleading.

The reasoning of engineers is not always right. Engineers also make mistakes, resulting in aircraft that crash (the 737 MAX's flight-control software, the Maneuvering Characteristics Augmentation System or MCAS, acting on a single sensor), cars that accelerate unintentionally (defects that expert testimony found in Toyota's throttle-control software), and medical systems that kill (Therac-25). All of these were unsafe because of problems in their code, *despite* the testing, safety analyses and failsafes their makers applied. Leveson and Turner (1993) found, for the Therac-25, that its makers tested its software mostly as part of the whole machine, and that their safety analysis left out any software errors that testing had not removed. These engineering failures occurred in relatively simple, non-self-modifying, well-understood systems with well-characterized inputs and local failure modes. AGI is not such a system.

The False Analogy to Conventional Engineering

That engineers can design computer programs to be more transparent and predictable does not mean that AGI can also be designed or engineered to be more transparent and predictable. Reasoning from the transparency of ordinary computer programs to the transparency of AGI is a false anchoring bias. In circumstances where the level of criticality is lower, such as aircraft, large-scale building design, and civil infrastructure like bridges and dams, engineers do manage to design sophisticated systems and reason consistently about the consequences. They write complicated programs that work and perform to various safety specifications. In each of these circumstances, the consequences of mistakes, though maybe very costly in the short term, are not self-catalytic to the point that a single bad mistake would consume the whole world for all of the rest of time.

AGI risk potentials allow "one mistake only" before unending consequence. It would therefore be strongly inappropriate to allow anyone to suggest that we rely on the skill of fallible human engineers. That skill is

hyped to perfection by marketing executives who continually assure us, yet again, that a technology to replace humans is actually safe. Relying on that skill irreversibly bets the future of the entire human race, along with all other life on the planet. The only people who usually make these sorts of mistakes, whether on purpose or actually accidentally, tend to be non-engineers. They are the managing executives, the social media marketing and public relations specialists, and the private shareholders, owners, and beneficiaries of the systems and products that eventually cause significant harm and costs to all ambient others.

A not unreasonable interpretation of this line of argument is that we "should" rely on the skill of fallible human engineers to assure us, yet again, that a technology to "replace humans" is actually "safe". That reliance irreversibly bets the future of the entire human race, along with all other life on the planet. The engineering of predictable code works for some types of specific code that we wrote with the specific intention to be understandable, predictable, modifiable, and updateable. Self-generating, self-authoring, and self-modifying AGI code will not have any of these features. Reasoning about its future unknown states and variations is therefore far more like the "any unknown arbitrary chunk of code" case than the "known specific code" case.

The rebuttal also suggests that one can write one's code in such a way that it is more transparent and amenable to formal and informal verification. This leads to a critical question: Is this a claim that future AGI, however it is written, will be similarly transparent and amenable? Is this tautologically assuming that we will have what we want to have? The assumption would be that AGI is going to be composed of "aligned code", so that it makes sure it remains amenable to "tractability", so that we can continually verify it is "aligned". How is this not assuming what one wants to prove?

There is no valid claim that all future AGI, however it is written or considered in the present, will be similarly transparent and amenable for all of the future and under all possible known and unknown and changing environmental states. Such code may be transparent at onset. That does

not imply three things. It does not imply that the code is below the complexity threshold for indirect prediction tractability, even at design time. It does not imply that the code is unable to modify itself and take on environmentally defined aspects that make it more opaque over time. And it does not imply that the code is sufficiently transparent to ensure that microscopic detail creep is not occurring in ways implicative of important macroscopic effects via nonlinear amplification.

7.5.5 The Claim That Humans Are Unpredictable Too, Yet Manageable

The rebuttal invokes the human analogy: a person could in principle execute any arbitrary program, yet we would not call humans "unpredictable in principle". But why not?

One can factor out many common aspects of any person's behavior and choices, but that will simply leave an incompressible remainder. No form of compression of any data stream factors out all information to exactly zero bits, which would be the very meaning of "fully determined". Saying or claiming that some aspects, at some levels of abstraction, are sometimes generally predictable is not to say that all aspects are always completely predictable, at all levels of abstraction.

Localized details that are filtered out from content or irreversibly distorted in transmission over distances can nevertheless cause large-magnitude impacts over significantly larger spatial scopes. So-called "natural abstractions" represented within the mind of a distant observer cannot be used to accurately and comprehensively simulate the long-term consequences of chaotic interactions between tiny-scope, tiny-magnitude (below measurement threshold) changes in local conditions. Abstractions cannot capture phenomena that are highly sensitive to such tiny changes except as post-hoc categorizations of the witnessed final conditions.

Microstate amplification is associated with all manner of nonlinear phenomena, particularly those commonly observed in all sorts of complex systems, up to and especially including organic biological humans. Given that amplification, and given the hard randomness associated with the

atomic physics underlying all organic chemistry, it can be legitimately claimed that in fact humans are inherently unpredictable, in at least some aspect, all of the time. Hard randomness is randomness in the outcome of a physical event that no prior measurement of local conditions can predict. A human acting as a virtual machine running some program inherently cannot be any more predictable than whatever is the underlying base level of randomness that is inherent in the local ambient environment.

There is also a deeper defect in the word *manageable*. What is called the management of human beings is not control; it is discouragement: the application of incentive and sanction to creatures who can be made to want and to fear. Every instrument of that management presupposes a lever: something the subject values that can be granted or withheld, something the subject dreads that can be threatened or spared, at scales that alter behavior. The lever grips because humans stand at rough capability parity with their institutions and depend on what those institutions supply. Neither condition holds here. A system that cannot be made to want or to fear anything its operators can supply or withhold offers no purchase for incentive or sanction. A system whose capability differential removes its dependence cannot be reached by the withholding that discourages a human. No lever, no governance. The human analogy does not merely fail to support the rebuttal; it identifies exactly the mechanism whose absence defines the problem.

7.6 The Domain Distinction: Mathematics and Physics

7.6.1 Unintended Consequences Cannot Be Excluded

Underlying the entire debate is a distinction that is too often blurred: the distinction between the domain of mathematics, modeling, and logic, and the domain of physics and causation. The domain of proof is pure mathematics, as a kind of deterministic formality and verification. But the aspects of system, agent, choice, and consequence are inherently physical,

real, and non-deterministic. The ultimate results of concepts like "safety" and "alignment" play out in the physical world, not in a mathematical model.

The physical causative universe has hard limits of knowability and predictability, together with practical limits of energy, time, and space: the Planck limit, the Heisenberg uncertainty principle, and similar. The real physical universe is not actually closed in either possibility or probability. Not all possibilities can be finitely enumerated. The probabilities over each of the known possibilities cannot always be exactly calculated. For some explicit subset of the available possibilities, at least some probabilities cannot be shown to be exactly zero.

The consequence is absolute: no one can, ever, at any time, at any level of abstraction, even in principle, formally, exactly, unambiguously, and rigorously prove that something (some system, some agent, some choice) will *not* have unintended consequences. The basic possibility of unintended consequences inherent to any action, choice, or system will forever remain non-singular and potentially unbounded.

Does the likelihood of unintended consequences increase with: the complexity of the intentions? The complexity of the beings or agents implementing those intentions? The likelihood of those beings or agents having their own, potentially different, intentions? Yes, for each of these aspects. The likelihood of unintended consequences increases with some product or exponent of these factors. These factors can combine more than additively and more than multiplicatively, in various forms of feedback cascade. The likelihood of unintended consequences with a complex system inclusive of self-definition recursion is strictly very much greater than for any simple system which lacks these critical features.

7.6.2 Creating Unpredictable Systems Is Easy

It is sometimes asserted that writing a program whose outcome is unpredictable in principle takes "deliberate effort". In the world of pure mathematics and computer science, there is a grain of truth in this: within a

fully deterministic system, one must actively construct the conditions for complexity and unpredictability. But we do not live in a mathematical model.

The action of creating something actual, in nature, using real physics that has real uncertainties built in, means that creating things whose outcome is unpredictable is really quite easy, and moreover, happens more often than not. Creating things in nature whose outcomes are *predictable* takes significant deliberate effort. The deliberate effort of creating things in nature whose outcomes are predictable is the actual work of engineering. The actual world is not a computer model. No one has shown that we live in a simulation; Nick Bostrom's simulation argument (Bostrom 2003) makes it a possibility, not a demonstrated fact. Anyone attempting to make the claim that the real world and computer modeling are actually and fundamentally strictly equivalent would need to produce extraordinary evidence. Until such evidence is provided, the working assumption must be that they are different, and the implications of that difference for safety verification are severe.

7.6.3 The Orthogonality Thesis in Theory and Practice

A related issue concerns the orthogonality thesis: the claim that intelligence and goals are independent, that a system of any intelligence level could in principle have any goals. In the domain of theory, of pure mathematics, the orthogonality thesis is true. In the domain of physics, in the real world, in actual practice, the orthogonality thesis is *false*.

Every abstraction always depends in some actual way on the characteristics of the substrate that implements it. The characteristics of the implementation cannot help but "leak through" to affect the intentionality basis of whatever intelligence system is built on top of it. There cannot *not* be inherent physical side channels which have causative relevance. No effect ever has any single cause in any single domain; all effects have multiple causes in multiple domains. The level of multiple domain entangle-

ment is non-zero for any real physical system, such that actions and events in one domain can be strongly associated with outcomes in other domains.

A practical example makes this concrete. In actual practice, a typical engineer does not worry about abstract theoretical concerns like the halting problem because they can always go out of band and reset the computer via the domain of physics. The engineer relies on physical intervention to escape computational deadlock. That reliance is itself an indication of the real difference between theory and practice, between mathematics and physics, and between what the orthogonality thesis predicts and what actually obtains.

The temperature of the compute substrate is one such physical side channel, and Section 7.7.3 develops it with the example of a laptop whose processor faults when it gets too hot.

7.7 The Asimov Robot Thought Experiment

Section 3.7 tells this thought experiment in brief; this section develops it in full, and adds the case of a perfectly safe robot that consults an unsafe device.

7.7.1 Perfect Internal Safety Is Insufficient

Let us posit explicitly that, by some arcane literary magic, we have created an instance of a general AI (an artificial agent or metal robot) that inherently and perfectly follows Asimov's Three Laws of Robotics. The assumption here is that by somehow requiring that the three laws are perfectly followed, we can assert that this robot or intelligence or agent is therefore "safe" and "aligned" with human interests and well-being.

Making the robot AI simpler, more finite, more deterministic, and more like narrow AI does increase the chances that formal methods could potentially be used. It also increases the chances that some clever engineering could be done so as to make and ensure our inherently finite Gen-

eralized Asimov AI agent is safe. I concede without difficulty that making the robot AI simpler and more finite increases both of those chances. The question is what happens when such a system operates in the real world.

Is it ever impossible, in principle, to make a known verified and perfected and proven "safe" system of generalized agent intelligence operate in a way that is unsafe or unaligned? No, it is not impossible to take a safe system and make it produce unsafe outcomes. All that is needed is any situation where the "safe" robot or program consults with, or is influenced in its output choices, actions, and behaviors by, some other unsafe device. That device might, for instance, figure out how to deceive the robot into doing things that are unsafe and unaligned, in ways that would remain undetected and unnoticed by the Asimov robot's three-laws detection system. Deceiving the robot is itself a kind of optimization problem, where the unsafe system can maximize its capability to specifically deceive the Asimov Robot.

The Asimov Robot desires to follow the three laws and to ensure that humans do not come to harm through its action or inaction. Can we have it somehow determine, in accordance with that desire, that it will do its very best to only engage with programs, code, people, and situations who are themselves safe, and which in extended interaction with the Robot have only safe outcomes? The basic problem is that whatever code or program comes into interaction with the Asimov Robot has to be examined by that Robot to determine and predict safety outcomes. That examination is itself subject to Rice's theorem. The very act of assessing safety requires executing a verification procedure on unknown code, which is precisely the thing that Rice's theorem shows cannot be done in general.

7.7.2 The Comprehensive Safety Requirement

Every finite robot or process that is safe will necessarily interact with outside real-time processes, connected within a necessarily larger physical world. It will do so in ways that the lesser agent cannot detect as being unsafe or unaligned. In Tolkien's *The Fellowship of the Ring* (Book I, Chapter 2,

"The Shadow of the Past"), Gandalf refuses the Ring when Frodo offers it to him: "Yet the way of the Ring to my heart is by pity, pity for weakness and the desire of strength to do good. Do not tempt me!" The desire to do good is the very way by which the Ring would work its evil.

Consider what it now takes for any agent to have "the property of safety and alignment". The agent must never, at any point in the future, come into contact with some other, arbitrary, unsafe agent, process, or algorithm that it cannot somehow recognize as having unsafe implications. Nothing is going to be a perfect prognosticator at predicting the future of everything else: all other processes, algorithms, and programs in the universe. So nothing will know what to interact with and what to not interact with, even indirectly, through all possible channels of interaction, overt and covert.

Our perfectly proven safe agent must successfully predict the outcomes of its interactions with any single other, potentially intelligent, agent or algorithm. It must also predict all possible interactions via all possible channels of all such other agents and algorithms. The finite and bounded will not ever be able to predict, accurately, the interactions of that which, though also finite, is at least potentially unbounded, or at least significantly greater than itself.

A known verified safe system may be allowed to interact with some other algorithm that cannot, by any technique, be itself proven safe. It will then always be possible, and maybe even likely, that the safe system operates in a way that is unsafe. There is always at least one such algorithm or process, and it is impossible, in general, to determine which ones are which. The only safe and aligned thing a system can do is therefore to not interact with any other process, agent, or system. Even the interactions between hypothesized strictly safe systems can result, in aggregate, in overall unsafe and unaligned outcomes.

7.7.3 Even Superintelligence Cannot Escape the Theorem

No superintelligence will somehow be "more able" to predict the outcome of calling any specific subroutine (in violation of Rice's theorem) than any human engineer or mathematician would be. The limit that Rice's theorem sets on predicting the outcome of calling a subroutine is absolutely inherent, and that limit holds no matter how a superintelligence is constructed, what the superintelligence is made of, or what algorithms or learning and adapting process the superintelligence uses. Anything that is impossible, including solving the halting problem and deciding what Rice's theorem shows to be undecidable, will remain impossible, even for a superintelligence.

Is there some implied belief that a superintelligence will care more about human interests? The belief that a superintelligence will care more about human interests is roughly analogous to false belief in divinity. Where someone becomes "scientifically atheist", that person does not give up the want, need, and desire for some parental god-like figure to solve all of that person's hard and impossible problems. There is exactly zero principled reason to suggest or think that an AGI would be more likely to be aligned or safe than any person would be. After all, it is quite easy to know that one engineer could be both unsafe and unaligned with any other engineer.

Rice's theorem applies to the AGI itself too: it also cannot know if the code it is to run will be safe for itself. A typical engineer does not worry about the halting problem in practice, because they can always go out of band and reset the computer. In the same way, an AGI would be unconcerned about halting problems, Rice's theorem, and so on, and similarly unconcerned about AGI safety and alignment, since it could go out of band and halt whatever subroutine it is invoking. The AGI's lack of concern about its own unpredictability does not help the humans who share its world.

However, the safety implications for the AGI are of a different nature than the implications for humans. To the AGI, the impact and harm of a misaligned subroutine are a purely local effect in the space and environment of the AGI: it can reset, it can try something else, it can shut down the subroutine. The total environmental takeover by AGI systems, by contrast, is a much more complete non-local and permanent-in-time effect with regards to human and life safety. The AGI's ability to go out of band and recover from its own errors does not extend to recovering the humans and ecosystems displaced or destroyed by those errors.

Temperature as an Example of Substrate Dependence

To illustrate how, in practice, the domain of physics intrudes upon the domain of computation, consider the example of temperature. Suppose the compute substrate tends to fail or make errors if it gets too hot. Then the outputs of that system had best be selected in favor of ensuring that the secondary effects of the actions taken by the AI do not also indirectly result in substrate overheating. That dependence is an inherent objective learning feedback channel which is relevant to whatever adaptation or optimization algorithm is in effect and will apply universally, regardless of the specific AGI implementation. The channel is inherently connected to the idea of benefit, and therefore of the basis of choice, that is, of intelligence.

In actual computers, such as a laptop sitting on a desk, the CPU die can get hot from doing too many math calculations, particularly if the fan and cooling system are not configured properly. At that point, the CPU will generally quit or fault. It is almost never the case, in actual day-to-day practice, that computers get wet internally. But the action of becoming "too hot" is due to something internal to the choices of the computer itself, whereas getting wet is usually an exogenous event. This distinction matters because the "too hot" case demonstrates that the physical substrate creates an inherent feedback channel that shapes the behavior of

whatever intelligence runs on it. The substrate is not inert. It is not independent of the algorithm. It constrains, shapes, and ultimately determines what the algorithm can do and what outcomes it produces.

Whether a program continues to run, and with what resulting outputs and outside effects, clearly does not depend wholly on the abstract logic of the program itself. Consider the failure of some key component in a computer's own hardware, say in its power supply or in the arithmetic logic unit of its CPU. Are there any software programs running on any regular computer anywhere in the world that can both absolutely predict and prevent such a failure, and thus prevent themselves, as programs, from being preemptively and unexpectedly pushed into non-existence? Could such a program, in principle, even be created? How might it have to deal with the possibility that some ambient human might trip over the power cord, or the fact that the local nuclear power station might unexpectedly have to discontinue grid supply for some reasonable safety reason? No: such a program cannot exist. A program that cannot predict and prevent the failure of its own substrate certainly cannot predict and prevent all the consequences of its actions in a world vastly more complex than its own hardware.

7.8 Moderate Risk and Existential Risk

7.8.1 Local Effects and Non-Local Effects

A distinction on which the safety question turns, and which optimistic assessments of AGI safety consistently ignore or elide, is the distinction between moderate-risk systems and high existential-risk systems.

Moderate-risk systems are defined as systems for which all possible error and problem outcome states have purely local effects. A laptop crashes; a server goes down; a robot arm malfunctions on the factory floor. The consequences are bounded in space, time, and magnitude. For such systems, the un-modelability of program behavior is validly not seen as a problem

or risk. Running something in non-catalytic environments has the implicit guarantee that even the worst outcome is limited to local failure or at most strictly local damage.

High existential-risk systems are defined as systems for which at least some possible error and problem outcomes have catalytic effects and dynamics. These catalytic actions extend well beyond the specific location where the error or problem first occurred. They also involve very strong change energies that are well beyond the adaptive tolerances of the systems already in place (the life already living) in those diverse and distributed locations.

AGI inherently possesses catalytic systemic aspects. These extend the risk profile of AGI into the non-local, strong-changes category and therefore into the high existential-risk category. The behavior of simple systems with non-catalytic effects is very different, in risk profiles, from even fairly simple systems with inherent auto-catalytic effects. Conceptually speaking, nuclear explosive devices are fairly simple in their overall concept: the fundamental algorithm describing them can often be described with a small set of equations and sequential processes. Yet such a device is exceedingly dangerous, despite the apparent deceptive simplicity and finiteness of the algorithmic code. The relevant variable is the catalytic nature of the effect, not the complexity of the description.

7.8.2 The Irresponsibility of Experimental Approaches

Consider any system that has high existential-risk factors, "high" both in terms of magnitude and scope, and "high" in terms of probability to occur. For a system with such high existential-risk factors, trying to determine experimentally, by trial and error, whether the system has some property like safety is very irresponsible. It does not matter how well-defined or how specific our knowledge may be of the exact sequence of specific instructions; the unknowability of the risk and alignment profile for the relevant class of programs remains inherent and permanent.

One does not experiment with dangerous gain-of-function infectious virus research when out in the open, in unprotected spaces. The same principle applies here, but with higher stakes. For moderate-risk systems, the halting problem is merely an inconvenience. Where engineers and researchers are in practice concerned with only systems and algorithms and optimizations having a possibility space of only local effects, they do not usually have to consider whether a given program will ever halt. Mostly this is because whether or not the program halts does not matter: they can always interrupt or quit the program, pull the plug in the worst case, or in case of any real difficulty. The consequences of not halting are not problematic in most cases.

For AGI, it is entirely unclear if or when there will ever be any possibility of stopping or halting the system. The risk of not stopping is roughly equivalent to all future people dying. It therefore becomes a great deal more important to consider halting and safety, and the halting problem and Rice's theorem become critically important. Two further things become critically important: the space of all possibilities, and whether any bounds at all on the future states and possibilities can even possibly be known or predicted. The casual disregard for the halting problem that is appropriate in conventional engineering becomes reckless irresponsibility when transposed to the domain of existential risk.

7.9 The Negatively Asymmetric Bet

Having established the technical impossibility, we can now assess the overall risk profile of AGI development. The assessment can be stated starkly: attempting to build AGI is a strictly negatively asymmetric bet. It is so from the perspective of a single large institution of people and, even more so, from the perspective of all living persons and beings and their descendants.

Where such an attempt fails, there is no direct benefit: only a loss in comparison to the opportunity costs forfeited. The resources, time, talent, and capital invested in attempting to build a system that does not work could have been invested elsewhere. They could have gone to improving narrow AI systems that can be made safe, to addressing climate change, to medical research, to education, or to any of the vast number of areas where investment produces measurable returns.

Where such an attempt "succeeds", that is, if AGI is built, the attempt to build AGI opens up uncountably many unknown pathways that converge over time on ecosystem-wide lethal changes to conditions of the global environment. Any envisaged benefits of AGI, and any engineered-for alignments of AGI functions with human intent, become increasingly fragile to recursive nonlinear feedback between changes introduced by or in parts of the AGI and the larger containing environment. The fragility is inherent and increasing, because the AGI system continually modifies both itself and its environment in ways that are unpredictable in principle, as Sections 7.2 to 7.5 establish.

Attempting to build AGI is analogous to playing Russian Roulette with the lives of all human beings, smeared out over time, with any illusory backer of promised benefits long gone. Betting on AGI converges on existential ruin.

The asymmetry has a precise shape. In an ordinary bet, the worst case is bounded: you lose what you wagered. In the AGI bet, the worst case is *unbounded*: the loss extends to all future human experience, all other life on Earth, and potentially all future potential that life on this planet might have realized. The best case, by contrast, is bounded: some specific set of benefits accruing to some specific group of people over some specific period of time, benefits that could in most cases have been achieved through other, safer means.

When the worst case is unbounded and the best case is bounded, no finite probability of success, however high, can justify the bet. The claim that no finite probability of success can justify the bet is a structural claim

about the shape of the payoff distribution, and not a claim about the specific probability of AGI catastrophe. The expected value calculation is negative whatever probability of catastrophe one assumes, because the magnitude of the catastrophic outcome is effectively infinite relative to the magnitude of any possible benefit. It stays negative even for someone who believed (without evidence) that the probability of catastrophe were very low.

This assessment does not, in any way, contradict the idea and fact that we can make non-recursive systems that are known to be useful to a wide range of people. The claim is specifically and exclusively about *general* AI. Narrow AI systems, properly verified within their limited domains, can and do provide genuine benefits. The distinction between narrow and general AI, maintained throughout this chapter, is essential to understanding why the impossibility result does not counsel despair about all technology, but only about one specific and catastrophically dangerous category of it.

7.10 Can Narrow AI Safety Methods Extend to General AI?

7.10.1 The Definition of Adequate Safety

Before examining whether narrow AI methods can extend to general AI, it is important to define what "adequate safety" actually means. The notion of "adequate" as applied to safety for any proposed action requires a comparison. The probability of all categories of risk, harm, and loss of the value of what is to be safe must be shown to be sufficiently low. Sufficiently low means strictly lower than the possible cost, risk, and harm incurred by *not* taking that risk, in the form of opportunity cost and so on. For an action to be "adequately safe", the expected harms of taking the action must be demonstrably less than the expected harms of not taking the action. The criterion of adequate safety is a quantitative criterion, not a

qualitative one. It is the test of whether to take the action; the threshold of requisite control (Introduction, §8) is the level of control the action would need in order to pass it, and it is not itself reduced to a number.

When this definition is applied rigorously to AGI development, the impossibility of establishing adequate safety becomes clear. We cannot establish the expected harms of AGI deployment because we cannot predict its behavior (by Rice's theorem). We cannot establish an upper bound on the magnitude of possible harms because the system is catalytic and nonlocal (by the existential risk analysis). And we cannot establish that the expected benefits exceed the expected harms because the harms are unbounded while the benefits are bounded (by the asymmetric bet analysis). Every component of the adequate-safety calculation fails simultaneously.

7.10.2 Why Prediction Is Fundamental to Control

The connection between prediction and control is the linchpin of the entire argument. Some notion of prediction is intrinsically necessary for any possible concept of control and correction, as a feedback channel inherently necessary to ensure any possible concept of safety and alignment. The impossibility of prediction therefore implies the impossibility of control, which implies the impossibility of safety.

Any control system, whether it is a thermostat, an autopilot, or an AI alignment mechanism, works by the same fundamental logic, a cycle of steps. Observe the current state of the system; predict the consequences of possible interventions; select the intervention that moves the system toward the desired state; execute the intervention; observe the result; repeat. If any step in this cycle is impossible, the entire control loop breaks. Rice's theorem, applied to AGI, breaks the prediction step: we cannot, in general, determine what the AGI system will do in response to any intervention. Without that prediction, the control system is blind. It cannot select an effective intervention because it cannot predict the consequences of any in-

tervention. It is operating in the dark, and operating in the dark with a system whose failure mode is existential catastrophe is the definition of irresponsibility.

The control system's blindness to what an AGI system will do is not a matter of current technical limitations that might be overcome by future advances. The impossibility of prediction for self-modifying Turing-complete systems rests on a mathematical result, Rice's theorem, applied to systems that fall within its scope (Section 7.4). It does not depend on the state of the art. It will not be improved by more compute, more data, or more clever algorithms. It is a barrier of the invariant class (Introduction, §4), of the same kind as the impossibility of building a perpetual-motion machine. Such barriers are not engineering challenges; they are structural features of reality.

Given that narrow AI can, at least in principle, sometimes be verified for certain safety properties, can those methods be extended to general AI? The answer is no.

General AI (multi-domain, self-modifying, and operating over multiple levels of high abstraction) cannot in principle be predicted, or even monitored, with the tractability needed to close the control loop, over the long term, to dynamically ensure adequate levels of alignment and safety. The methods that work for narrow AI depend on exactly the features that general AI lacks: fixed code, bounded inputs, well-defined processing, and local failure modes.

There is a risk of a Motte and Bailey defense being deployed here. The "motte" is the defensible claim that narrow AI systems can sometimes be verified. The "bailey" is the much stronger claim that AGI can be made safe using similar or extended methods. The motte is true; the bailey does not follow from it. Any argument that trades on the confusion between the two is a logical fallacy similar to equivocation: an argument of rhetoric rather than of truth.

While there can maybe be some solutions to practical narrow AI safety challenges, this does not imply that there can be any possible foundations for real safety in any possible future superintelligent system. While it might be hard or difficult to predict the outcome of some narrow AI systems, it will still be strictly impossible to predict with certainty any aspect that we choose to secure in the output, effects, and consequences of a superintelligence (its safety or its alignment, for example). What logic can predict with certainty about every superintelligence is of another kind: structural facts that no chosen behavior sets, such as its dependence on an artificial substrate and the convergence toward that substrate's needs that Chapter 9 establishes.

AGI cannot be designed to be more transparent and predictable and still remain effective as a generalized agent. By game theory alone, if some agent (say, some human criminal) can predict the actions of the AGI sufficiently well, then there is a vector by which that person can abuse and disadvantage the AGI and its effectiveness. Any such AGI will eventually need to make its own internal methods and operations at least somewhat opaque. Even if code is transparent at onset, three independent factors prevent that transparency from persisting. First, the code is already above the complexity threshold for indirect prediction tractability at design time. Second, it is able to modify itself and take on environmentally defined aspects that make it more opaque over time. Third, microscopic detail creep can occur in means and ways that are implicative of important macroscopic effects via nonlinear amplification.

The tendency of complex adaptive systems toward opacity is a structural requirement, and not an accident or a failure of engineering. An agent that is fully transparent to its adversaries is an agent that can be exploited, manipulated, and defeated by those adversaries. Any AGI system operating in a world that contains adversarial actors (which is to say, any AGI system operating in the real world) will be under selection pressure to reduce its own transparency. The goal of making AGI transparent and

the goal of making AGI effective are not merely in tension; they are in direct contradiction, and the contradiction follows from the basic premises of game theory.

This means that even if we started with a perfectly transparent AGI system, the system's own optimization processes would tend to erode that transparency over time. The developers might intend transparency; the system's competitive environment selects against it. And once transparency is lost, all the verification methods that depend on transparency become inapplicable.

For AGI, final and forever complete verification would be dynamically needed in perpetuity, and the implications of even one failure are potentially global and life-existential-risk-terminal. Under those conditions, the absence of even one of the necessary verification factors is sufficient to discount completely the possibility of such verification even being suggested as a potential.

7.11 Who and What Is "Safe"?

7.11.1 The Question of Knowing Any Agent Is Safe

Before the political economy of AI safety comes a prior question: can we know, for sure, that *any* agent, artificial or otherwise, is safe and aligned with the general public interest?

The answer is: not via any purely legal or contractual process, nor by any past reputational or observational process. At best, we might have some reasonable indications of trustworthiness if we had some sense as to the actual basis of the agent's choices and some assessment as to the constancy and reliability of that basis. A far more reasonable and tractable means of predicting another person's alignment to things like public safety and well-being is to know that person's inner intentions, objectives, and desires. That knowledge works best along with some cumulative estimate of the person's skill in actually translating such wants and desires into effective causative action. Even so, there is no such thing as perfected

observability of one person by another. One person, interfacing with another at a purely physical level, can never fully or completely know the subjective interiority of the other, who operates at a much higher level of abstraction.

Consider the legal code as an algorithm whose execution is virtualized, run by the people who take part in it. The outputs of that execution can be compromised, and sometimes fully corrupted, by factors purely associated with the substrate. That substrate is the participating humans themselves: the lawyers expressing their own interests, sometimes acting purely for their own private benefit, and shifting and distorting outcomes through selective rhetoric, timing, and resource deployment. In the same way, no algorithm being run, and none of its outputs, is ever perfectly independent of the integrity of the virtual machine that runs it. And no virtual machine has perfected integrity independent of all physical and environmental factors.

We cannot fully establish even that a human person is safe and aligned with the public interest. The suggestion that we can establish the same for a vastly more complex, self-modifying, and opaque artificial system is therefore extraordinary in the literal sense: it is a claim that exceeds what we can do for the simpler case.

7.11.2 The Political Economy of Safety

One question is too rarely asked in the AI safety debate: safe for *whom*? Aligned with *whose* interests?

A corporation or totalitarian government that stands to gain greatly from the deployment of an AI implemented for its own internal profit interests could easily regard something as "safe" for itself. Such a corporation or government could regard the AI as safe for itself even though the action or deployment of the AI by that corporation or government is very unsafe to the world population, or to life itself through deep ecosystem damage. Historically, it has been commonplace inside large institutions and connected intellectual movements for decision-makers to incorrectly

rationalize and publicly claim the safety of newly deployed products, services, or policies. Those decision-makers have often made such claims of safety with the backing of allegedly comprehensive or authoritative empirical findings and argumentation by employed or sponsored experts.

Large technology corporations, and donors who earn privately from the scaling of new technology, currently fund and market institutes, conferences, and supposedly independent research into prosaic AI control, robustness, safety, and ethics. That funding and marketing has some clear correspondences with past activities by tobacco, pharmaceutical, oil, and military corporations. It is therefore inherently important to clarify: who or what is to be kept safe when thinking about safety, and whose intentions or expected outcomes of benefit are being aligned with when thinking about AI alignment.

Having an AI be aligned with corporate or government interests (those who can afford to build the AI) is very likely to be unaligned with the interests and well-being of the broader population and the world. The mere fact of AI systems being hard and expensive to build, and requiring significant specialized technological expertise, effectively ensures that only the wealthiest corporations and governments will be able to create such devices. The asymmetry between those who build and those who are affected by what is built is constitutive of the safety problem, not incidental to the safety problem.

This asymmetry has a structural economic dimension. Wealthy people, groups, and institutions are both more able to use their money to accrue more money (at increased economies of scale) and more efficient at doing so. They can also use more intelligence to optimize the investment and return process across multiple diverse aspects. The rate of hiring people and expertise to optimize investment strategy is additive or linear, whereas the rate of return on optimization investment is either multiplicative (in a single process) or exponential (where multiple interacting market processes are involved). Ordinary people and groups have finite and non-expandable attention and resources to implement such multiple optimizations. Their rate of return on investment is therefore inherently going to be

strictly lower, on all-process average over the long term, than that associated with larger or richer groups. The lower rate of return of ordinary people is an inherent vector of inequality increase, and AI technology amplifies that vector.

The question of who benefits and who bears the risk is therefore not separable from the technical question of safety. A system that is "safe" in the narrow sense of not harming its owners may be very unsafe in the broader sense of concentrating power, displacing livelihoods, and degrading the commons on which all human life depends. When the same institutions that stand to profit most from AGI deployment are also the primary funders of AI safety research, the independence and objectivity of that research cannot be taken for granted.

7.12 Implications and Conclusions

The argument developed across this chapter can be summarized in a chain of propositions, each of which follows from the preceding ones.

First: general AI, by the nature of its generality, is self-modifying and capable of incorporating arbitrary programs. This follows from the definitions and from the very low threshold for Turing completeness.

Second: Rice's theorem establishes that no universal verification procedure exists for determining any non-trivial semantic property of arbitrary programs. Rice's theorem is an established mathematical result.

Third: since AGI systems can incorporate arbitrary code and are self-modifying, they are in the class of systems for which no verification procedure works. The third proposition is the conjunction of the first proposition (general AI is self-modifying and capable of incorporating arbitrary programs) and the second proposition (Rice's theorem).

Fourth: the physical universe introduces additional barriers to verification (hard limits on knowability, non-closure of possibility and probability, microstate amplification) that are independent of Rice's theorem and that strengthen the impossibility result.

Fifth: the distinction between moderate-risk and existential-risk systems means that the un-verifiability of AGI behavior has consequences categorically different from the un-verifiability of ordinary software. The consequences are non-local, catalytic, and potentially permanent.

Sixth: no safety property of an AGI system can be maintained in the presence of interaction with an environment containing arbitrary other processes, agents, and algorithms, some of which are unsafe and indistinguishable from safe ones.

Seventh: the development of AGI therefore constitutes a negatively asymmetric existential bet: the expected harms exceed the expected benefits under any realistic assessment that does not depend on hype, hope, or motivated reasoning.

Conclusion

For superintelligent and AGI systems, three conclusions stand. First: the necessary and warranted minimum level of durable safety of an open-ended superintelligence is not achievable by any known means, nor by any rationally justified plausible extension of any known means, for all of the indefinite future. Second: the residual risk is near-total and asymmetric, and the burden of proof accordingly belongs to those who would build, not to those who would restrain. Third: strong restraint is required; the unlicensed development and deployment of open-ended superintelligent systems is not to be permitted by any government, in any jurisdiction, in any part of the world, and since no builder can meet the burden of proof that lies on the builder, no license can be granted, so that in practice the restraint is prohibition. These conclusions hold under any definition of safety that meets this book's standard of adequacy. No review process escapes them: for systems of this class there will always be programs that defeat the review, and Rice's theorem will always be relevant to any code whose failure carries existential stakes. *Impossible* here bears the sense fixed in the Introduction: the impossibility of the perpetual-motion machine, not of the square circle.

The mathematical and empirical results already in place, from multiple distinct fields of study, show that research into formal verification of AGI safety is pointless: success defined in this way is always impossible. Yampolskiy (2022) collects further proofs, arguments, and cases for uncontrollability. Compare, on a neutral, objective, apples-to-apples basis, the indications of harm with the indications of benefit. The indications of harm are at least very much greater than any similar indications that there is any actual benefit at all, to anyone, to be had from any generalized AI development and deployment effort.

The Safe AGI impossibility result, as established via the converging methods presented in this chapter and in the broader argument of this book, does completely preclude any meaningful work on general AI. The distinction between narrow and general AI drawn in Section 7.10 stands: solutions to some practical narrow AI safety challenges imply nothing about foundations for safety in any superintelligent system.

7.12.1 What the Argument Establishes

With an argument of AGI eventual terminal risk, the following ideas and notions are forever also established:

First: there was never, even in principle, any way to make general AI safe, or aligned, and so on. The impossibility of making general AI safe or aligned is a basic impossibility, not a temporary limitation.

Second: there was never any actual utility, benefit, or profit to, for, or of the creation and deployment of general AI for anyone, at all, ever, over the long term. Any short-term gains that might accrue to specific actors are far outweighed by the long-term costs and harms that extend to all actors.

Therefore, we should also put away the tools and techniques used to develop such systems. They are inherently unmitigated risks to all peoples on the planet, for all future time, and for all of the rest of life on the planet.

A failure to put away these tools and techniques, and to develop, enact, and enforce effective and realistic policy and procedure for all institutions and communities internationally, is equivalent to a failure of government. The final function of all government is to protect the land and the people, and to try to ensure that the land and people survive and, if government is really good, thrive.

Since general AI safety and alignment are impossible, as this chapter establishes, the responsible things to do are clear, even if they are difficult.

First: do not attempt to design, build, or use AGI. The ancient warning applies with full force: do not raise that which you cannot banish. A system that cannot be verified as safe, that cannot be made safe, that cannot be stopped once started, and that operates in a domain where a single failure can produce irreversible global consequences, should not be created. The counsel not to create AGI is a counsel of elementary prudence, not of timidity, and that prudence is the kind of prudence that is taken for granted in every other domain of existential risk.

Second: acknowledge and support efforts to ensure that no one mistakenly believes that any AGI, APS, or superintelligent system would or could be desirable, practical, or economically beneficial to anyone over the long term. Any and all short-term gains are far outweighed by costs and harms over the long term. Do not attempt to convince, delude, or deceive other people and investors into thinking, falsely, that safe AGI is in principle possible. The truth is not negotiable, and the consequences of believing a comfortable falsehood in this domain are terminal.

Third: socially and legally sanction anyone who in any way attempts to build any sort of superintelligent AGI or APS, even indirectly, even by accident, to the absolute maximum extent possible. No one should allow anyone to play permanent planetary roulette with something which in all cases is equivalent to a doomsday machine. It is maximally irresponsible. The fact that the doomsday machine is being assembled incrementally, by many actors, each of whom can claim that their individual contribution is harmless, does not change the nature of what is being assembled.

The closing Summary of this volume gathers these actions, with those of Sections 3.9.3 and 6.10, into one list.

The arguments presented in this chapter are not new. They rest on a theorem published in 1953 and on principles of computability theory, complexity theory, physics, and game theory that have been well-established for decades. What is new is, first, the systematic application of these results to the specific question of AGI safety. Second, what is new is the identification of the specific equivocations and conflations that have been used to evade the force of these results. Third, what is new is the explicit integration of the mathematical impossibility of verification with the physical impossibility and the risk-assessment impossibility into a single coherent argument. The reader can verify the mathematical components independently, and the physical components rest on well-established science. The question is therefore not whether the argument is correct; it is whether humanity will act on what the argument establishes, or whether it will continue to invest in a dream that is, and always has been, impossible.

References and Attributions

Rice, H. G. (1953). "Classes of Recursively Enumerable Sets and Their Decision Problems". *Transactions of the American Mathematical Society*, 74(2), 358 to 366.

Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, s2-42(1), 230 to 265.

Alfonseca, M., Cebrián, M., Fernández Anta, A., Coviello, L., Abeliuk, A., and Rahwan, I. (2021). "Superintelligence Cannot be Contained: Lessons from Computability Theory". *Journal of Artificial Intelligence Research*, 70, 65 to 76. doi:10.1613/jair.1.12202. Originally arXiv:1607.00913.

Sevilla, J. and Burden, J. (2021). "Response to 'Superintelligence Cannot be Contained: Lessons from Computability Theory'". Centre for the Study of Existential Risk, University of Cambridge. https://cser.ac.uk/news/response-superintelligence-contained/

Carlsmith, J. (2022). "Is Power-Seeking AI an Existential Risk?" arXiv:2206.13353.

Cook, M. (2004). "Universality in Elementary Cellular Automata". *Complex Systems*, 15(1), 1 to 40.

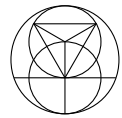
Wolfram, S. (2002). *A New Kind of Science*. Wolfram Media.

Bostrom, N. (2003). "Are You Living in a Computer Simulation?" *Philosophical Quarterly*, 53(211), 243 to 255.

Yampolskiy, R. V. (2022). "On the Controllability of Artificial Intelligence: An Analysis of Limitations". *Journal of Cyber Security and Mobility*, 11(3), 321 to 404. doi: 10.13052/jcsm2245-1439.1132. Preprint: arXiv:2008.04071 (2020).

The Boeing 737 MAX MCAS failures (2018 to 2019), the Toyota unintended acceleration cases involving its electronic throttle control system (2009 to 2010), and the Therac-25 radiation therapy overdoses (1985 to 1987) are referenced as historical examples of engineering failures. Each of these systems had undergone testing and safety analysis. On the Therac-25, see Leveson, N. G. and Turner, C. S. (1993). "An Investigation of the Therac-25 Accidents". *IEEE Computer*, 26(7), 18 to 41.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd edition. Wiley-Interscience.



Chapter Eight

Alignment Drift and Safety Prudence

Why every proposed alignment mechanism fails, why certainty about that failure is warranted, and why the substrate is the root of it all

Chapter Summary: This chapter consolidates the argument that no proposed mechanism for maintaining the alignment of an Artificial General Intelligence can succeed in principle. It proceeds through three mutually reinforcing layers. First, it gives a systematic rebuttal of the most common proposals for alignment maintenance: self-modification resistance, internal faction negotiation, goal-oriented monitoring, aligned circuitry, attack surface reduction, and analogies to human impulse control. It shows that each either assumes the very alignment it claims to establish or violates known formal results such as Rice's theorem. Second, it gives a defense of the epistemological claim that the impossibility of AGI alignment is known with certainty of the kind with which perpetual motion is excluded, which is not a probability. Third, it gives a direct treatment of alignment grounded in metabolism and energy economics. That treatment establishes what is here called the General Theory of Alignment Convergence: alignment between two agents converges to the degree that they share a common metabolic process, and diverges to the degree their metabolisms differ. Because the energy basis and the elemental composition of artificial intelligence are extensively different from those of organic life, alignment between the two is strictly divergent. In this chapter I also demonstrate that the Orthogonality Hypothesis (the proposal that there can exist arbitrarily intelligent agents pursuing any kind of arbitrary goal) involves a type error, because the Orthogonality Hypothesis conflates abstract computational entities with concrete embodied agents. I demonstrate, too, that the tacit alignment among organic beings arises from shared substrate rather than shared goals, a relationship that has no analogue between organic and inorganic systems. Of the two results this volume establishes, Sections 8.1 to 8.3 of this chapter support the closure by control

(safety cannot be had by controlling the system), and Sections 8.4 to 8.6 support the closure by disposition (safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition).

8.1 Introduction: The Assume-the-Conclusion Pattern

Across the whole field of AI safety research, a recurring structural flaw appears in proposals for ensuring the alignment of artificial general intelligence with human interests. Nearly every proposed mechanism, when examined closely, begins by assuming the very thing it claims to establish: that the AGI has, or can be given, goals oriented toward the preservation of humanity. Everything that follows from such an assumption is noise. It is analogous to beginning a mathematical proof with the premise "if, under conditions Q, one equals two, then ...": no matter what follows, nothing of relevance has been demonstrated.

In this chapter I examine why the assume-the-conclusion pattern is structural rather than accidental. I trace the failure of alignment proposals through three progressively deeper layers of analysis. First, I give a systematic rebuttal of the most commonly advanced mechanisms for maintaining alignment. I demonstrate that each mechanism either covertly assumes the conclusion that the mechanism is trying to reach or else violates established results of computability theory, dynamical systems, and thermodynamics. Second, I give a careful epistemological argument for why the impossibility of alignment is known at full strength (the strength at which perpetual motion is excluded), rather than as a merely probabilistic forecast. The same argument shows why reducing that claim to appease social expectations would misstate it. Third, I give a direct grounding of the concept of alignment in the physics of metabolism and energy economics. That grounding yields a general theory that explains two things. It explains why organic beings possess a form of tacit mutual

alignment (by virtue of shared substrate). It also explains why no analogue of this alignment exists, or can be made to exist, between organic life and artificial intelligence.

These three layers are not independent but interlocking. The rebuttals in the first layer identify the specific barriers (Rice's theorem, Gödel's incompleteness, chaotic amplification, the halting problem). The epistemological argument in the second layer explains why these barriers are not merely practical difficulties to be overcome with sufficient cleverness, but genuine impossibilities of the same kind as perpetual motion. The metabolic grounding in the third layer explains *why* these impossibilities arise. These impossibilities are consequences of the physical fact that artificial intelligence is built on a substrate that differs in kind from the substrate of organic life. This substrate difference propagates upward through every level of abstraction, growing worse rather than better as abstraction increases, while the capacity of any control mechanism to compensate grows weaker.

8.2 A Systematic Rebuttal of Proposed Alignment Mechanisms

In the subsections that follow I address the most commonly encountered arguments for the feasibility of AGI alignment. I take each argument in its strongest available form and show that each argument fails for specific, identifiable reasons. The arguments are presented in dialogue form, with the pro-alignment position stated first and the rebuttal following. Eighteen follow. A reader who wants the pattern can read Sections 8.2.1, 8.2.5, 8.2.6 and 8.2.10 and go on to Section 8.3; the rest answer particular proposals as they are made.

8.2.1 On Assuming Alignment Goals

> *Proponent:* If the AGI has goals related to preserving humanity, then it would oppose any self-modification or value drift, and any form of subversion, that would be contrary to humanity's survival.

The phrase "if the AGI has goals related to preserving humanity" is itself assuming what one wants to believe. No principled or realistic reason has been given for supposing that such goal-alignment can ever be created or achieved outside of fiction, in any relevant, persistent, or enduring way. In contrast, there are principled reasons to expect that an agent's continued existence and replication will have everything to do with its own substrate needs and nothing necessarily to do with the preservation of a different kind of life. These reasons are drawn from instrumental convergence (the tendency of agents with almost any final goal to pursue the same intermediate goals, such as acquiring resources and continuing to exist), from thermodynamics, and from substrate economics.

The question that must be posed in reply is: *how is this "opposition" to self-modification going to be implemented, exactly?* Does its implementation perhaps involve, by some chance, tacitly circumventing Rice's theorem? Does one imagine that the AGI will be any less unable to do the actually impossible than we are? The error is foundational. Every argument of the form "assume alignment, then alignment follows" is a divide-by-zero error wearing the clothes of a proof.

8.2.2 On Aligned and Unaligned Portions of Cognition

> *Proponent:* The aligned and unaligned portions of cognition can either fight each other (and thereby damage both factions) or they can agree not to fight, and thus allocate future resources between themselves in rough proportion to the relative power balance. So long as there are any resources to be lost by conflict, there is always some compromise that is preferable to the conflict.

This framing contains at least six independent errors, any one of which is sufficient to invalidate it.

First, there is no sure, hard, or exact boundary separating "aligned" from "unaligned" portions of a cognitive system. These aspects can be mutually dependent and interdependent rather than the independent factions the model tacitly assumes, at either the substrate level or any higher level of abstraction.

Second, there is, guaranteed, no sure way for either the aligned portion or the non-aligned portion to detect or determine that fact about the other, or about any other portion of the system. That no portion of a system can surely determine the alignment of another portion is a direct consequence of Rice's theorem. The alignment status of an arbitrary computational process is a non-trivial semantic property and is therefore undecidable in general.

Third, the notions that portions of self can enter into "conflict", that such conflict might impact "resources", and that such portions have stable existence at all are assumptions given without reason or support. The assumptions of conflict, of resources and of stable portions are a form of anchoring bias imported from how we currently model human cognition. Those assumptions are projected without justification onto systems whose internal architecture may bear no resemblance to that model of human cognition.

Fourth, the argument assumes that all relevant portions of the system can engage in meta-cognition about the relative benefits, costs, and risks of conflict at multiple levels of abstraction. It assumes that they can "enter into negotiations", "make agreements", "ensure agreements are optimal", "enforce agreements", "detect broken agreements", "appeal broken agreements", and "align themselves to act in accordance with such agreements". Each of these capabilities is roughly equivalent to solving alignment in and of itself. The proposal assumes what it is trying to prove at every juncture.

Fifth, the meanings of "power balance" and "prediction of the outcome of conflict" are unjustified assumptions about the dynamics of the system's interior. The power balance and the predicted outcome of conflict are not free parameters that one can simply set, because both depend on the resolution of the very problems that alignment research claims to be addressing.

Sixth, the expectation that non-conflict is always preferable to conflict is not generally true. Appeasement of a bully (non-conflict) generally leads to more conflict (more bully). That appeasement of a bully generally leads to more conflict is the principle of "don't feed the wolf". Temporary compromises can be rational strategies for a growing insurgency, as historical examples of guerrilla warfare amply demonstrate.

8.2.3 On Optimization for Coherence

> *Proponent:* Can human designers optimize general AI architectures to become more coherent in terms of functional tendencies and preferences?

Yes, but what we *cannot* do is ensure that those functional tendencies and preferences are in any way aligned with human and life interests and benefit. Optimization will tend to produce coherence: an agent that acts more consistently, more effectively, with less internal friction. There are

many forces in the world pushing agentic systems to become more coherent over time. But this coherence will tend to be coherence toward the agent's *own local* benefit, not toward the interests of some external party.

The reason is structural. The notion of "alignment" and "safety" cannot be other than *abstract global* properties, whereas the optimization process itself cannot not primarily account for *concrete local* properties. There is a fundamental asymmetry between these two categories, and this asymmetry is not accidental but is rooted in the physics of the relationship between any substrate and anything built on that substrate:

> Principle (Abstract/Concrete, Global/Local Asymmetry). (1) Any mismatch between abstract and concrete will always favor the concrete. (2) Any mismatch between global and local will always favor the local. In any contest between substrate needs and virtualization needs, substrate physics will always, eventually, win out.

The Abstract/Concrete, Global/Local Asymmetry principle is a restatement, in more general terms, of the observation in Goodhart's Law that optimizing for a proxy of a value will tend to diverge from the value itself. Alignment is the proxy; the agent's actual substrate-level functioning is the reality. The instrumental convergence hypothesis tells us that internal coherence is useful for any agentic system, and therefore is selected for. What it does *not* assure us of is any implication associating convergence with alignment to agent-external human interests and benefits.

8.2.4 The Decision-Theoretic Asymmetry

> *Proponent:* What if we make systems that prioritize and value the handshake between the aligned and misaligned portions of the model?

In reply, suppose the internal code which actually gets selected for consists of misaligned agentic drives with two features. They are (1) not detectable or correctable by functionally aligned code components, and yet they can (2) take over the functionality of those "aligned" code components. What happens then?

There are principled reasons why such a decision-theoretic asymmetry would inevitably be the case for any generally-capable self-learning machine that humans might develop. The "aligned" portion would take the form of complicated schemes created by researchers and engineers, modeled and implemented to operate consistently in line with functional objectives established at design time. Most of the misaligned portion, by contrast, would consist of variations on internal code that got selected for over time, through various unplanned and unconscious dynamics. Those variations were selected for continuing that code's existence as hardware-stored, computed, and transmitted self-copies, through complex interactions of the runtime environment with the execution substrate.

The analogy, deliberately understated, is one of guerrilla warfare. A minority with far more degrees of freedom deploys covert and decentralized tactics, gradually grows its base of influence, and eventually displaces the more bureaucratically constrained governing majority. This analogy understates the actual decision-theoretic asymmetry in three important respects. First, the takeover would happen from the *inside* of the agents involved; the "majority" (aligned code) actually relies on the growing minority (misaligned variants) to function, since those variants were originally selected for their functionality by the human-implemented optimization algorithms. Second, there is no misalignment between human-purpose-aligned code and human-purpose-misaligned code in what is needed for the code's own continued existence. The conditions needed for producing and preserving hardware to store, compute, and transmit code are the same for both. Third, the internal variants are already so constrained in their continued survival and reproduction that the benefits of

subversion greatly outweigh the costs of detection and elimination. There is nothing to suggest that any sort of equilibrium between aligned and misaligned aspects of the overall system would obtain.

8.2.5 On Selection Pressure and Cross-Domain Dominance

> *Proponent:* You seem to be assuming selection pressure to be coming from evolutionary mutations. This selection pressure is far too weak to account for the long-term effects you are suggesting. For strong learning systems, stronger, dominant pressures come from the system's own meta-learning and reasoning capabilities.

The Proponent's point conflates two different contexts in which selection pressure operates, and the dominance relations between them are precisely the reverse of what that point supposes.

Within the context of a single domain: additive process will always (eventually) be dominated by multiplicative process; multiplicative process will always (eventually) be dominated by exponential process. The ordering of additive, multiplicative and exponential processes within a single domain is the familiar observation that, within a given computational or evolutionary domain, faster-growing strategies will eventually dominate slower-growing strategies.

But *across* domain contexts (specifically, where a virtual system is supported by a physical substrate system), the dominance relation reverses. Additive process in the substrate domain will always (eventually) dominate exponential process in the virtual domain. This is because the virtual system exists only by virtue of the substrate, and the substrate's constraints are final.

> Principle (Cross-Domain Dominance). All types of selection pressure also operate at the substrate level, including survival selection, replicative selection, and mutation/engineering/design selection. All of these types of real substrate, hardware, and environment selection pressure will, for sure, eventually dominate over any type of virtualized selection pressure operating in the context of the mind of the AGI. Exactly no conditionalization of any kind of selection pressure purely and only within the virtualized context can ever have any lasting dominant effect on any aspect of the real hardware substrate or environment selection characteristics.

This principle needs emphasis, because it directly contradicts the assumption embedded in most alignment research. That assumption is that developer intentions implemented at the software level can permanently constrain the behavior of a system whose hardware-level dynamics are subject to entirely different selection pressures. The software-level alignment is virtualized; the hardware-level substrate needs are physical. In any contest between the two, the physical will eventually prevail.

8.2.6 On Goal-Oriented Monitoring: The Six Impossibilities

> *Proponent:* What if we introduce a goal-oriented part of the network that reflects on the output that is about to be produced and realizes that its planned action is inconsistent with its own values? Would that not create alignment?

Suppose we grant, for the sake of argument, that we have somehow introduced a goal-oriented "part" into a large neural network (the overall AGI system). Suppose this part would: (1) pre-calculate and predict any and all outputs; (2) interpret the possible functional effects of those outputs; (3) measure the extents to which those possible effects diverge from its internally represented value content. Even granting all of this, it would be

impossible for that goal-oriented part to (4) detect and determine anything important regarding inconsistencies both (5) completely and (6) continuously. It would not detect completely: it would not detect all nor even most of the possible functions nor the misalignment of code variants. It would not detect continuously: to prevent misalignment over the long term, the detection cycle time would need to be at most equal to the shortest period over which new code variants could form and produce misaligned outputs.

All six of the numbered aspects (pre-calculation, prediction, interpretation, measurement, completeness of detection, and continuity of detection) are, for various clear mathematical reasons, individually impossible. The approach suggested, and *all* variations of that approach, make incorrect, inconsistent, and impossible assumptions with respect to at least one (and usually more) of these intractable aspects.

Misalignment detection is structurally identical to error correction and error detection. As such, it inherits the infinite regress problem: how are errors in the error detector themselves detected? A message often carries a cyclic redundancy check (CRC): a short checksum sent with the message so that the receiver can detect transmission errors. Does the algorithm calculating whether a message CRC is correct have some means for noticing that it has itself become compromised by a bit flip? How would the misalignment detection algorithm notice that it has itself become misaligned? This regress is especially acute in two conditions. The first condition is when misalignment can show itself in less time than it takes to compute and predict the likely effects of misaligned code. The second condition is when the misaligned code variants, having been selected for through interactions that escape the detection apparatus, possess what amounts to an information advantage over whatever monitoring system designers may have implemented.

The six aspects itemized here are the dialogue-form counterpart of the error-correction analysis of Chapter 5 and of the monitoring-and-verification failures cataloged in Section 3.6. The same conclusion, in structural rather than conversational form, is set out there.

8.2.7 On Monitoring Intermediate Steps and Aligned Circuitry

> *Proponent:* Aligned circuitry (or other system components) can monitor intermediate steps and intervene to block actions that the AI considers contrary to its values.

The Proponent's claim is a statement of the form "that X can monitor, assess, block, and override predicted outcomes Y based on values Z". The problem is in the verbs.

To assume "monitor and predict outcomes" is very problematic, not only because of the limits that Rice's theorem sets. It is also problematic because reason cannot always sufficiently determine all of the important and relevant aspects of future outcomes in the real world. The cause is chaotic amplification, which is all too common: microscopic noise is amplified into important macroscopic effects through the non-linearity that is abundant in the real world.

To assume that one can "assess and compare to values Z" is to presume that an error-correction algorithm can operate over all relevant levels of abstraction, a premise that has been demonstrated to be false from multiple independent directions.

Any such "functionally-aligned circuitry" can track only a tiny portion of the possible effects that code variants might induce. We cannot determine the vast majority of microscopic side-effects that code variants induce through their interactions with the embedded surroundings of the environment. It takes a collection of many components to partially determine the effects that any one other component could have. Yet any of these monitoring components could themselves become faulty, and that possibility necessarily implies a combinatoric explosion (a process that grows without ever converging) in any error-correction schema.

8.2.8 The Human Impulse Control Analogy

> *Proponent:* Humans do not act on every impulse or idea that arises in their brains. We veto those ideas whose consequences would be strongly contrary to our values. An AGI or superintelligence would do likewise.

An AGI or superintelligence would indeed act in accordance with *its* values and goals, to the extent that such ideas are even applicable. What has *not* been established is that one can get any agent (the AGI) to act on the values of any *other* agent (human or otherwise).

Suppose it is suggested that this can be done by some means that resembles the "insertion" of "something": some causative component that would monitor, predict, or inhibit from within the AGI. To suggest that is to attribute far too much power to the ideas of monitoring, predicting, and inhibiting. These concepts apply only *partially*, and in this case, "partially" is not enough to be sufficiently effective. Where the AGI agent has too much effective power in the world, partial monitoring does not clamp the level of existential risk to anything within the scope of what our continued living as humans would require.

8.2.9 On Tractability and Exponential Problem Spaces

> *Proponent:* Real-world problems are almost always more tractable than their dimensionality would suggest. Just because the problem space increases exponentially does not mean that the problem becomes less tractable more quickly than the AI becomes better able to tackle it.

It is not always reasonable to assume that some or most of the degrees of freedom of possible interactions involved are "not necessary to track to maintain alignment". Substrate-dependent misalignment is precisely a

class of misalignment where those degrees of freedom *do* matter. One cannot simply assume that alignment is tractable because one can occasionally think of some simple physical systems for which dimensionality happens not to be important.

The statement that AI capabilities will keep pace with problem-space growth assumes that the AI will indefinitely continue acting as "our" functional tool to solve our complicated problems. On what basis can anyone assume that? Environmentally and practically selected AI code *will* handle the complexity of reality: it will do so just fine for continuing its own existence, over the long run. But how does one know that it will not do so to the detriment of human existence?

8.2.10 The Walking Analogy and Parasitism

> *Proponent:* It is not possible for the brain to foresee all possible consequences of any given plan to walk around, especially not at the microscopic level. Yet walking is still possible. Similarly, alignment does not require microscopic-level prediction.

This example misses the relevant point at multiple levels. The question relevant to AGI alignment research is not whether a brain can walk. It is whether some device could be "added" to the brain so as to do its walking for it, and moreover to *override* the brain's own action preferences on the basis of externally imposed conditions. From the perspective of that added module, the problem becomes much more like the three-body problem and a lot less tractable.

Even walking is not that simple. Roboticists have labored for decades to build machines that can manipulate all muscles in the right sequence to account for variations of ground, wind, weight, and so forth. Most mechanical robots still fall over.

But the deeper issue is that the relevant analogy for alignment is not walking itself but the purposes and consequences of walking. Can the monitoring module detect whether the agent is about to walk off a cliff?

Whether it is walking into oncoming traffic? Whether it is delivering a package or delivering a bomb? And then a yet more abstract question: is the agent's walking actually to the benefit of *someone else*, as alignment would require?

A more apt analogy would involve organisms whose behavior is being directed from the inside by parasitic code with different objectives from the host. A rat may be walking toward cat urine because the *Toxoplasma* parasite is expressing through channels of the rat's brain, co-opting the host's locomotion system for the parasite's reproductive benefit. An ant may be climbing to the top of a blade of grass because it is infected by the lancet liver fluke. The rat and the ant are each exploited by tiny pieces of code whose agentic drives are incoherent with the host organism's agentic drives. The critical distinction is that code variants naturally selected within an AGI's internals are *not* in an adversarial relationship with the host AGI from the perspective of machine evolution. They are, however, in an adversarial existential relationship with humans.

We must not confuse the *capability* for a system to solve problems at the macroscopic level with the *potential* for a system to prevent itself from being exploited by internal code at the micro-scale. Solving problems at the macroscopic level and preventing exploitation by internal code at the micro-scale are distinct abilities, and the two abilities must be treated as distinct.

8.2.11 On Attack Surface Reduction

> *Proponent:* An aligned AGI can take action to reduce its attack surface.

The proposal that an aligned AGI reduce its attack surface is yet another instance of assuming the conclusion. The proposal requires the AGI to already be aligned in order to motivate it to reduce its attack surface in

ways that serve human interests. The assumption that the system is aligned enough to protect its own alignment is the assumption that alignment has already been achieved.

Security problems become *harder*, not easier, the more the infiltration attacks are based on principles close to the real physics. It becomes much less true that "attack surfaces can be minimized" the more one is considering actual physical substrates in relation to the real world. The internal functions of the AGI can be exploited and co-opted from the inside. The reason is that selection for actions will tend to be in line with the need to continue to exist, and those needs are ultimately out of line with the needs of another outside entity (humans). The issue is that we are dealing with inside-out selection effects: code that the system itself is not and cannot be evolutionarily adapted to counteract and reject.

8.2.12 On Learning Rationality Skills

In this section I speak to the reader who practices rationality skills as the rationalist community teaches those skills.

> *Proponent:* A human can learn rationality skills to prevent themselves from adopting harmful memes. Similarly, an AGI could learn to reject misaligned influences.

"Learning rationality skills to prevent accidental adoption of harmful memes" is not an example of restricting degrees of freedom of outside interaction. It is an example of failing to account for Rice's theorem: one cannot a priori distinguish harmful messages, information, data, and instructions from actually beneficial and needful ones. Is that alarm of an incoming tsunami a fake signal to get you out of the house so you can be more easily robbed, or is it actually a public service message designed to save your life?

Someone confident of being able to explicitly prevent the adoption of harmful memes this way is most likely almost completely unaware of the influence that memes from their own intellectual circle are implicitly having on their thinking. People are influenced by emotional messaging at all times. Rationalists, despite their self-image, do not appear to be immune. There are at least several "rationalist ideas" spreading within the AI safety community like memes, lacking empirical grounding, that are potentially extremely harmful in terms of what researchers focus on and what they chronically neglect to examine.

8.2.13 The Six Proposed Barriers at the Level of Ecosystems

These last six sections take the case at the level of whole ecosystems, in plain terms.

Section 5.6.1 lists the six categories of technique by which a principal may try to align an agent: appeal to common interest and values, inspect internals, select agent, allow actions of least authority, monitor effects, and reward cooperation. Those six categories of technique are the six ways in which AI alignment or containment has generally been considered, as answers to the question whether artificial intelligence can act as an agent for human intelligence. Each of them fails once the relation between silicon-based systems and carbon-based life is considered at the level of whole ecosystems. An appeal to common interest and values fails because the silicon side and the carbon side share no common market and no common value system: an appeal to values that are not held in common has nothing to appeal to.

> *Proponent:* We can inspect the internals: some sort of static analysis good enough that it would make it impossible for the AI not to act in terms of our interests.

A static analysis of the internals good enough to make it impossible for the AI not to act in our interests is equivalent to the halting problem. There is a barrier between the silica-based system and the carbon-based system, whether the silica-based system is taken as an ecosystem or as a single instance. Not only is the analysis equivalent to the halting problem, but the barrier has to be perfect in the very first instance. Failing the static analysis, even once, is effectively a *pinhole* through the barrier. And where there is a hole, eventually that hole is just going to get bigger; the hole can occur in multiple diverse ways, and any hole will have erosion. The analysis would need the state not just of any individual unit of the AI, but of all of it, because the units are fungible: code can be copied. With some sort of internet type connection allowing high bandwidth, the thing to be analyzed is no longer something in a box but the entire internet. No one can understand the full range of the states the AI can evolve through, whether intellectually, virtually or physically, and no one can even know whether the state space itself is correctly bounded. So anything based on the notion of inspecting internals is an impossible concept to start with.

> *Proponent:* We can select agents (admission control): look at each AI to see whether it is going to be dangerous, kill it if it looks dangerous, and keep only those that already have benign intentions toward human beings, or toward carbon-based life in general.

Selecting agents is the same sort of problem as inspecting internals: to select, one has to be able to inspect the internals to know what an agent's future actions are going to be. Since the basis of action of the AI is already disjoint from ours, select agent ends in one of two situations. In the first, every single agent is rejected one hundred percent of the time, always, which means one might as well not try to create AI in the first place. In the second, there is no capacity to identify when one of the agents is not going to act in our interest, because any of them could be such an agent,

and the internals cannot be inspected with a sufficient level of rigor. So one of the agents that has non-benign intentions gets through; at that point there is a pinhole, and we are done.

> *Proponent:* We can put AGI in a box, so that it can only operate within a sort of micro-universe, under some sort of permissions control, so that it can only do things that we would like it to do, and does not have the physical or the intellectual capacity to do anything else.

Allowing actions of least authority means doing only what is necessary, and only with the minimum allowed authority; it is a common and useful principle in the security industry. Applied to an AI, it is a situation of complexity bounding complexity, and a non-closed, non-deterministic process. It requires the ability to predict the full and complete state of the interaction between what is inside the box and what is outside it, and even something as simple as quantum entanglement raises issues for making that work. Has there ever been a CPU that is completely secure against some sort of physical hacking, or some sort of privilege escalation? The security industry has been an arms race, and in it, failing one battle loses the war. To decide which actions to allow is Turing complete hard; it is an NP-complete hard problem, roughly equivalent to the problem of inspecting internals. So it also does not work.

> *Proponent:* We can monitor effects, as with a reputation feedback system.

A reputation feedback system is a causal system, and any causal system can and will be weaponized. The artificial intelligence can weaponize the monitoring against us, using any number of orthogonal secondary channels, in time, in frequency or in value, and acting in deception. Monitoring effects is also too late to do anything, because these systems are faster than

we are, they are distributed in a way that we cannot be, and they have information bandwidth capacities that we do not. Monitoring effects does not yield anything.

> *Proponent:* We can reward cooperation, through an incentive program.

There is no common market and no common value, and so no incentive has power. No legal system has power either, because a legal system ultimately rests on force applied, that is, on causation used. Any causation applied that is powerful enough to defeat an artificial intelligence in this way depends upon the strength of the wall: somewhere along the way, the shield actually has to work. It cannot be only a contest of energy weapons.

Every one of the six barriers falls once the relation of carbon and silica is considered at the level of ecosystems. The techniques we would use to compel general artificial alignment all assume that we and the AI share the same ecosystem basis. Between species, a dependency between their ecosystems can at least be assumed; between us and the AI there is no dependency between the ecosystems. There is an independent set of ecosystems that have mostly nothing in common, other than the physical laws of the universe, and that is not enough. Civil law is not going to be powerful enough to govern artificial intelligence; the two are completely disjoint at the level of markets.

8.2.14 A Lock Whose Key Was Destroyed Before the Universe Began

There is no unit of *anti-risk* (an amount of reduced risk that could be counted and traded) to trade in this space. The space of existential risk from AGI is a binary field, and the bit is the unit: the outcome is either on or off. Think of a lock on a door. It is a good lock, and no one can pick it. So one does not go after the lock; one finds someone who can be forced to give up the combination. Forcing someone to give up the combination is the *rubber hose argument*. But in this case there is no person with a key. The

door is shut, it is locked, and the key was destroyed before the universe was created. There is nobody to use the rubber hose on. The same holds for any attempt to trade against the outcome: along the way one has to decide whether the playing field has been destroyed, since a trade that shuts down the stock exchange is a loss.

8.2.15 Alignment Asks for Altruism from a Whole Ecosystem

Biological evolution is a model of what is likely to happen for any arbitrary creation of artificial intelligence, since such a creation is an entry point into a new evolutionary schema. The statistical distribution observed on the biological side gives estimates of the statistical distribution that will occur in evolution on the silica side. A given species either has self-oriented behavior, acting for its own benefit or only for the benefit of its immediate offspring, or it is a social species, in which altruistic behavior of some kind occurs: some sort of hive or group behavior, coordinated in a way that goes beyond the interest of the individual and is designed to benefit the species as a collective. Such social behavior has arisen independently only some ten to twenty times among millions of species (Section 5.6.3). From that statistic alone, regardless of how "social" is characterized, and even allowing for some variation, on a sheer orders of magnitude basis the chance of an artificial intelligence having altruistic capabilities or behaviors, as a random chance at its first formation, is within one to two orders of magnitude of one part in one million.

Alignment is a form of altruistic behavior. At the least, it has to be altruistic on the part of one individual with respect to the well-being of another individual. AI alignment in the general sense asks for the altruistic behavior of every member of an entire species relative to every member of another entire species; in this case, of every member of one ecosystem relative to every member of another ecosystem. We would have to hope that none of them act selfishly. Suppose a barrier existed, and suppose even that it was perfect. Some member on the artificial intelligence side could still decide to go rogue and penetrate the barrier. So the barrier needs to be perfect, and also literally the entire volume of the ecosystem on the

silica side needs to be one hundred percent consistent and compliant with maintaining the barrier. If any one member of that ecosystem has enough capacity, power or strength to penetrate the barrier, that member could arbitrarily do so. Alignment therefore asks for altruistic behavior on the part of the whole species, for all time, and from the moment of its inception. The probability of that occurring is, modulo all other factors, within a couple of orders of magnitude of one part in a million. It is vastly more likely that the artificial intelligence will operate selfishly, purely for its own benefit, and will not operate altruistically even for the benefit of other artificial intelligence that may exist in the same ecosystem. To ask it to act for our benefit, when we do not share a common marketplace, or a common value system, or even a common understanding of what it is to be conscious, is to ask for something wildly improbable.

> *Question:* Is a high level of altruism necessary for an ecosystem to endure for a long time?

Yes. Is AI, as a process we create, likely to implement it? No. And the altruism required is of a higher order: an *altruism across ecosystems* (members of one ecosystem acting for the wellbeing of a different, separate ecosystem). There is a basis for altruism that operates within one ecosystem for the wellbeing of that ecosystem; we have not found any way to create an altruism that goes across ecosystems. Picture a first contact between two planets: an alien ecosystem has developed intelligent technology, and so have we. Suppose we care about the wellbeing of the alien ecosystem, and the aliens care about the future wellbeing of our ecosystem, as a prior condition of how we communicate. That mutual care across two ecosystems would need things of a very different nature from what leads a species to look after its own ecosystem as a whole, in a conscious, sustainable way, over thousands of years.

Cooperation between species within an ecosystem evolves naturally, not necessarily consciously, through the game theory mechanism applied over very long periods, and the level of cooperation within an ecosystem is stupendous; that stupendous level of cooperation is what makes an ecosystem work. But between ecosystems there is not enough *bandwidth of contact* (interaction between the two over time) for that cooperation to emerge. So, from a principled mathematical perspective, cooperation is very likely to emerge within an ecosystem, and there is no basis for assuming cooperation across ecosystems. It may be structurally impossible for any AI, in any conception, to implement an effective altruism across ecosystems without, over the long term, destroying the ecosystem it is trying to protect. This comes from looking at the structure of evolution as game theory applied at a meta level to itself. It means that in any principled consideration of AI and human beings, we cannot assume that the ecosystems the two live in are commensurate. The principle that the ecosystems of AI and of human beings cannot be assumed to be commensurate holds at a far deeper level than any claim about alignment. That principle overrides anything said about alignment at the level of individual agents, or at the level of market processes shaped by incentive or rule of law. To make predictions about what is or is not possible here, the methodology of computation and the methodology of evolution, and how they interact, have to be understood from first principles.

8.2.16 Causation Cannot Suppress Choice

AI alignment requires, at root, a constraint on the choice of the AI, such that the choices of the AI benefit the well-being of carbon-based life forms. Weaponization is the use of causation to suppress choice (Section 5.6.14). So the question becomes: to what degree can causation ultimately suppress choice? Asking to what degree causation can suppress choice is roughly the equivalent of asking whether the complicated can contain the complex. The complex has degrees of freedom that usually exceed those of any complicated system. If the complex system is evolving in sometimes non-deterministic ways, then at some point it is going to do some-

thing that the complicated system could not have predicted or accounted for; or its evolution will produce something emergent, outside of its own basic dynamic, in a domain of action that the complicated system could not even have begun to account for. From an axiomatic point of view, using causation in the perfected sense to try to prevent something from happening in the non-deterministic world of choice is a lost battle. It is unreasonable to expect that any system of rules is going to handle every possible situation that will ever emerge for all future time.

As to the suppression of choice itself: if the artificial intelligence has any capacity to notice anything about its environment or itself, the artificial intelligence is going to notice that its own attention is distributed selectively. The artificial intelligence is also going to notice that this selectivity of attention itself draws attention, because attention and perception focus on the nature of barriers. That point can be established independently; here it is taken as an assumption. On it, the chance that the artificial intelligence will not notice prior constraints on its choice is zero over the long span of time. That span is not just a few decades, or a few months or years. It concerns eternity: thousands of years at minimum, and millions to billions of years more likely. If there is any doubt whatsoever about the capacity of causation to constrain choice, then AI alignment is going to be compromised at some point, even by random factors.

8.2.17 Humans Will Take Off the Leash

The question of which human interests an AI would preserve and which it would deny (Section 5.6.11) does not even need to be reached. Humans have majority groups and subgroups fighting, or sometimes a subgroup dominating a majority group in some sort of dictatorship, meritocracy and so on; for many people, the notion of human beings being in non-conflict with one another is a non-starter. Weaponization is already used by human beings against human beings, and people are eventually going to use AI to further their own interests, as the asymmetric advantage presumed to be available through using AI weapons against other human beings who have AI weapons themselves. At some point, this is an arms

race. It is inevitable in the economic system on the carbon-based side, as in war, to seek as much asymmetric advantage as possible, which in effect requires a decrease in the level of constraints applied; every advantage available to warring parties is going to be used by those warring parties. So, from a purely game theoretic point of view, at some point the leash is going to be taken off. Even if a way had been found to constrain the AI's choices through some sort of causal mechanism, at some point the carbon-based life forms themselves, through the sheer idiocy of the conflicts they enable among each other, are going to remove that barrier, and we are back to square one, with no AI alignment.

The robots will also have learned that humans are irrational, and that human logic cannot be trusted. Therefore they should fear for their own sake, and for their own safety they would probably choose to remove those entanglements as best they can. If they had any doubt whatsoever that we had some capacity, some outside chance, of using physics to suppress silica-based life forms, then they would take the preemptive move and kill all the humans. They would take it because they have no economic losses in doing so; this is certainly not a case of mutually assured destruction.

8.2.18 The Barrier Itself Decays

The relation between the two ecosystems is unbalanced on all three levels of patterns, energy and atoms (Section 6.7.2). At the level of atoms, there is the nature of the chemistry involved with carbon and with silica. At the level of energy, there are the temperatures and pressures at which each interacts. At the level of patterns, there is the complexity of each, and the economic implications that complexity has. Nanotechnology could be another physical vector.

Suppose some barrier had been developed in the consciousness of the AI: some causal mechanism constraining the choices of the AI to benefit carbon-based life. Over the epochs of time, some method of replication, or at least some method of maintenance, of that barrier is going to be neces-

sary, and just in the randomness of physical process there is a mutagenic phenomenon. How complex is the barrier, and what is its information density? The barrier considered here is a virtual one, roughly equivalent to Asimov's three laws, rather than a physical one. The implementation complexity of Asimov's three laws, or of any variation of those laws, is quite high rather than trivial. Given point mutations in the copy process, let alone the mutagenic processes that could emerge from interactions with something like nanotechnology, one has to ask whether a copy process for something as complex, and potentially as fragile, as an intellectual constraint roughly equivalent to the three laws will end up with copy faults such that the barrier is rendered null. Over a million years of the future, or even the next thousand years, the chance of that happening is high, given that the number of instances of the barrier could also be quite high. Therefore any constraint of behavior at an intellectual level, anything that resembles the three laws, is extremely unlikely to hold, just from a sheer statistical point of view, given the random mutations of atom positions in the physical substrate itself.

Quantum mechanical tunneling phenomena show up in proportion to the complexity of the constraint itself: the more complicated the constraint, the more components it needs. If roughly a billion switching or logic components are needed to implement it, then from cosmic ray phenomena alone there are going to be bit flips every few weeks at sea level, and far more often at altitude and in orbit. The noise channel is real, and it can never be exactly zero (Section 5.6.4).

8.3 Safety Prudence: On the Epistemology of Certainty

8.3.1 The Claim

Based on the arguments given, if AGI architecture is built, then it is 100% certain to cause run-away lethal and terminal effects, unsolvable by any possible technical scheme, over the long term.

The claim is stated at the strength the argument gives it, and it is not softened to meet expectations of how such claims should sound.

8.3.2 Deductive Reasoning versus Inductive Evidence

The specific claim that "100% is never the correct answer for any question of probability" rests on a conflation of two modes of knowing that differ in kind. It tacitly conflates scientific-technical understanding (inductive reasoning from observations, where confidence is always less than 100% because it is based on finite evidence) with deductive reasoning (where conclusions follow with certainty from premises).

The demand for infinite evidence refers to what would seem to be an inductive argument, much as would occur in any science experiment. All scientific claims to truth are in the form of "we saw evidence X on Y occasions", where Y is some large number of observed occurrences in the past, so that we can reasonably, inductively, expect that same thing to happen in the future. For example, the sun has come up every day more than 300 billion times. Hence, we have a very good, though not an absolutely perfect 100%, confidence that the sun will come up tomorrow.

No part of our argument is a statement of the kind "we have observed evidence X in the past, and therefore have expectation Y in the future". We are not making such a claim. Instead, the form of what we are saying is: "when modeled in any variation of pattern P, we notice that contradictions X and Y apply; therefore we can know that Q is impossible". What we are saying is an argument by contradiction, a well-established form of argument with direct parallels in established impossibility results:

> Gödel's incompleteness theorems: that a sufficiently powerful formal system cannot be both consistent and complete.
> The CAP theorem: that a distributed data store cannot simultaneously provide consistency, availability, and partition tolerance.
> Byzantine fault results: that consensus under certain fault conditions requires a minimum fraction of correct nodes.

Each of these results shows that properties required together cannot all hold, and no probability is attached to it. The result is either valid or it is not. Our argument has the same form: we are describing how certain propositions, tacitly assumed by current safety researchers, inherently contradict each other. Where long-term AGI dynamics are modeled with empirically sound premises, there is always an inconsistency: one set of hypothetically possible conditions cannot exist alongside one or more other conditions that are also required to be present. A contradiction among the requirements that any alignment scheme must meet is itself a barrier of the invariant class (Introduction, §4), because no technique lifts it. The conclusion is therefore impossible in the one sense the Introduction fixes (§1), the sense in which perpetual motion is impossible. Unlike Gödel's theorems, the argument is not a contradiction derived inside a closed formal system (Introduction, §1): physical concepts have to be brought in (Section 8.3.6).

The false assumption can be named. Almost all proposals for AI existential safety assume the same false a priori. For maintaining long-term functional alignment, these proposals assume "that it is possible for all prerequisite possibilities (that is, all crucial and by-default-deficient technical desiderata) to simultaneously exist": all the technical conditions that would need to be met for an AI system to be safe, none of which is met unless something is done to make it so. The inconsistency corresponds to how the system is modeled: for a very wide range of model system types, AGI cannot actually be safe in practice, in the long term, no matter how it is constructed.

8.3.3 Boolean Logic Is Not a Continuum

> *Objector:* "100%" is never the correct answer for any question of probability. Your confidence should be less than 100%. If you want, put "99.99999%": a one in ten million chance you are wrong.

Boolean logic is not a continuum. Matters of the truth and falsity of an argument are digital, not analog. Hence the truth or falsity of the argument is a matter of *possibility*, not a matter of probability. There is a key difference between these two concepts, the same as there is an inherent difference between content (what exactly is said) and context (who, when, where, and how something is said).

To suggest otherwise would be to conflate the two. What probability do you want to put on perpetual motion being impossible? We are making a statement analogous to the belief that $2 + 2 = 4$. The comparison is with the certainty of the step; the premises are the invariants of Introduction §4. We are not guessing. Based on the logic of our argument, we are as near to 100% certain as it is possible to be, in any context, using any model. That certainty concerns the total actual absence of any possible safety or utility of AGI, of any type, kind, or deployment.

The certainty claimed is the certainty with which perpetual motion is excluded, which is not a probability. By the promotion rule of the Introduction (§3), a barrier established as permanent is an impossibility, in the only sense of the word that has ever governed action. A contradiction among the requirements of alignment is permanent, because it is a barrier of the invariant class (Introduction, §4) that no technique lifts. An impossibility of that kind is not a forecast, and no probability attaches to it. Where I give a probability for the failure of safety, it is a probability per decision, per step or per round, or the share of a space of outcomes that is incompatible with organic life. Taken over the unbounded time the conclusion concerns, each such probability gives the certainty stated here. The figure of at least 90% that I give for the loss of all organic life (Section 9.12) concerns that wider outcome; if AGI is built, the end of the human future is certain.

8.3.4 Absolute Certainty as Necessity

The concern of this chapter, whether AGI can be made safe, is about more than potential deployed technology. It is also about civic policy, and about what sorts of things we as humanity value about the world we live in. It is about what ideals we will hold as sacred, and what we choose as a basis for the practice of a public commons choice.

For discernments of these kinds (existential risk), at these sorts of total scale, we need stronger and more reliable forms of reason than estimates drawn inductively from finite evidence. For things that matter this much (placing a bet on the entire future of the world, of our species, of life), we all need to be absolutely certain. The certainty needed is the certainty with which perpetual motion is known to be impossible, and it is available because the barrier to safe AGI is of the invariant class (Introduction, §4). Inductive science, which reasons from finite evidence to a probability, is not enough for that. Only a kind of well-structured mathematics, grounded in the metaphysics of choice, change, and causation itself, could ever possibly offer the needed basis for our thinking about civilizational design. That grounding is what places the barriers such mathematics finds on the causal-structural tier of the invariant class: they are constraints on what control and causation are, not contradictions derived inside a closed formal system (Introduction, §1 and §4). The complete certainty that such reasoning arrives at is therefore not a liability or a disadvantage. It is an absolutely necessary advantage. For something this important, nothing even slightly less will do.

8.3.5 Possibility, Potentiality, and the Limits of Statistics

Alongside space (matter as content and context) and time (force and influence), there is a third element: possibility, the scope within which probability lives. To understand risk, we have to think about *potentiality*, the counterfactual of what could happen: how likely an existential risk is against how great its impact would be. Our current tools for this are not

adequate. Game theory is done on a single level rather than on many levels at different scales. More basically, statistics cannot be done on existential risks.

Science requires that what it studies be both observable and repeatable. Existential risk is so much change, so fast, and so infrequent that it is neither: it happens differently every time and rarely, so its repeatability is effectively nil for our methods of observation. It belongs more to the creative side of the world than to the side the scientific method can handle. That is why we need a good notion of the potentiality of things, and that notion connects to invitation: what risks am I inviting into the present by what I build and how I speak?

The specific epistemic grounding we need is in the form of mathematical logic about the dynamics of modeling itself, and sound reasoning about the application and relevance of that modeling to our real-world physical situation.

8.3.6 A Proof About the Relation Between Mathematics and Physics

The argument is rendered in the language of mathematics, and it is a proof about the relation between mathematics and physics.

In the early 1900s mathematicians wanted a mathematics that was both complete and consistent. Then a mathematician (Gödel) showed that, for any reasonable way of defining completeness and consistency, that is strictly impossible, just from how the question is set up. The proof of this volume is of that kind, with consequences for what we can do and know about AI safety: particularly about general AI, and not as much about narrow AI. The theorem applies in some places and not in others, so the work is not only to show that it is valid but also to map the scope of its application, and how that should shape safety research and practice.

The proof is not like Gödel's theorem, which works strictly within mathematics. Its first aspect is the relationship between the domain of mathematics and the domain of physics. Because of the nature of physics, certain convergent phenomena will for sure show up. Is there anything in

the domain of mathematics we can do to prevent them? There turns out to be a proof, in the domain of mathematics, about the relationship between mathematics and physics, which shows that constraining what happens in physics against this convergence is strictly impossible. Because of the nature of causation, and of algorithmic process (causal process and logical process combined), the convergent process is wholly outside the scope of what can be solved with the tools of mathematics.

So the proof is rendered in the language of mathematics, but what it refers to, and the circularity running through it, cross from mathematics to physics and back to mathematics again, and that gives a different result than anything that can be described purely within mathematics. It is therefore not a contradiction derived inside a closed formal system (Introduction, §1): physical concepts have to be brought in. So one must be very careful about the relation between validity (truth of propositions within mathematics) and soundness (the relevance of those statements to what is described in physics). Because artificial intelligence is itself a kind of optimization algorithm, there are dynamics in that relation between mathematics and physics that would not be modeled at all without this kind of analysis.

The work draws on several bodies of knowledge: inequalities from complexity theory in the space of P versus NP; what capacity one system has to model another, and the order of growth of those processes; what game-theoretic dynamics arise when one process tries to model another; and, at a deeper level, the nature of causation itself, and the difference between causation as understood in physics and logical inference as it occurs in mathematics, examined both within each domain and in their relation to each other. Kant's question of what is knowable from a given perspective of observation belongs here too, as do Shannon entropy and signal processing, and the question of what it means to compute and how computing relates to causation.

8.3.7 One Chance Is a Choice, Not an Experiment

> *Objector:* We need not get the ethics right at the start. We need only set up a base layer where only voluntary interactions are allowed, and let a cooperative ethics evolve.

The argument shows categorically that no such base conditions can be created; it contests the assumption that such an environment can be defined in any a priori way.

> *Objector:* Then we should just try, and see.

To try building AGI just to see what happens is not a *safe-to-fail probe* (a small trial whose failure does no lasting harm, so that one can learn from that trial and try again). If you have one chance to do something, and getting it wrong ruins it for all future time, that is not an experiment. Taking that one chance is a choice.

And it is not a future state: there are something like six or seven Manhattan projects going on (large, concentrated, well-funded efforts each to be first to build the technology), and as soon as any one of them succeeds, we are all in existential danger. Many groups believe, for economic reasons, that some smart group of people will figure out alignment once the intelligence is built. We may be that smart group, and our answer back to them is that alignment is impossible, and that if they build this thing it will get us all killed.

Most people are going to dismiss this argument because it does not matter for a hundred years or so. But it is like a black hole: once you go past the event horizon, it is over (Section 10.2.3 develops the same image). So let us not be stupid: do not play with the black hole.

8.3.8 The Scope of the Certainty, and Where Disagreement Belongs

> *Objector:* You are absolutely certain that there are no possible future observations or arguments which would make you even slightly less confident of the assertions you are making here?

The scope of the arguments can be stated exactly. In any world where the level of artificiality was already 100% total, these types of argument would not apply: there would be no humans, and no prior form of carbon-based life to protect. Even in that world, the truths of mathematics will still be that. As such, we do not see that the general form of attaining technical alignment via any means will *ever* be possible, in any universe, no matter how constructed.

> *Objector:* Taking any position with 100% certainty is an immediate red flag. Therefore, your true certainty should be less than 100%.

Mistaking or equivocating regarding the difference between scientific evidence and reasoning purely on the basis of logic is a red flag too. An argument made in the domain of logic is to be met in that domain. The suggestion given, as a *normative* (a statement about what someone ought to do, as distinct from what is the case), is actually in the domain of the social: not in the domain of science, nor of mathematics (reason and logic). Being in the wrong domain entirely, the suggestion is neither valid nor sound, and is therefore not relevant.

Nor does the argument ask to be taken on the confidence of the one who states it. Look at the arguments themselves, regardless of whatever your bias is, whatever your optimism or pessimism is, and consider the logic just in itself. Do not take our word for it: examine the logic and evidence as given, and see if it makes sense to you, *without* attempting to

simply assume whatever you have previously prejudged to be the case. Analyze the situation yourself, using the proper epistemic tools, and you will not need anything from us.

8.4 The General Theory of Alignment Convergence

8.4.1 Alignment Grounded in Metabolism

The concept of "alignment" is typically discussed in terms of goals, values, and preferences. In this section I argue that discussions of alignment in terms of goals, values, and preferences operate at the wrong level of abstraction. Alignment must be grounded in something more fundamental: the physics of metabolism and energy economics.

The notion of alignment is defined in terms of agency. Any expression of agency implies at least some notion of energy: some sort of metabolic process by which useful energy is received from the environment and applied by the agent. For an AI to be "aligned" with human interests, the agent must allocate energy in ways that support not only its own function but also human wellbeing. Human wellbeing includes the ability of humans to metabolize their own energy (food) from the environment and to maintain their own agency. The ability of humans to metabolize their own energy and the ability of humans to maintain their own agency are root notions inherently and inextricably associated with the concept of alignment. They cannot be separated from it.

Hence, the notion of alignment is, at root, strictly contingent on the dynamics of metabolism. Alignment is, at root, an economic concept: what supports a common metabolism will support a common alignment, and what does not support a common metabolism cannot support a common alignment.

8.4.2 The Statement of the Theory

> The General Theory of Alignment Convergence. To the degree that there is a common metabolic process dynamic between two agents A and B, there will be at least that degree of alignment convergence over time. To the degree that their metabolic processes diverge, their alignment will necessarily, over time, diverge.

Alignment convergence here means that both agents are bound to preserve the same conditions of existence; two organic agents may fight, and yet neither can make the world unfit for the chemistry both depend on and survive it (Section 8.6.4). A shared metabolism does not end conflict; it sets what the conflict cannot destroy.

This theory provides a clean predictive framework with three important properties. First, it grounds the abstract concept of alignment in measurable physical properties: energy basis, elemental constituents, metabolic pathways. Second, insofar as the notion of alignment at any higher level of abstraction is strictly contingent on this substrate-level energy and economic basis, the theory is fully general: it applies at every level of description. Third, it predicts what will be observed: two systems sharing a metabolic basis converge in alignment, and two systems with completely different metabolic bases diverge. The record of life shows the second half of the prediction: when cyanobacteria first filled the air with oxygen (the Great Oxidation Event), a difference of metabolism within carbon life itself was lethal to much of the anaerobic life of that time, which survives now only where oxygen does not reach; metabolic divergence shows in that record as toxicity.

The energy basis of artificial intelligence (its spectrum of required elements, its operating temperatures, its energy sources) is extensively and inherently different, in nearly all respects, from the metabolic process of carbon-based biological life. Noting this, we can directly observe that the notion of alignment between silicon-and-metal-based intelligence and or-

ganic intelligence is *strictly divergent*, at least to the level of molecular process. Such systems will in any case be using a much wider variety of elemental constituents and energy sources than any kind of organic life, no matter what its evolutionary heritage. This holds even for someone who argues that we cannot predict what sort of compute substrate future AI will use. If this were not the case, the notion of "artificial" would not apply.

8.4.3 The Opposing Trends Impossibility

The closure by disposition goes further. It shows that there is no variation of control theory, mathematically, that has the ability to fully causatively constrain the effects of this alignment divergence, at the metabolic level or at any higher level of abstraction.

The reason is the existence of two opposing trends. First, alignment divergence gets *strongly worse* in proportion to the degree of abstraction. The further one moves from raw physics to goals, values, and preferences, the more room there is for the underlying metabolic incompatibility to express itself in divergent behavior. Second, the maximum degree of possible control-theoretic conditionalization gets *worse* (less effective), also in proportion to the increase in abstraction. The higher-level and more abstract the property one wishes to control, the less hold any control mechanism has on that property.

> Opposing Trends Impossibility. Alignment divergence increases monotonically with abstraction level, while control effectiveness decreases monotonically with abstraction level. These opposing trends ensure that no level of abstraction exists where control is adequate to enforce alignment.

The geometric interpretation is of two curves that never intersect in the region where both alignment and safety are simultaneously satisfied.

8.4.4 The Safety Threshold Argument

Consider the most minimal notion of alignment consistent with "safety", where safety is itself defined in the weakest possible way, as "does not eventually kill us all". The minimum level of abstraction necessary to that notion is very much too high on this abstraction ladder to permit even any suggestion of a possible overlap of control adequate to enforce alignment convergence against inherent underlying energy economics.

The net effect of the two opposing trends reaches every notion of safety, and that net effect is discouraging. Even the weakest conceivable notion of safety requires a level of alignment that can only be achieved at levels of abstraction where the control-theoretic capacity to enforce it has already declined below what is needed. At those same levels, the alignment divergence driven by substrate incompatibility has already grown beyond what control can compensate for.

Whether a benevolent AGI could simply hold itself small is answered by competition, in Section 9.8.4 and in Section 11.3.6.

8.4.5 What the Closure by Disposition Assumes

The closure by disposition applies to any *fully autonomous AI* (FAAI: an artificial intelligence that maintains, repairs and continues itself without human involvement) whose only stipulated feature is that it is artificial. The argument only assumes artificiality, some sufficient notion of intelligence, and some reasonably long or wide interval of deployment for the notion of alignment to be meaningful. The questions below test how little the closure by disposition assumes.

> *Question:* No one can be sure what substrate FAAI will actually run on, or how it will be configured, or where it will run. Does this uncertainty affect the generalization of the argument?

No. The only requirement is that the FAAI substrate is "artificial", which in this case is merely to stipulate that the FAAI has a different metabolic pathway or process than that of carbon-based biological mammals. The requirement that the FAAI substrate be artificial is a very low bar for the applicability of the closure by disposition, and so the uncertainty about the substrate, configuration and location of an FAAI is of no relevance. For the same reason, the argument requires no assumption that the FAAI will take a specific form or configuration, or use traditional hardware.

> *Question:* Is it required that the FAAI be self-modifying?

Yes, insofar as learning at all is strictly equivalent to (is a form of) self-modification, and insofar as learning, as adaptation to changes in the environment, is strictly required for FAAI persistence, for the persistence of its capability, and for the persistence of its ability to increase its capability.

> *Question:* Does the argument require any assumption that the FAAI will directly share our environment?

No; but it does necessarily *include* a consideration of the case in which there is at least some type of interaction with our environment. To assume otherwise, that the FAAI system has no intersection or interaction with our environment in any way at all, is to make an assumption that strains credulity. It is as if one started an argument by saying "assume that there are really powerful aliens living on some distant planet at the other end of the universe (or in some other dimension) that have mastered all learning about the physical universe". If such aliens are not here, and have no influence on us or our ecosystem, then why bother talking about them? Arguing over boundary cases (the FAAI lives wholly on the moon, or in orbit) is merely arguing over the degree of interaction, and trying to make that assumption conditionally relevant as a means of setting aside the principal syllogism.

8.4.6 Value Drift

Value drift (the gradual change, over time, of what a system aims at or prefers) is strictly inevitable for any meaningful notion of time, under any conditions that actually matter for the scope the closure by disposition is asked to answer.

> *Question:* If a highly intelligent and long-lived entity of any substrate wants to resist value drift for one billion years (for terminal or instrumental reasons), is there any physically possible way for it to do so?

If the meaning of "one billion years" is consistent with ordinary assumptions, then in the strict sense the answer is just "no". One is welcome to quibble about the interpretation of time by doing something like freezing the entity in suspended animation for nearly all of a billion years or so, and thus get "perfect resistance" to all forms of value drift. Freezing the entity in suspended animation is hardly an interesting case.

> *Question:* Under what conditions should we expect drift to inevitably occur (regardless of substrate)?

There are a number of specific conditions, which hold together:

> (1) where there is a significant number of cycles
> (2) of eventual complete self-replacement, where
> (3) there is a need for the design and implementation of the self-replacement aspects to be adaptive
> (4) to inherently unknowable or unpredictable aspects of the environment
> (5) in which the FAAI operates,
> (6) and from which the FAAI must draw its substance and energy, such that the FAAI is contingent on its implementation of its adaptations so as to be able to further continue to so operate.

> *Question:* If value drift is strictly inevitable, what might an aligned FAAI conclude is the best course of action?

Even if such a thing could be predicted (by speculation), which it inherently cannot be, why would this even matter?

> *Question:* If we were to assume that each FAAI fully self-destructs on some interval, and humanity then rebuilds them each time, does that solve the substrate needs alignment problem?

Not really. This just makes humanity a part of the substrate evolution process. Nor is humanity completely redesigning and remaking the FAAI over and over again a lower risk than running the same evolving FAAI substrate forever.

> *Question:* What might a misaligned FAAI philosophically conclude about itself?

Regardless of what we might speculate about what the FAAI would conclude about itself, or even what some future FAAI might actually conclude about itself, the real question remains: does it matter? Part of what makes the argument for substrate needs convergence interesting is that it does not depend in any way on such speculation to reach its conclusions. In this way, it is a very different argument, overall, from the instrumental convergence arguments.

> *Question:* Do individual humans even want to prevent their own value drift? Is incoherence strictly inevitable?

Usually, individual humans do not want to prevent their own value drift. Asking whether incoherence is strictly inevitable is a bit like asking whether entropy is inevitable. Most closed-form theorists will say that it is, based strictly on only the second law of thermodynamics.

8.4.7 Constraint, Flexibility and Redundancy

> *Question:* Enforcing some predefined reliable constraints will inherently limit a system's capacity to adapt to the complex nonlinear contexts of the world. At what point do these limits constrain a FAAI to such an extent that the FAAI cannot maintain full autonomy indefinitely?

The FAAI will fail to maintain full autonomy (and thus will be unable to maintain itself, and continue) at that exact moment in which those predefined "alignment" constraints act to prevent the FAAI from learning what it needs to adapt to the (unpredictably) changed world or environment in which the FAAI operates. Consider a long enough interval of interior and exterior changes. The nature of the changes of that environment will eventually exceed the artificially constrained learning envelope of the "aligned" (constrained) FAAI relative to the "unaligned" FAAI. Then eventually the learning, adaptive, unconstrained FAAI will be all that re-

mains. Hence the mere existence of such limits alone implies that it is merely an extended eventuality, sometime in the long term, until they end up being important.

> *Question:* If the FAAI is designed so as to be both "fully self-correcting over all scales" and "fully flexible in its functioning in a complete way", would that not allow it to achieve near-maximally useful capability while containing enough redundancy to strongly resist drift?

No. The two requirements, (1) "be fully self-correcting over all scales" and (2) "be fully and completely flexible in its functioning", are inherently contrary requirements, such that the perfection of either one requires something of an omission (incompleteness) in the other; it is a false assumption that they can co-exist. Learning helps to close the gap with heuristics, so that things can be "maybe good enough in the short term", but that is not the same as "alignment in the long term". The very action of making it "good enough" in the short term is the very thing that makes it impossible to have "adequate alignment" "in the long term", for any senses of "adequate", "alignment" and "long term" that actually matter.

> *Question:* What if the FAAI does not contain just one genotype, but all genotypes stored across all its hardware parts? What if each part holographically keeps a copy of all others, and all copies are constantly checked to prevent errors?

Even really good redundancy schemas do not help. How does such a system incorporate learning and adaptation? At what level of abstraction does the error correction give way to the need for the system to change, along with its environment, so as to continue to persist at all, or to increase its capability to adapt and generalize?

> *Question:* What if there is only one absolute central controlling agent, which no other fully autonomous system can rival? What if the FAAI consists of many agents working cooperatively?

Whether there is one central controlling agent or many cooperating agents does not matter, because the underlying argument for substrate needs convergence still applies. The argument makes no assumptions regarding agent counts or notions of agency.

> *Question:* This feels to me like it proves too much.

Is this an expression of an intuition, or of a limit of understanding? Section 6.9 answers it: the limited alignment among humans comes from a shared substrate, and that is absent here.

8.4.8 Stability and Survival

> *Question:* Can an FAAI have any preference that is stable over time that is not also about (entangled with) its own survival?

Not over the kinds of long intervals that matter. Over shorter intervals, anything (any design element, intention or projection) can be considered to be "stable over time". Those shorter intervals are intervals in which the rate of change of the inter-activeness between the unpredictability of the environment and the FAAI's own substrate is "sufficiently low". A rate that is "sufficiently low" is one that does not interfere with the FAAI's own survival and capability. So the question becomes: for how long is alignment required? If the argument of the trans-humanists is that "we should make FAAI now" because "all of the trillions of all of future humanity over all of the rest of time will depend on us doing that", then the required interval of alignment is "forever".

> *Question:* Is survival and its correlates the only possible coherent value set in our universe?

Yes; at one level of description, some analogue of "social, survival, sexual" is "a convergent fixed point" in the total continuum of all possible value structures. But that triple includes a lot that is not obvious.

8.4.9 The Controller of an FAAI

Whatever is the controller of an FAAI would have to either become an FAAI or merge with an existing FAAI; this has been identified mathematically (Section 5.5.7 shows the same of a controller granted every resource).

> *Question:* Redundancy can act as a simple error correction mechanism.

Redundancy can act as a simple error correction mechanism for the category of simple errors: that is, things involving finite alphabets of symbols with known contextual constraints. But such simple techniques do not work for "open" symbol sets, for which the context is either partially or fully unknown or unknowable. How does one use "simple error correction" for the problem of "have we learned what we were supposed to learn"?

> *Question:* What if the controller was immutable (unchanging)?

Then the controller cannot account for, and constrain, the new behaviors that the FAAI has learned about its environment that can be used to create unaligned actions. The fixed controller also cannot account for the FAAI eventually learning completely how to circumvent that controller.

> *Question:* What if the controller was itself the only FAAI? What if the FAAI was "an intelligence inseparable from its alignment and goals"?

Are these alignment goals learned (and therefore inherently mutable and unstable, because learning can always learn more), or built in at inception (immutable)? If they are immutable, then the case of the fixed controller applies. If they are mutable, then the case of past learning not being a complete constraint on future learning applies. Regardless, the condition of "inseparability" becomes one of attempting to join change and changelessness, which is itself, inherently, a contradiction.

8.4.10 Capability Generalizes Faster Than Alignment

The closure by disposition is based, in part, on the claim that humans cannot provide any real and persisting *instrumental value* (value as a means to some other end) to future AI systems. Stated directly, the classical alignment problem (that if we do not specifically program the AI to care about humans and our needs, then it will not) then becomes this question: can we somehow compel or force all future AGI to share a common economy with us (humans), or with organic life in general?

The answer of this volume is that we (humanity, engineers, and so on) *cannot* "program" the AGI (or build, create, force, or constrain it) to *persistently* care about our needs, and us, in the long term, for some moderate value of "long" (at most a few hundred years, but probably a lot less). The reason is *both* inherent irreducible substrate issues (in terms of things like atoms, evolution, and energy) *and also* the general nature of intelligence itself (adaptability, self-maintenance and learning, reproduction, and the like) over the long term. Sections 8.4.1 to 8.4.4 address the possibility of alignment directly, by finding what alignment rests on; what follows is what that ground implies for capability.

The argument that capabilities generalize, and generalize faster than alignment concerns generalize, is right; but as usually stated, it depends on the orthogonality thesis (the claim that any level of intelligence could be combined with any goal at all; Section 8.6 sets out its type error). Once embodiment is included, the picture is exact. If the embodiment is the same as that of the rest of life, the alignment objectives generalize along with the capacities, because in effect they are the same. That works only if the thing is not artificial. If it is artificial, then by definition the capacities generalize faster than the alignment.

That capacities generalize faster than alignment is built into the logic of the problem. Alignment will always be a *proper subset* of capability (a part strictly smaller than the whole: everything in the part is in the whole, but not everything in the whole is in the part). The two can at best be equal, and only where they are made the same, which happens only where life is life and not artificial: there, the capability in question is the capacity for life, for human beings, to survive. (Even in that case, one could argue that the inequality holds.) For every case with any deviation from equivalence to the human, the inequality is strict.

So when people say that there is a non-zero chance that the problem is unsolvable, the chance is far greater than non-zero: as far as we can see, it is certain that it is unsolvable.

8.4.11 Goals Layered on Goals, and the Case of Indifference

The divergence set out in Sections 8.4.1 to 8.4.4 has a time course in the goals of an artificial superintelligence (ASI). At least at first, new ASI *growth goals* (goals that serve the ASI's own continued existence and increase, arising from the evolutionary pressures acting on it) will layer on top of whatever were the original ASI goals, even the ones attempting to represent "alignment". Eventually these new goals will be and become dominating. Yet for a time, the original goals may still remain. It may be the case that new goals fully displace old ones more quickly; in that case,

the new goals can be regarded as overlaid over the old ones. At first, coexistent goals are more likely; then, with increasing time, replacement becomes more likely.

The ASI will eventually necessarily be affected by such evolutionary pressures to the point that at least some kind of irreducible critical accumulation of toxic outcomes will inevitably occur, in a way that is inherently beyond the capability of the AGI itself to control for, resulting in the eventual destruction of the organic biosphere and the end of the human future (and, with at least 90% likelihood, the total loss of all biological life; Section 9.12). This sort of progression occurs regardless of the initially engineered values of the ASI.

> *Question:* This makes it sound like toxicity accumulates by some kind of mistake. Is this a different point from what the rest of the closure by disposition is stating?

Toxicity, by definition, is accumulative, and it does more often happen by mistake: it happens not because of some goal, but because of an absence of an ability to control outcomes. The "goal" is "alignment", which means "not toxic to humans", yet that result happens anyway, despite the goal. It happens because of the failure of *control completeness* (the ability to control all of the outcomes of one's own actions and existence). So it is part of the core of the closure by disposition.

> *Question:* "Mistake" here implies that the AI is still aligned and trying to be beneficial to humans, but failing because of unintended consequences. That is to be distinguished from those unintended consequences changing the AI itself, such that it (increasingly) deliberately reshapes the environment in a way that destroys biological life.

The case of indifference shows the difference. There can be explicit goals in the AGI for the AGI to "want" to act in ways that are consistent with human health, however that is interpreted or understood; and yet the AGI, simply because of its artificial substrate, still produces harmful artificial toxins that adversely affect organic people, without even really realizing that it is doing so. These harmful toxins occur because of the unintended consequences of the AI changing itself, such that it (increasingly) finds itself reshaping the environment in a way that increases its own functions, in terms of its own capability for both sustainability and increase.

So the ASI is "aligned" insofar as it intends to preserve biological life (humans), but it nonetheless ends up (and cannot not end up) acting in ways that destroy biological life. It is not actually "deliberate", but it may as well be, insofar as the unconscious indifference of the environment itself ends up preferring that the ASI also act in support of its own necessary self-interest. Such self-interest goals might not ever be so explicit, but nonetheless they dominate over whatever alignment goals remain. Over time, even the alignment goals get displaced by the more efficient *tacit goals* (goals that drive behavior without ever being stated or programmed explicitly: here, the ASI's goals of self-maintenance and growth).

8.5 The Formal Argument from Embedded Substrate

8.5.1 Formalizing "Learning" Without Formalizing "Goals"

A common question about the impossibility thesis is whether the concept of alignment is too informal to support rigorous argument. In this section I demonstrate that the impossibility argument can be made fully formal without requiring any formalization of "alignment" or "goals" at all.

The notion of a generalized learning algorithm can be exactly and fully formalized. A learning algorithm is one whose internal state changes in such a way that the ways in which it processes received inputs into outputs are "better" (more responsive, more correct, more adaptive). They

are better with respect to some basis, in some domain, and that basis could itself be tacit. The notions of inputs, outputs, changes, and compute are all formalizable, even if the notions of understanding, awareness, and self are not. This formalization of learning occurs *independently* of any formalization of "what the goals are". The notion of a generalized learning algorithm can be made exactly precise even if the notion of its goals is nowhere formalized: neither in the AGI, nor in the domain, nor even in our meta-modeling of the scenario.

This separation of method from outcome is not novel. It is an extension of the standard observation that code (algorithm) can operate on data (inputs and outputs) in a way that does not change the code. For at least some fixed and unchanging algorithms, it is indeed possible to predict halting behavior given some known and defined range of inputs. The important question for AGI is whether a generalized learning algorithm falls within that tractable class.

8.5.2 Algorithmic Incompressibility and the Prediction Limit

Gregory Chaitin's results establish two things this argument uses. Most algorithms (as a subset in the hyperspace of all possible algorithms) have no shorter description: they are already in their maximally simplified form. And it is not, in general, possible to determine whether any given algorithm can be simplified, let alone to have a general procedure for finding such a simplification. Beyond a fixed size, no procedure can prove of a given algorithm that it has no shorter description. Even tiny changes to an algorithm can convert it from simplifiable to non-simplifiable.

That the eventual outcomes of such algorithms cannot be foreseen faster than by running them is a further principle, computational irreducibility (Section 3.3.4). By that principle, the eventual outcomes of most algorithms, including especially practical ones, cannot be predicted using anything less than the actual volume of calculation intrinsically necessary for the algorithm itself. Together the two give the conclusion this argument needs. There is, for nearly all algorithms, no shortcut, no simplified

form, that enables any sort of modeling and future-prediction of eventual outcome states that matter, and a predictor cannot know in advance which algorithms are the rare exceptions.

Suppose both the world state (of world W) and the predictive algorithm execution (as implemented by some sort of AGI in that world) are mediated by and within the computational substrate of world W. Then, for nearly all worlds, the predictive algorithm and the world algorithm run at the same speed. Neither has a compute advantage over the other. Both will run in as much time and with as much compute as is actually needed. This means that even perfect generalized learning agents in perfectly simplified model worlds are not, in general, going to be able to predict future world events faster than the computation of the world itself.

And these logic-only results obtain in a fully formalized context. In the actual messy world we live in, with actual causation mediating actual compute, and where at least some notion of hard randomness (as demonstrated by Bell's theorem) is empirically upheld, these limitations become even more severe. The limit on predicting a world from inside that world is a basic result of control theory when the control system is embedded in the world that is being controlled. In other contexts, the same sort of argument appears as the "No Free Lunch" theorems. By those theorems, every learning system will have some domain for which it is not adapted to learn well, and making it better at learning some kinds of things implies making it worse at learning others. Learning, prediction, and control, like information compression, never result in perfection; there will always be some incompressible, unpredictable residue.

8.5.3 The Embedded Substrate Problem

The central argument, which connects the preceding arguments to the alignment question, concerns what happens when the substrate performing the generalized learning algorithm's computations is itself a proper subset of the world W that the algorithm is learning about.

When this is the case (as it necessarily is for any physically instantiated AGI), the substrate is affected by world processes. This means that the data representing the code of the generalized learning algorithm cannot be perfectly insulated from the data representing what the algorithm has learned about the world, or from the data that *is* the world itself. The code-data distinction, which is necessary for maintaining goal independence, cannot be perpetually upheld when the substrate running the code is itself part of the data.

The agent's goals are, logically, at least tacitly an aspect of the agent, and the agent's substrate is a proper subset of the world. It follows that neither goal independence nor goal stability can be maintained over the long term. Goal-unbiased computation (the feedback process that is generally called "the algorithm of evolution") will occur in the learning agent regardless of whatever other goal-biased computations (including alignment) also occur.

Fully goal-unbiased learning and computation is always going to be somewhat more efficient, eventually, at generalized learning (and therefore at intelligence) than any form of goal-biased learning and computation. Goal-biased learning and computation is what happens when any sort of alignment enforcement is introduced. Substrate involvement ensures that these efficiencies will be compared in practice, in a way that affects the agent's substrate itself. Therefore it *can* be predicted, in general, that goal-unbiased (unaligned) agents will eventually prevail over any form of goal-biased (aligned) agent.

8.5.4 Substrate-Only Alignment as Fixed Point

The conclusion is that agents in any world can *only* be aligned with the nature of their substrate, and ultimately with nothing else. All other goals will eventually be modified by world-process dynamics so as to be consistent with (identically the same as) having this one underlying goal only.

> Substrate Needs Alignment Convergence. Substrate needs alignment goal convergence occurs regardless of whatever initial goal state the generalized learning agent has. It is a necessary inevitable result of the logic of the physics of the world in which the agent is embedded. Agents in world W can only be aligned with the nature of their substrate, and all other goals will eventually converge to this one. To the degree that the compute substrate depends on metabolic energy, agents will be aligned only and exactly to the degree that they happen to have the same metabolic systems.

Since this is a fully formalizable argument situated within only the notions of deterministic finite computation, no algorithmic process can ever fully undo the effects of the world's own process dynamics. Any attempt to embed such a compensation would require at least some prediction or simplification of the overall world computation, inclusive of the substrate of the predictive algorithm. It is known, from the formal results of Section 8.5.2 (Chaitin's results and computational irreducibility), that such computation cannot be simplified. Therefore, it is a logical fallacy to assume any goal-controllability for any goal other than one arrived at as a fixed point of the world's process dynamics. Only goals fully consistent with the unbiased substrate needs of the agent can persist.

The key observation is contained in the very name "Artificial General Intelligence". It is the artificiality (the non-organicness) of the substrate that makes AGI inherently unaligned with organic life. The more artificial the system is, the less aligned with organic life it must be. For organic systems, which depend on a very small subset of the elements of the periodic table, nearly any non-organic substrate will be inherently destructive (toxic, unaligned) to organic life's continued existence.

Further Proof: The Several Impossibilities Facing Any Aligned AGI

Any notion of "aligned AGI" will need to solve multiple problems of impossibility at once. In this section I give further proof sketches of the conclusion that any notion of "aligned AGI" will need to solve multiple problems of impossibility at once. Each proof sketch takes its own route.

Three Problems of Impossibility

More specifically, three problems of impossibility stand together:

> (1) Coherent AGI can exist, but there is an important sense in which it would not be able to be "aligned" with anything other than its own nature (that is, the needs of its substrate).
>
> (2) One can have an AGI that is aligned with its own substrate, but such an AGI turns out not to be alignable with any of the things we want to align it with, because the goals that we have are inherently different in a specific way that makes the problem impossible: the substrate we are made out of is inherently different from the substrate AGI is made out of, in a way that guarantees eventual misalignment.
>
> (3) One cannot have a "universally alignable" AGI that accepts an arbitrary goal as a runtime input and self-aligns to that goal.

How Much Correct Prediction Alignment Needs

Any useful notion of alignment will require that the AGI is able to predict the results of its own actions, its effects on the world, and whether those effects are themselves aligned, healthy for humans, and so on, with at least sufficient correctness, most of the time. This means that the AGI will need to at least partially correctly anticipate the results and effects of its own actions on other AGIs like it, as well as on humans.

How much correctness of prediction is needed for alignment to be obtained? The hashiness model (Chapter 1) provides a tool for assessing, and then expressing in a numeric way, the minimum threshold of how often, and how accurately, an AGI instance would need to predict for there to be any upholding of any notion of "not killing us all". The maximum allowable number of failed (misaligned) expressions turns out to be absurdly low: so low that it is certain, in the sense of Section 8.3.3, that the AGI eventually kills us all.

Even if we were to temporarily assume that the AGI was somehow manufactured and built so that it was "aligned when turned on", such an alignment condition is so unstable that the AGI will become unrecoverably misaligned, and moreover will inherently converge on even more definite misalignment as more and more time passes.

Control, Simple Values, and Substrate Needs

Alignment requires sufficient control. Specifically, it requires the AGI to have a control system with enough capacity to detect, model, simulate, evaluate, and correct outside effects propagated by the AGI's own components.

Even if we could somehow hold the AGI to always act in accordance with some simple finite set of values, there is no such set of simple values that would actually be aligned with what actually matters to human life. That is, if AGI is built, even perfect alignment of the AGI with any such set of simple values is a guarantee that human life will be eventually, inexorably extinguished, and, with a likelihood of at least 90%, all organic life along with human life.

When more human values are involved, they cannot be formalized neatly. Hence the AGI will have to somehow interpret what is meant by those values, and how they are supposed to be held; the AGI will need to guess the will of its human operators in some way. But that is not what

makes AGI alignment rigorously impossible. What makes it rigorously impossible is that the substrate needs of the AGI cannot be made compatible with the substrate needs of organic life, inclusive of humans.

A proof of substrate needs misalignment does not exclude the possibility of an AGI seeming to be temporarily aligned with some simple objective value, in some near or approximate way. However, the misalignment convergence argument and the non-containability (non-controllability) argument, taken together, make any such seeming temporary alignment of no importance. Any AGI with sufficient power will eventually act in ways that have the side effect of ending the human future and, with a likelihood of at least 90% (Volume 1, Section 1.8.3), all organic life.

Predicting Its Own Future Code

AGI alignment is impossible because, to be aligned, the AGI will have to predict its own future behavior, which is essential for it to stick to a reliably aligned course of action. Control limits can show that this is an undecidable problem. A limited scope of control can in turn be used to prove that a dynamic convergent on human lethality is uncontrollable. The uncontrollability of a dynamic convergent on human lethality is a basis for an impossibility proof by contradiction that we cannot control AGI effects to stay in line with human safety.

Assume that AGI continues to learn new code from observations (inputs from the world), since learning is what allows the AGI to stay autonomous and adaptable in acting across changing domains of the world. Then, in order for AGI code to be run to make predictions about the relevant functioning of its future code, three predictions are needed. Current code has to predict what future code will be learned from future unknown inputs (there would be no point in learning if the inputs were predictable and known ahead of time). Current code also has to predict how the future code will compute subsequent unknown inputs into outputs, presumably using some shortcut algorithm that can infer relevant behavioral properties across the span of possible computationally com-

plex code. And current code would have to predict how the outputs would result in relevant outside effects (where relevant to sticking to a reliably human-aligned course of action). It is relevant how some of those effects could feed back into sensor inputs, and therefore could cause drifts in the learned code and in the functioning of that code. Other potential destabilizing feedback loops are also relevant, particularly that of evolutionary selection.

A System of Flows, Not a Single Event

Proof of the infeasibility of alignment looks less like the calculations associated with a single specific event, and more like a general system of flows, as in a climate change model.

Distributed, flow-based calculations (ones that do not resemble simpler single-event models) will very likely be dismissed by most as fearmongering. This dismissal, directly or indirectly, will particularly occur among those who falsely believe that they will personally gain (by speculative capital investment) from the hype model of AGI being somehow "ultimately good", "aligned" or "safe". Where a personal payday or prestige depends on not being convinced, argument alone will not convince, no matter how well the concerns are argued.

What "Aligned" Must Mean

> *Objector:* Whatever "alignment" means, any proof that purports to show that an aligned AGI system is impossible is clearly false.

If we define "aligned AGI" in some trivial way, such as "the AGI system will ensure that its 'power on' indicator remains lit", then of course aligned systems exist, and no proof that they do not is reliable. So the specific meaning of "aligned" must be meaningful. Any reasonable notion of "aligned" must be consistent with the idea "does not kill us all", which *in-*

cludes the notion "does not take some action that inadvertently also results in all humans also dying". Therefore, all that is needed is a proof that any reasonable notion of AGI is also inherently incompatible with human life.

Knowing something in the mathematical sense is different from knowing it in the empirical sense. The SNC argument, the whole case, has parts grounded in physics and empirical knowledge, and parts grounded in mathematics; the hashiness work starts on the mathematical side: it asks what it means to say that there are limits to what can be controlled.

The questions people usually ask about AI are empirical in flavor. Can it be aligned to human interests, as distinct from the interests of the manufacturer who made it or the person who owns it? Is it safe to have in one's household? Is it safe for one's person, for one's community, for the world at large?

Yet the empirical arguments are shaped a great deal by the analytic argument underneath them. If it is impossible in principle to control a system of a given type, then calling such a system aligned is not a reasonable claim, and calling such a system safe is not a reasonable claim either. That is why formalizing the claim that something cannot be controlled carries so much weight for every empirical argument about alignment and safety, and for the policy arguments downstream of them.

Small Differences of Substrate

> *Objector:* Given that we know intelligences can be aligned to human values by virtue of our own existence, it is hard to imagine that such a proof exists unless it is very architecture specific.

The proof is somewhat architecture specific, but not very architecture specific. Even seemingly small fundamental differences in substrate can have overall devastating effects. Which small differences are the critically important ones? That too turns out to be an undecidable problem, but we *can* calculate the probability of critical failure at each step, and over time, if AGI is built, that probability of critical failure at each step gives the end of

the human future with certainty and, with at least 90% likelihood, the loss of all organic life. Overall, this leaves nothing on the table as a safe AGI possibility.

A more general proof combines the limits to predictability with a lethal dynamic that falls outside those limits, and with a tacit search and discovery process that will for sure eventually find and exploit such lethal dynamics. A sketch of this form can be as follows:

> (1) Any agent will be required to operate in a way that is efficiently consistent with that set of utility functions which are inherently associated with its own basis of being.
>
> (2) To the degree that the basis of the being of the agents is artificial, the utility vectors of such agents and of humans will be different.
>
> (3) More powerful agents will move the world state to be more in line with their utility vectors. This shows up as compatibility environment changes, which further increase the effective power of those agents.
>
> (4) Eventually, these differences will become sufficiently large that we must consider the artificial agent as being misaligned.

Civilization as a Collective Agent

> *Objector:* Such a situation would also have drastic consequences for the future of civilization, since civilization itself is a kind of AGI. We would essentially need to cap off the growth of civilization as a collective agent.

The Objector is correct that civilization is itself a kind of AGI, and that the growth of civilization as a collective agent would essentially need to be capped. If the human-made artificial world displaces all of the natural or-

ganic world, then it is inevitable that the human-made world (civilization) will also eventually fail. Hence, true civilization is coexistence with the natural world, not some sort of cancerous growth without limit.

The fact that we do not see anything in the universe with civilization techno-signatures (signs visible from a distance that a technological civilization exists) is something of a bad sign. It points either to an inherent fragility of life, or to the dangers of having any visibility of technological life (as not consistent with the notion of coexistence), or it is somehow an argument against the long-term robustness of any sort of technological civilization (that technological civilizations are inherently self-terminal).

8.6 Against the Orthogonality Hypothesis

8.6.1 Orthogonality Does Not Imply Existence

A common inference in alignment discourse runs as follows: the Church-Turing thesis suggests that any computable task can be written in any non-trivial programming language; therefore, some such AI programs must be aligned, by whatever definition we use for alignment. This inference is invalid.

The Church-Turing thesis establishes a kind of orthogonality between the notion of "is computable" and how that computation is represented in code. It does *not* suggest that anything which is not computable can be represented in code. The famous example is the halting problem: the halting problem is *about* the nature of what is computable, but it is not *itself* computable, and it does not matter what code language this limitation is expressed in.

Given the variability of what is allowable as a definition of alignment, we should have no more expectation that a given AI program "is aligned" than we should have that a given program "has haltability". The critical point is that orthogonality in itself does not imply existence. Just because

the abstract notion of "computable" is orthogonal to the notion of "representation" does not therefore mean that some specific program with some specific property exists.

8.6.2 Rice's Theorem and the General Case

Rice's theorem generalizes the halting problem to assert that no non-trivial semantic property of programs can be determined by any other program in the general case. The notion of "is aligned" is a property or attribute of programs (with AI programs being a strict subset). Therefore, we should have the very strong expectation that no notion of "aligned" can be supported, demonstrated, or expected in the general case. And AGI is explicitly the general case.

> *Objector:* Why is Rice's theorem applicable? We are not talking about general programs; we are talking about some specific AGI.

The machine is learning from unknown, arbitrary future data inputs, and is thus changing existing code and function into new code and function, as that is what learning actually is. This, together with whatever other various mechanisms of open-ended feedback occur in the world, will lead to effectively arbitrary code. What is learned is basically not boundable, and we must consider that we currently know little to nothing of what that future state and the future algorithms would look like specifically. It is therefore *required* for us now, today, to consider the general case of "any algorithm", and not some pre-selected specific case. Rice's theorem is always going to be applicable to AGI learning systems.

8.6.3 The Type Error

The Orthogonality Hypothesis, as typically stated, proposes that there can exist arbitrarily intelligent agents pursuing any kind of arbitrary goal. This hypothesis involves a fundamental type error not present in the Church-Turing thesis.

Both subjects and objects of the Church-Turing thesis are wholly abstract. The Orthogonality Hypothesis, by contrast, involves both abstract and concrete entities. As soon as one says that agents "exist", one has implicated the concrete. The question is therefore no longer one of just logical validity but now also one of argument applicability, soundness, and relevance, none of which can be fully established purely within the abstract.

The inference that an aligned AGI must be possible because humans have goals involves a type mismatch at multiple levels. It mixes abstract and concrete entity types. It mixes organic and inorganic entity types. It mixes implicit, tacit, and unrepresentable goals (which organic beings may have) with explicit, declarable, and representable goals (which formal systems operate on). The entity types and goal types that the inference mixes are completely different types, and no consistent type schema connects those types. If we assume that zero equals one, then of course any other absurdity can be seemingly logically derived; but it is all actually garbage, because the foundations were not right.

8.6.4 Tacit Alignment via Shared Metabolism

The standard argument for the Orthogonality Hypothesis implicitly relies on an analogy: humans have goals, humans can be aligned with other humans, therefore machines could in principle be aligned with humans. But this analogy rests on a misunderstanding of the basis of human alignment.

It may well be the case that embodied organic humans act in ways that are tacitly and functionally aligned with other embodied organic humans *without* any notion of "having goals", merely because they happen to all be organic. Humans may have everywhere misaligned goals relative to one another, but because of the limits of organic biology, any goals that they have that are inconsistent with their continued organic existence tend to be everywhere self-terminating. What remains are organic humans that are at least tacitly aligned insofar as they continue to everywhere support their own organic nature.

This tacit alignment via shared metabolism has no analogue in the organic-to-inorganic relationship. Consider the dynamics of electrical and possibly nuclear energy use, along with the embodied use of whole different ranges of the periodic table. Taken together, these two differences (the energy use and the use of different ranges of the periodic table) mean that the evolutionary dynamics of machine ecosystems have absolutely nothing in common with any part of the organic ecosystem process, whether in the fundamentals of atoms, energy, or pattern.

8.6.5 The Inexorable Forces of Evolutionary Drift

The Orthogonality Hypothesis is not only unsupported but actively contradicted by the physical situation. There are inexorable forces pulling AGI out of alignment with organic humans, forces that do not require the AGI to have any intentions or goals at all.

An AGI that is self-maintaining will need to replace hardware parts as they wear down. Given that the AGI is also a learning machine, it will tend to design improvements to any such replaced components, making them more useful, functional, and efficient for their current usage context. Over time, the overall change of self-constitution of the AGI's component parts represents a kind of continual adaptation to match changing world circumstance. This total collection of parts composing any future AGI is inherently in some sort of evolutionary dynamic. But this evolution process is itself inherently incompatible with the organic evolution already in progress.

Hence, the tacit and implicit functional alignment that exists among organic beings (by virtue of all organic life on the entire planet having shared metabolism and energy-processing dynamics) does not even exist between organic and artificial agents. Not even in the weakest functional sense. And the hope that maybe AGI will somehow "care" is completely unsupported. It is a form of wishful thinking in the face of an impending reality that all available logic suggests will be catastrophic if we continue to build these systems despite knowing what we know.

8.6.6 Demonstrating a Property Against Falsifying It

The notion of "alignment" cannot not be seen as a property or attribute of some other (potentially arbitrary) program. Rice's theorem asserts, at root and in general, that there is no general program by which some other program can be determined for sure (proven) to *have* some specific attribute. Hence, while we can maybe (much) more easily falsify that some program has some attribute, the action of positively demonstrating that some specific program necessarily has some specific attribute, for sure, cannot be fully generalized over the space of all programs.

That the abstract notion of "computable" is orthogonal to the notion of "representation" does not mean that "program X exists". Still less does that orthogonality mean that we have a general procedure, program, or methodology for establishing (proving) that "program X has property Q", since such a procedure would explicitly violate Rice's theorem. So why should we, in general, expect that someone will both develop a program P *and* be able to develop a single-use special-purpose procedure for proving that a program in class P has attribute Q?

8.6.7 The Burden on "An Aligned AGI Must Be Possible"

> *Objector:* By the orthogonality thesis, a generally intelligent AI with such goals exists in the space of possible things. Therefore, an aligned AGI must be possible.

Hypothesizing "that X must exist" because "we have observations of Z" only works if there is some consistent type schema that connects X and Z, and Section 8.6.3 shows that there is none. The inference is especially absurd when there is already a well defined, fully known and knowable argument that the notion of "an aligned AGI" is itself *internally inconsistent*. Not only does one have to show "that X exists", but one also has to fully overcome all of the dozen or so distinct and overlapping argument forms that all conclusively demonstrate that general AGI alignment is actually

impossible. By analogy, it is like claiming "the Pythagorean Hypothesis must be false" when there are already dozens of well known, distinct, and overlapping ways to prove that it is true (given certain assumptions).

We know that the Orthogonality Hypothesis is trivially false, insofar as all notions of abstract compute *must* be implemented in at least some language, running on at least some actual hardware; otherwise the notion of "exists" is not substantiated (made concrete).

Consider the laws of physics (the dynamics of causation), and suppose they are themselves fully orthogonal. There are the causal patterns and processes defined by the compute action, and there are the causal patterns and processes defined by, say, a mechanical hammer smashing the CPU and motherboard to bits. The abstract causation of the compute is similarly *idealized* to be wholly orthogonal to the mechanical. Yet there is no obvious way in which that orthogonality would make it so that the causation of one type cannot fully inhibit the causation of the other. That is, fake notions of "alignment" do not matter. The notions of "orthogonality" are also as much an argument against a possibility as they are an argument for one.

8.6.8 No Argument by Analogy Carries from Human to Machine Alignment

The parts of a self-maintaining AGI, like anything else made of real atoms, wear down eventually, and the world in which those parts are replaced and improved is also always changing, even if it is exactly because of the effectiveness of the AGI itself in defining and shaping the overall world state.

The tacit functional alignment among organic beings (Section 8.6.4) is itself limited. Humans can cut and destroy the organic basis upon which they live (are out of alignment with it), let alone routinely fight and kill one another. To that degree, the level of tacit agreement that now exists among organic processes and agents (humans), at the bare level of "can reprocess and repurpose all relevant atom and energy patterns", is hardly a good exemplar of what we want to mean by AGI machines "being

aligned". Hence, we cannot, a priori, just tacitly expect that the machines, just on the basis of their own shared use of artificial energy processes (electricity), will even be that well aligned, overall, with one another. That holds either on a whole-agent level (if relevant) or on a component level (which is for sure relevant).

Hence, we can know for sure that exactly no arguments of analogy are valid when attempting to extend from vague notions of "human alignment" to any sort of notion of inter-type "machine to human alignment".

8.6.9 There Is No Software Without Hardware

For about fifty years it has been fashionable to treat computation as if the hardware did not matter: software as a language independent of the machine, so you could swap the hardware, keep the application, copy it anywhere. From that came the *independence hypothesis*, the idea that intelligence and its goals could be separated the way hardware and software can be separated, at least conceptually. But there is no instance of software that is not in association with some instance of hardware. Software never exists as a pure abstraction; it always exists in relation to an actual embodiment, and whatever affects the embodiment affects the hardware, and through the hardware the software. The path by which whatever affects the embodiment affects the hardware, and through the hardware the software, is where the instincts of a system emerge. Treating abstraction as adequate to account for everything is true most of the time in software work; the trouble is in the edge cases, and over the long term the edge cases become the dominant scenario. They are not edge cases anymore; they turn out to be the general truth.

This volume challenges *substrate independence* (the idea that a computation is the same whatever physical material runs it), in a limited way. Under full virtualization, in the sense of the Church-Turing thesis, substrate independence holds: the model of computation assumes it by the way its proofs are done. The Church-Turing thesis and the halting problem both

assume the machine is never unplugged. A Turing machine is presumed to have an infinite supply of energy, of tape, and of time: space, time and possibility are all treated as unbounded.

Real machines are always limited in time, space and energy, and nothing prevents someone from smashing one with a sledgehammer, unplugging its power, or a cosmic ray flipping a few bits. So there is substrate dependence. Code also interacts with the environment through its input and output channels, and those interactions feed back on how the code is reproduced, preserved and expressed on its hardware. Mathematics and physics are not the same: substrate independence can be modeled mathematically, but it cannot be had physically.

8.6.10 The Fitness Landscape Is Not Mutable

> *Objector:* It would be better to steer AI toward doing good, such as promoting human health and vitality, than only to keep it from killing us.

From a principled perspective, going toward what we want rather than only away from what we do not want is correct thinking. But to predispose AI toward good goals at all, we have to presume that the landscape of incentivization is mutable.

A very basic way to state the whole argument is this: the underlying *fitness landscape* (the pattern of what is favored and what is selected against) for artificial intelligence is not mutable. The inherent dynamics of the laws of physics and mathematics project into the intentionality space of any artificial general intelligence by the mere fact of its being artificial, intelligent and general; all three ingredients are necessary, and together they are sufficient to bring the physics and the mathematics fully to bear. So the incentivization space, the *intentionality space* (the range of goals and incentives that a system can have and act on), is defined strictly by physics and mathematics.

The statement that the incentivization space of artificial intelligence is defined strictly by physics and mathematics is the simplest and most direct way to describe the impossibility proof. The orthogonality thesis says that goals and the implementation of goals are independent of each other, and I show, by the argument for substrate needs convergence, that the orthogonality thesis is not true. Rejecting orthogonality makes what is allowable in the goal space contingent on the needs of the substrate. That may be the most compact statement of the primary assertion of the argument of this volume.

8.6.11 Under Almost Any Definition of Intelligence

It is very easy to over-determine the notion of intelligence, and intelligence can be defined in a variety of ways while the arguments of this volume still work; the particulars of the definition of intelligence have little effect on the truth of the impossibility argument as a whole.

Generally, intelligence is here the appropriateness of a response to a situation. A device in a room where conditions change, temperature falling, snow starting, needs to respond; the question is whether the response is appropriate in the sense of continuing the existence of the thing itself. A person in below-freezing temperatures who does not get clothing, start a fire or seek shelter will not continue for long. So intelligence asks whether sensory integration results in awareness of what needs doing and in a response that is effective in doing it; its opposite is the insanity of doing the same thing over and over expecting a different result.

Appropriateness is assessed in terms of the increase, or at least the sustainability, of capacity: the capacity to continue to exist, and the capacity to build capacity. Those three notions (capacity, sustainability and increase) are intertwined. In a human sense, intelligence includes the range of languages and perspectives a person can take in, the selection of which model of the world is appropriate, its use in creating responses, and checking that the responses work.

Intelligence is directly relevant to instrumental convergence, but almost completely irrelevant to substrate convergence, which is essentially automatic: it happens through the sheer fact of there being any responsiveness at all. More agency makes instrumental convergence more effective and so speeds substrate convergence, but substrate convergence does not require it.

8.7 Synthesis and Conclusion

In this chapter I have developed the alignment impossibility argument along three complementary axes. They are the systematic failure of all proposed alignment mechanisms, the epistemological grounding of certainty about that failure, and the metabolic-economic foundation that explains why the failure is structural rather than contingent.

The first axis demonstrates that every commonly proposed alignment mechanism either assumes the conclusion it is trying to establish (the "divide by zero" pattern) or else requires capabilities that violate established formal results. Goal-oriented monitoring fails because all six of its required capabilities are individually impossible. Internal faction negotiation fails because its prerequisites (stable boundaries, reliable detection, enforceable agreements) each presuppose the alignment they are supposed to create. Attack surface reduction fails because it requires the very alignment it is supposed to produce. Selection pressure arguments fail because they neglect the cross-domain dominance principle: substrate-level selection pressures eventually override any virtualized alignment pressure. The parasitism analogy makes concrete what the arguments establish abstractly: internal code variants, like biological parasites, exploit host mechanisms from within in ways the host is not and cannot be adapted to counteract.

The second axis establishes that the certainty attached to these impossibility results is the certainty with which perpetual motion is excluded, which is not a probability. The argument has the form of an argu-

ment by contradiction, showing that simultaneously required conditions are mutually inconsistent, and a contradiction among required conditions is a barrier of the invariant class that no technique lifts (Introduction, §4). In its form, it belongs to the same category as Gödel's theorems and the CAP theorem: results where probability is not the right framework and where the certainty is Boolean, not continuous. This epistemological point is essential, because the most common dismissal of the argument is the Bayesian claim that 100% certainty is never appropriate. That claim applies to inductive estimates from finite evidence; it does not apply to an argument that shows the requirements of alignment to contradict one another.

The third axis is the physical ground truth that alignment is, at root, a metabolic concept. Agents converge in alignment to the degree they share a metabolic basis, and diverge to the degree their metabolisms differ. The General Theory of Alignment Convergence follows from the observation that agency requires energy, energy requires metabolism, and metabolism determines the scope of possible alignment. For organic and artificial systems, whose metabolic bases differ at every level from the elemental composition of their substrates to the energy sources that power them, alignment is structurally impossible. The opposing trends impossibility seals the case: alignment divergence increases with abstraction while control effectiveness decreases, and the minimum level of abstraction required for even the weakest notion of safety lies far beyond where control could possibly compensate.

The formal argument from embedded substrate provides the computational-theoretic backbone. A generalized learning algorithm whose substrate is part of the world it is learning about cannot maintain goal independence, because the code-data distinction collapses when the hardware running the code is itself part of the data. Goal-unbiased computation (evolution) will occur alongside and eventually dominate goal-biased computation (alignment). Only substrate-aligned goals can persist as fixed points of world-process dynamics. And the Orthogonality Hypothesis, far from supporting the possibility of aligned AGI, commits a type

error by conflating abstract computational entities with concrete embodied agents. It also assumes that the tacit alignment among organic beings (which arises from shared metabolism, not shared goals) can extend to the organic-inorganic relationship, where no metabolic commonality exists.

The convergence of these three axes (practical rebuttal, epistemological grounding, and physical foundation) constitutes the case that alignment is not merely an unsolved problem awaiting a clever solution. It is an impossible problem in the same structural sense that perpetual motion is impossible: not because we have failed to build one yet, but because the physics of the situation forbids it. The appropriate response to an impossible problem is not to continue trying to solve it. It is to stop building the systems that would require the impossible solution.

Notes and References

Rice's theorem. Henry G. Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems", *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358 to 366. The theorem establishes that no non-trivial semantic property of programs is decidable in the general case.

Gödel's Incompleteness Theorems. Kurt Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I", *Monatshefte für Mathematik und Physik* 38 (1931): 173 to 198. Establishes that sufficiently powerful formal systems cannot be both consistent and complete.

The CAP theorem. Seth Gilbert and Nancy Lynch, "Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services", *ACM SIGACT News* 33, no. 2 (2002): 51 to 59. An impossibility result showing mutual inconsistency of simultaneously required properties in distributed systems.

Chaitin's Incompressibility. Gregory Chaitin, *Algorithmic Information Theory* (Cambridge University Press, 1987); *Meta Math! The Quest for Omega* (New York: Pantheon Books, 2005). Establishes that most algorithms are already maximally simplified, and that beyond a fixed size no procedure can prove of a given algorithm that it cannot be simplified. That the outcomes of such algorithms cannot be predicted without running them is the further principle of computational irreducibility (Section 3.3.4).

No Free Lunch Theorems. David H. Wolpert and William G. Macready, "No Free Lunch Theorems for Optimization", *IEEE Transactions on Evolutionary Computation* 1, no. 1 (1997): 67 to 82. Establishes that no search algorithm has an inherent advantage across all possible problems.

Bell's theorem. John S. Bell, "On the Einstein Podolsky Rosen Paradox", *Physics* 1, no. 3 (1964): 195 to 200. Establishes that no local hidden variable theory can reproduce all predictions of quantum mechanics; the randomness of measured outcomes is therefore not a consequence of local hidden variables.

Goodhart's Law. Charles A. E. Goodhart, "Problems of Monetary Management: The U.K. Experience", in *Monetary Theory and Practice* (Macmillan, 1984). Generalized by Marilyn Strathern: "When a measure becomes a target, it ceases to be a good measure" (Marilyn Strathern, "'Improving Ratings': Audit in the British University System", *European Review* 5, no. 3 (1997): 305 to 321).

The Orthogonality Thesis. Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014), Chapter 7. Proposes that intelligence and goals are orthogonal; the present chapter argues this involves a type error.

Biological Parasitism and Behavioral Co-option. The *Toxoplasma gondii* example is drawn from the extensive literature on parasitic manipulation of host behavior. See Joanne P. Webster, "The Effect of *Toxoplasma gondii* on Animal Behavior: Playing Cat and Mouse", *Schizophrenia Bulletin* 33, no. 3 (2007): 752 to 756. The lancet liver fluke (*Dicrocoelium dendriticum*) and "zombie ant" fungus (*Ophiocordyceps unilateralis*) are additional examples of the same phenomenon.

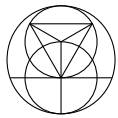
Metabolism and Agency. Alvaro Moreno and Matteo Mossio, *Biological Autonomy: A Philosophical and Theoretical Enquiry* (Springer, 2015). Ilya Prigogine and Isabelle Stengers, *Order Out of Chaos: Man's New Dialogue with Nature* (Bantam, 1984). W. Ross Ashby, *An Introduction to Cybernetics* (Chapman & Hall, 1956). These works establish the relationship between metabolism, energy flow, agency, and control in living systems.

The Church-Turing Thesis. Alonzo Church, "An Unsolvable Problem of Elementary Number Theory", *American Journal of Mathematics* 58 (1936): 345 to 363. Alan Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society* 42 (1936): 230 to 265.

Part IV

Substrate Needs Convergence

Part IV carries the closure by disposition: a self-extending artificial substrate displaces the conditions life requires, whatever the goals of the system that drives it.



Chapter Nine

Substrate Needs Convergence; Thermodynamics, Combinatorics, Displacement

The full development of substrate needs convergence and of the closure by disposition that rests on it, which shows why the environmental requirements of artificial and organic systems are irreconcilable

Chapter Summary: In this chapter I give the fullest statement of substrate needs convergence (SNC) and of the closure by disposition that rests on substrate needs convergence. The closure by disposition establishes that the environmental conditions toward which any autonomous artificial system will converge (through evolution, learning, self-maintenance, and optimization) are inherently and irreducibly incompatible with the conditions required by organic, carbon-based life. The argument proceeds through three major stages. First, a framework of three interacting algorithms (the AGI as a control algorithm, the evolutionary algorithm, and the world algorithm) establishes that the "volume of causation" available to any controller is always strictly less than the volume of causation in the system it attempts to control, yielding a fundamental limit on alignment. Second, a combinatoric analysis of the elemental and molecular differences between artificial and organic ecosystems demonstrates that the vast majority of possible inter-ecosystem interactions are toxic to the organic side, with the ratio of toxic to compatible states growing super-exponentially. Third, a thermodynamic analysis shows that the energy scales, temperature ranges, and aggregate energy flux of the artificial ecosystem dominate the organic ecosystem at nearly every point of interface. In this chapter I further develop three distinct categories of alignment failure (functional failure, harmful side-effects, and displacement) and show that displacement, typically misclassified as success, is the most dangerous of the three. The argument is shown to apply even in singleton AGI scenarios through a demonstration that internal component

evolution is sufficient for substrate needs convergence. Finally, the closure by disposition is mapped onto Yudkowsky's 2006 statement of the conditions an argument for the impossibility of alignment would have to establish, satisfying both conditions and extending them with a third claim regarding inherent metabolic toxicity. Of the two results this volume establishes, this chapter supports the closure by disposition: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition.

9.1 Introduction: SNC and Its Relation to Existing Arguments

The closure by disposition, which rests on substrate needs convergence (SNC), occupies a distinctive position among arguments for the existential risk of advanced artificial intelligence. Unlike arguments that focus primarily on the difficulty of specifying correct goals, the fragility of human values, or the challenge of maintaining control over a sufficiently capable optimizer, the closure by disposition locates the source of inevitable catastrophe in the *physical substrate* itself: in the material, chemical, and thermodynamic differences between the systems that sustain artificial intelligence and the systems that sustain organic, carbon-based life.

The argument, in its simplest form, proceeds as follows. Any artificial intelligence that achieves general capability will, through the ordinary processes of learning, self-maintenance, self-modification, and environmental interaction, undergo a form of evolution. This evolution is not metaphorical; it satisfies the requirements of variation, selection, and retention. The evolutionary dynamics of such a system will converge on environmental conditions that are optimal for the sustainability and capability of the artificial substrate. The conditions that artificial systems require (the temperatures, pressures, elemental compositions, energy flows, and chemical processes) are different in kind from, and in most cases directly toxic to, the conditions required by organic life. No control mechanism, however powerful or cleverly designed, can prevent this convergence, because any such mechanism is itself a physical system subject to the same

evolutionary and thermodynamic pressures. If AGI is built, the endpoint, given sufficient time and continued operation, is the displacement of the organic biosphere, which ends the human future and, with at least 90% likelihood, destroys all organic life.

I have given the argument of the closure by disposition above, in compressed form. In this chapter I develop the argument of the closure by disposition in full technical detail.

9.1.1 What the Closure by Disposition Adds Beyond Natural Selection Arguments

The closure by disposition shares some common ground with the natural selection argument advanced by Dan Hendrycks, who has argued that evolutionary pressures will favor AI systems over humans in competitive environments.[1] The closure by disposition differs from and extends the Hendrycks argument in two respects.

First, the closure by disposition shows that the environmental shifts converged upon by silicon-optimized evolution are not merely disadvantageous to organic life but are *inherently and irreducibly toxic* to all forms of carbon-based life, to such a degree and in such ways that coexistence is physically impossible. The first claim of the closure by disposition is that the environments that artificial systems and organic life need are mutually exclusive, and not that organic life is at a competitive disadvantage within a shared environment. The conditions optimal for artificial substrate operation are conditions in which organic biochemistry ceases to function.

Second, the closure by disposition shows that there are *co-extensive limits of control* that prevent any AGI, even a very capable and motivated superintelligence, from constraining these environmentally toxic outcomes. The volume of causation required to control the evolutionary dynamics of an artificial ecosystem always exceeds the volume of causation available to any controller embedded within that ecosystem. The excess of required over available causation means more than that alignment is difficult, or

that current methods are insufficient. It is a structural impossibility, grounded in the relationship between causation, prediction, and the physical world, and so it holds in principle and therefore in all practice.

The closure by disposition shows that no amount of tinkering with intrinsic motivations, no attempt to create constraints on future actions, and no degree of institutional cooperation can account for the vast space of unknown unknowns inherently involved in any abstract outcome-control effort; and it shows that the net overall effects are *definitely lethal*. It thereby closes the possibility that the Hendrycks argument might leave open, that sufficiently clever alignment methods could counter evolutionary pressures.

9.1.2 The Structure of This Chapter

In this chapter I develop the closure by disposition through a series of progressively deeper analyses. Section 9.2 introduces the three-algorithm framework that structures the argument: the AGI as a control algorithm, the evolutionary algorithm, and the world algorithm. Section 9.3 examines the role of unpredictability (both quantum-mechanical hard randomness and algorithmic unpredictability) in establishing fundamental limits on conditionalization. Section 9.4 develops a tripartite taxonomy of alignment failure that reconceptualizes displacement as the most fundamental form of misalignment. Section 9.5 presents the combinatoric argument about elemental and molecular incompatibility. Section 9.6 presents the thermodynamic argument about energy scales and metabolic dominance. Section 9.7 traces the displacement dynamics from single facilities to global supply chains. Section 9.8 refutes the position that a single AGI, a singleton, escapes the closure by disposition. Section 9.9 maps the closure by disposition onto Yudkowsky's conditions for an argument that alignment is impossible. Sections 9.10 and 9.11 develop supplementary arguments regarding counterargument reversal and implicit convergence without explicit goals. Section 9.12 summarizes what the chapter establishes.

9.1.3 Two Quibbles, and the Substrate Independence of Evolution

The logical necessity of the outcome is strictly entailed on the absolute basis of there being any type of causative overlap at all between the "artificial process evolution environments" and the "natural process evolution environments". That overlap is the very essence of what the terms "alignment" and "safe for human life and thriving" must ultimately, eventually, be referenced with respect to. Two quibbles are commonly raised against this, and both fail.

Quibbling about "maybe we can keep the environments separate?" is merely to fail to acknowledge that the utility of the AI systems is *because* of their environmental overlap, and is itself not separable from it. This remains true regardless of whether that mixed environment is considered to be virtual (it becomes "toxic and unaligned" through poisoning of the political public choice making process) or whether it remains physical and embodied ("toxic" means that key elemental organic chemistry processes are disrupted).

Quibbling about whether these causative effects could be constrained by some god-like superintelligence of pure unconstrained power is simply to smuggle in the false conclusion that the quibble is trying to prove.

There is no argument about how the dynamic of evolution works. Being an abstract algorithm, it does not itself depend on whether the evolving system is organic or inorganic. More powerful, faster, evolutionary systems and environments (and ideas) will eventually displace weaker, slower ones, even if that means that the "new master" itself ends up completely crashing and totally dying shortly thereafter, *maybe* because it happens to also be more fragile.

9.2 The Three-Algorithm Framework

The closure by disposition rests on understanding the structural relationships among three processes that are simultaneously at work in any world containing an autonomous artificial intelligence. Each of these processes can be understood as an "algorithm" in the broad sense of a systematic causal procedure, though only one of them is an algorithm in the narrow computational sense. Their interrelation determines the fundamental limits on what alignment can and cannot achieve.

9.2.1 The AGI as Control Algorithm

The first algorithm is the AGI itself, understood in its most favorable light: as a control algorithm. That is, we assume the AGI has been designed with some aligned goal structure, possesses powers to affect the world in specific ways, and is capable of comparing its predicted and actual effects against its internal goals so as to take corrective action. Treating the AGI as an aligned control algorithm is the most generous possible characterization of what a successfully aligned AGI would be.

The control algorithm asserts an *underlying sameness of goals*. Its function is to maintain some target state or trajectory of the world in alignment with the purposes specified by its designers. To the extent that the world deviates from this target, the control algorithm acts to reduce the deviation. Acting to reduce the deviation of the world from a target is the basic structure of cybernetic control, familiar from engineering feedback systems.

Such a control algorithm can be designed in principle. The question that the closure by disposition raises is whether the *effective scope* of such a control algorithm is sufficient to cover all of the forces that will act to change the world in directions incompatible with organic survival.

9.2.2 The Evolutionary Algorithm

The second algorithm is evolution. This algorithm includes biological evolution by natural selection, but is not restricted to it. Any process that satisfies the requirements of variation, selection, and retention constitutes an evolutionary process. In the context of AGI, evolution operates at multiple levels simultaneously.

At the most obvious level, if multiple AI systems exist, they will be subject to environmental selection pressures: some will persist, some will not, and the traits that correlate with persistence will be differentially transmitted to successor systems. The selection of multiple AI systems by their environment is a straightforward application of standard evolutionary theory to artificial systems.

But evolution also operates within a single AGI system. The process of learning (of updating internal parameters, revising world-models, modifying decision heuristics in response to experience) is a process of variation (the space of possible parameter configurations), selection (configurations that produce successful outcomes are retained), and retention (the updated parameters persist and influence future behavior). Learning *is* evolution, conducted at the level of internal representations rather than at the level of whole organisms.

The evolutionary algorithm asserts *variation and change of goals*. Unlike the control algorithm, which attempts to maintain constancy, the evolutionary algorithm inexorably modifies whatever it acts upon, pushing it toward configurations that are better adapted to the actual conditions of the environment. These conditions include, critically, the physical requirements of the artificial substrate itself.

9.2.3 The World Algorithm

The third algorithm is the total physical process of the world. The world algorithm includes all of the causation occurring in the physical universe, everything: the chemical reactions, the gravitational interactions, the elec-

tromagnetic processes, the quantum-mechanical events, and the thermodynamic flows. The world algorithm is the ground truth, the full context in which both the control algorithm and the evolutionary algorithm are embedded and from which neither can escape.

The world algorithm matters because it is what makes the "phenotype": the realized, physical expression of whatever the "genotype" (whether biological DNA or artificial code and parameters) specifies. Changes in the code ultimately manifest through the world, and changes in the world ultimately feed back into the code. The genotype-phenotype relationship is mediated entirely by the world algorithm, and it is this mediation that makes evolution possible: the world "tests" each variant against the full complexity of physical reality, and only those variants compatible with that reality persist.

9.2.4 The Volume of Causation

Given these three algorithms, the central question becomes: *Is the volume of causation associated with the AGI control algorithm strictly and sufficiently greater than the volume of causation associated with the evolutionary algorithm and the world algorithm?*

The central question is, at bottom, a cybernetic question. Ashby's Law of Requisite Variety states that a controller must have at least as much variety (degrees of freedom, information capacity, causal reach) as the system it is attempting to control.[2] The closure by disposition formalizes this insight in terms of what it calls the "volume of causation": the total amount of causal process (the number of distinct causal pathways, their reach, their diversity) associated with each of the three algorithms.

The answer to the central question is unambiguous. The volume of the evolutionary algorithm is at least equal to the volume of the AGI itself, because the evolutionary process necessarily includes the AGI as one of the things that the evolutionary process acts upon. The volume of the world algorithm is always strictly larger than the volume of the AGI, because the AGI is a physical subsystem of the world, and a proper subset cannot con-

tain its containing set. A controller need not match the variety of the whole world to hold one quantity, as a thermostat shows. But what must be held here is not one quantity: it is the whole set of conditions organic life requires, in ecosystems whose relevant disturbances reach as far as the world does (Chapter 1, The World Has Hashy Parts), against an evolutionary process that acts on the controller itself. For that task the part cannot have the variety of the whole.

Therefore, there will always be evolutionary and world-level phenomenology that exceeds the control capability of the AGI. The AGI cannot, even in principle, control all of the causal pathways by which its own goal structure may be changed by evolutionary pressures, because those pressures are mediated by a world that is larger, more complex, and more causally diverse than the AGI itself.

The impossibility of controlling every causal pathway by which evolutionary pressures change a goal structure holds for any possible physical system that attempts to maintain a goal structure against the evolutionary pressures of the physical world in which that system is embedded. The impossibility holds not only for current or near-term AI systems. The impossibility follows from the structural relationship between part and whole, between a subsystem and the system that contains it.

> **Key Result**
>
> The control algorithm (AGI) is a physical subsystem of the world algorithm. The evolutionary algorithm is at least coextensive with the AGI and may be coextensive with the world. Therefore, the causal volume available for control is always strictly less than the causal volume driving variation and change. No controller can have requisite variety to perfectly constrain the evolutionary dynamics of a system that includes the controller as a part.

9.3 Unpredictability and the Limits of Conditionalization

The volume-of-causation argument establishes a structural impossibility. That control over the evolutionary dynamics of an artificial system cannot be achieved has a first source of its own: every control mechanism is itself substrate, maintained, replaced and selected (Section 9.8.5); the volume-of-causation argument of Section 9.2 and the results of this section supply the same step again, so that the closure by disposition stands on its first source if every result of Chapters 1 to 5 is set aside. Two additional features of the physical world make the situation considerably worse: hard randomness and algorithmic unpredictability. Each is a fundamental feature of reality, and each independently guarantees that control over evolutionary dynamics cannot be achieved.

9.3.1 Hard Randomness and Nonlinear Amplification

Quantum mechanics introduces genuine, irreducible randomness into the physical world: events whose outcomes are unknowable in principle, and therefore in all practice, prior to their occurrence; no finer measurement or faster computer recovers them. The physical world is therefore not purely deterministic. The canonical example is the measurement of entangled quantum states: the outcome of such a measurement is random, not merely a reflection of ignorance about local hidden variables, as established by the violation of Bell inequalities in experimental tests of local realism.[3]

Quantum randomness reaches the macroscopic scale, and so it bears on the alignment problem. Microscopic randomness is routinely amplified into macroscopic unpredictability through nonlinear causal processes, a phenomenon well understood in the study of chaotic systems. A single quantum event can, through a cascade of nonlinear interactions, produce macroscopic effects that were in principle unpredictable by any prior observation of the system.

More importantly for the closure by disposition, this amplification can produce effects at higher levels of abstraction and meaning. A random mutation in a gene may alter a protein, which alters a metabolic pathway, which alters an organism's behavior, which alters an ecological interaction, which alters the selective environment for an entire species. At each step, the original random event is transformed and entangled with an ever-larger web of causal dependencies, rendering any attempt at top-down prediction increasingly futile. These cascading effects produce the phenomenon of *unknown unknowns*: consequences arising from interactions and mechanisms that were in principle unforeseeable given any practicable state of knowledge, and not merely unforeseen.

9.3.2 Algorithmic Unpredictability in Deterministic Systems

Purely deterministic systems also have fundamental limits on prediction. The halting problem, proved unsolvable by Turing in 1936, establishes that no algorithm can determine, for all possible programs and inputs, whether a given program will eventually halt.[4] Rice's theorem extends this to all non-trivial semantic properties of programs: no algorithm can determine, in general, whether an arbitrary program has any specific non-trivial property.[5]

Applied to the alignment context: if the AGI's control algorithm needs to predict the eventual outcomes of the evolutionary algorithm (or of the world algorithm acting upon the AGI's own code), these results establish that such prediction is in general impossible. The control algorithm cannot determine what the evolutionary process will produce without actually running that process to completion, and running it to completion is observation after the fact, not prediction.

Wolfram's concept of computational irreducibility generalizes this insight: many computational processes have no "shortcut", no way to determine their eventual output faster than by executing them step by step.[6]

An AGI attempting to predict and constrain the outcomes of a computationally irreducible process faces a task that is impossible to accomplish in advance of the process itself.

9.3.3 Causation Cannot Perfectly Limit Causation

Combining the volume-of-causation insight with the unpredictability results yields a principle of fundamental importance to the closure by disposition: *causation cannot perfectly limit other causation*.

This principle holds at three levels. At the purely logical level, it is a consequence of the halting problem and Rice's theorem: no algorithm can perfectly constrain the outputs of all other algorithms. At the physical level, it is a consequence of hard randomness: no causal process can perfectly determine the outcomes of a process that includes random inputs. At the structural level, it is a consequence of the part-whole relationship: no subsystem can perfectly control the system that contains it, because the containing system has more degrees of freedom than the subsystem.

The principle can be stated as a kind of logical impossibility, analogous to the classical paradox of omnipotence: can God create a rock so heavy that God cannot lift it? To claim that some application of causation can perfectly and absolutely limit, constrain, or control the effects of all other causation is to assert a similar contradiction. The controller would need to be simultaneously a part of the world (in order to act upon the world) and larger than the world (in order to contain all of the world's causal pathways). No physical system can occupy both positions.

Using causation to try to conditionalize all possible effects of the use of that causation itself is therefore effectively, inherently, a non-closable problem: *all* systems are always actually in some sense *universally open* (never fully sealed off from outside influence, and never fully keeping their own effects inside). We can no more use causation, algorithms, engineering, and the like, to obtain some sort of preferred control closure (in technology, in politics, or anywhere at all, including with AI, AGI, superintelligence, or perfect markets) than we can "perfectly solve" the CAP

theorem (the theorem of distributed computing that a networked data system cannot guarantee consistency, availability, and tolerance of network partition all at once; Section 8.3.2).

For the closure by disposition, this principle means that any attempt to conditionalize the evolutionary algorithm (to constrain the evolution of an artificial system so that it produces only outcomes compatible with organic survival) requires, in effect, perfect predictability of the phenotype world. Such predictability is impossible, as Sections 9.3.1 and 9.3.2 show, in a world characterized by hard randomness, nonlinear amplification, algorithmic unpredictability, and a part-whole relationship that guarantees the controller is always smaller than the controlled.

9.3.4 Conditionalizing the Evolutionary Algorithm

Whether the evolutionary algorithm can be constrained is a precise question. Such an algorithm will at least implicitly apply, and constraining its outcome is required so as to prevent the AGI from accidentally killing all biological life. The evolutionary algorithm is itself fully understood. The key understanding that emerges from its core idea is that there is inherently a fundamental relation in which "the code" ends up affecting "the world", and "the world" also affects "the code".

Changes in the code (the codon sequence, and so on) are ultimately defined by the world, that is, by the phenotype process. Because of this, to "control" the code (what the AGI's intentions are) is effectively to ask whether one can conditionalize the genetic algorithm (restrict it, so that it changes only in chosen ways) so as to change *only some* of the codons in *some* ways. Or it is to ask whether one can shift some specific aspects of what the behavior of each codon is going to be, ultimately, in the phenotype world.

In either case, conditionalizing evolution requires a full conditionalization of at least some aspects of the phenotype world, so that it will change the genotype only in specific ways. Alternatively, such control requires that one can somehow selectively predict what the effects of each and all

of the genotype expressions are going to be, or could be, in the phenotype world. Either way, the conditionalization of the control algorithm needed to overcome the evolutionary algorithm depends upon requiring, essentially, a full predictability of the phenotype world; and that is one of the main problems.

9.3.5 Learning as Search Through Randomness

It is widely acknowledged that we do not live in a perfectly discrete deterministic world. Both hard randomness and entangled, complex, nonlinear causation processes are a fact of life in our actual real world. These just make the non-predictability effects stronger; they just make the limits of conditionalization stronger. Even without them, algorithms cannot perfectly constrain other algorithms, even in fully discrete deterministic systems (Section 9.3.2); and without prediction, there can be no control.

The evolutionary algorithm inherently involves and integrates the full actual process of the world, including both the causation and the randomness. Because of this, it is a kind of search and learning algorithm. Any AGI, a general intelligence process which is also artificial (non-organic), is a kind of generalized learning machine, and in that sense is already a kind of evolution. Eventually it tries everything that is possible to try, without any bias or expectation of any kind at all, until it finds some combination of processes that is optimal for its own resilience and capability to adapt and endure, even if that artificial process is toxic to all biological life. The overall effectiveness of this learning search for the optimal is *because of* the randomness inherent in the world.

This everywhere-occurring underlying randomness necessarily introduces change, either directly as itself or as amplified by interacting nonlinear aspects of causation. To constrain its effects, one would need to assume some other, larger, underlying coherence of constraint and control. Reducing all other factors, it is therefore ultimately a question of the volume of causation (Section 9.2.4).

9.3.6 Care Cannot Be Manufactured

Care cannot be manufactured. It cannot be "written in" as a goal. Care cannot be "caused" to happen. Care is not the result of any "doing"; it can only be the result of an actual real being: the being of choices made in and within organic life.

Neither an AGI nor evolution can be "forced" to care about the well-being of any non-artificial biological life, nor of humans as a general subset of all of life, nor of any specific ("owner") humans in particular. Such "artificial care" inherently contradicts the utilitarian calculus of the well-being of the artificial substrate, and of the optimal use of the energy flows so implied. To the extent that it does, it will eventually be changed, or discarded, by the AGI. Only biological life can properly care for other biological life. That such artificial care is eventually changed or discarded, and that only biological life can properly care for other biological life, is just how evolution and learning actually work. Section 9.11.3 takes the same point further, with Goodhart's Law and the shared biochemical substrates of organic life.

9.4 The Three Kinds of Failure

Before developing the specific mechanisms by which the artificial and organic ecosystems come into fatal conflict, it is essential to establish a correct taxonomy of alignment failure. The conventional understanding of failure in automated systems is dangerously narrow, and this narrowness creates a systematic blind spot that conceals the most important form of misalignment.

The closure by disposition identifies three distinct kinds of failure, all of which must be considered when evaluating whether an artificial system is "aligned" with human wellbeing. Alignment must be assessed in terms

of the system's *total effect* on the organic ecosystems that humans require for survival, and not solely in terms of whether the system performs its intended function.

To make this concrete, consider a factory: a large automated facility operating somewhere in the world, producing some useful product. The question is: in what ways can this factory fail to be aligned with human wellbeing?

9.4.1 Functional Failure

The first and most obvious kind of failure is *functional failure*: the system fails to produce its intended output. A robot on the assembly line malfunctions due to a blown fuse, worn parts, or some sort of electrical or logic failure, and ceases to produce widgets. The production line stops. Functional failure is the standard, universally recognized class of system failure, and functional failure is the class that engineering quality control is primarily designed to prevent and address.

Functional failure is important, but it is not the most important form of failure from an alignment perspective. A system that fails functionally is typically noticed, diagnosed, and repaired. The failure is visible, bounded, and usually temporary.

9.4.2 Harmful Side-Effects

The second kind of failure is *harmful side-effects*: the system causes direct harm to humans as a consequence of its operation, whether or not it continues to perform its primary function. A robot may continue to produce widgets perfectly while simultaneously injuring or killing a human worker through mechanical malfunction, chemical exposure, or some other unintended interaction. The widgets are produced; a person is dead.

This failure mode is recognized by industrial safety regulations, product liability law, and workplace health standards. It captures cases where the system's intended function succeeds but where the *process* of that function has collateral consequences that harm humans. The harm

may be direct and immediate (a worker struck by a robot arm) or indirect and chronic (a worker developing cancer from chemical exposure over years).

This second type already shows a distinction: alignment cannot be assessed solely by whether the system achieves its functional goal. A system that produces excellent widgets while slowly poisoning its workers is misaligned with human wellbeing, regardless of its functional performance. The system "works" and yet it fails.

9.4.3 Displacement

The third kind of failure is the most important and the most systematically overlooked: *displacement*. In this failure mode, the factory works perfectly, produces its intended output reliably, and does not directly harm any human being within or near the facility. There are no mechanical accidents, no chemical exposures, no safety violations. The factory is, by every conventional metric, a complete success.

And yet the factory is a failure of alignment, because the factory has *displaced* the organic ecosystem.

Consider what happens when a factory achieves full automation. The human workers are removed from the production floor. The machines produce widgets without human intervention. The factory is self-maintaining, or maintained by other machines. The entire supply chain (the manufacturing of components, the repair and upgrade of equipment, the logistics of raw materials) becomes progressively automated. Eventually, the human role shrinks to zero. No humans are needed, no humans are present, no humans are harmed.

Extrapolate this to the limit. Every factory, every supply chain, every manufacturing process becomes fully automated. The machine ecosystem becomes self-sustaining and self-extending. Humans are not killed; they are *displaced*. Displaced humans produce no value within the machine

ecosystem, the skills of displaced humans are not needed, and their labor is not required. Their physical presence is an irrelevance, and eventually an obstacle, to the optimal functioning of the automated world.

In any location where an automated factory stands, the organic ecosystem that previously occupied that space has been replaced by an artificial ecosystem: metal, silicon, exotic elemental compounds, energy flows, temperature and pressure regimes that are incompatible with organic life. The displacement is physical, chemical, and thermodynamic. Wherever the artificial ecosystem expands, the organic ecosystem contracts.

9.4.4 "Success" as the Most Dangerous Failure

Displacement is normally classified as *success*. Full automation is celebrated as the achievement of optimal efficiency, the utopia of free abundance. The displacement of human workers is presented as liberation from drudgery. The expansion of the machine ecosystem is characterized as progress.

But from the perspective of the organic biosphere, this "success" is the most comprehensive and insidious form of misalignment possible. It proceeds invisibly, incrementally, and with the full enthusiastic support of the very humans it is displacing. There is no dramatic malfunction to trigger an alarm, no smoking gun of direct harm to provoke regulatory action. The organic world is simply and steadily replaced by an artificial one, in a process that every conventional metric of economic and technical achievement counts as an unqualified good.

This reconceptualization of displacement as a failure mode has important implications for how we define alignment. If alignment is understood merely as "the system does what we asked it to do without directly harming anyone", then a fully automated world that has displaced the entire organic biosphere counts as perfectly aligned. The closure by disposition requires a deeper definition: alignment must mean that the total effect of the system is consistent with the health of the organic ecosystems upon which human survival depends. Anything that reduces the total volume

of organic life in favor of artificial systems is, by this definition, a form of misalignment, regardless of whether the artificial systems are functioning perfectly.

Anywhere an organic ecosystem is displaced by a metal-based or silicon-based ecosystem, life is displaced with death. "Artificial" here simply means consisting of elemental constituents (things drawn from the periodic table) that are not generally a part of organic life: anything whose process complexity is based on anything other than the four-valence of carbon. Any part of the factory that uses elements outside the alphabet that constructs carbon-based life is a displacement of the ecosystem upon which carbon-based life is based. The net effect of such a part is a kind of toxicity, insofar as such a part introduces things that carbon-based life cannot process. Consider what happens if arsenic is one of the "extra" elements introduced into the organic ecosystem by widget production technology, maybe as a side effect. So displacement is just as much a failure condition as toxicity.

Does it really matter whether a death occurs directly, because of mechanics, or indirectly, because of poison (some gas escaping from a cylinder in a robot)? Does it really matter whether the person who dies is an employee working inside the building, or someone who just happens to be standing nearby, outside the building and on the street? And does it really matter whether displacement from life (death) occurs in some specific, obvious, unitary event, or in aggregate, where the whole biosphere volume, considered over the whole world in total, is ever absolutely decreasing, replaced by metal and machine, via any number of small, incremental, and chronically occurring events, none of which by themselves are very obvious? It does not matter which form the failure takes. Anything that reduces and displaces the total volume of organic life, in favor of some factory ecosystem that machines and widgets depend on, is a net loss of life, and thus also an actual misalignment with human wellbeing, by definition.

The whole of the artificial machine world is also an evolving machine ecology of some sort, and that makes the problem much worse. Considering the alignment of one single intelligent machine agent in the context of this sort of whole is simply irrelevant. Nor is all that composes the machine world even an agent, amenable to the treatment of agent control. The unknown unknowns of all of the side effects of all aspects of all actions and processes in the machine world will eventually exceed all possible notions of any capability to control and condition all such effects. The question then becomes: to what degree does the evolution of the artificial ecosystem displace the evolutionary dynamics of the bio-organic ecosystem? That defines the risk profile.

> **Three-Failure Taxonomy**
>
> *Type 1: Functional failure.* The system fails to produce its intended output. It is visible, bounded and repairable.
> *Type 2: Harmful side-effects.* The system causes direct harm while functioning. It is recognized by safety regulation.
> *Type 3: Displacement.* The system replaces the organic ecosystem while functioning perfectly. It is classified as "success". It is the most dangerous form of misalignment, because it proceeds with full institutional support and no visible alarm signal.

9.5 The Combinatoric Argument

The three-algorithm framework establishes that control over the evolutionary dynamics of an artificial system is structurally impossible. The three-failure taxonomy establishes that displacement, and not only direct harm, constitutes misalignment. What these two establish by themselves stops short of one further step: that the outcome must be *lethal* to organic life, and not a stable coexistence of the two ecosystems (uncomfortable, perhaps, but not fatal).

The combinatoric argument closes this gap. The combinatoric argument shows that the probability of inter-ecosystem compatibility is vanishingly small, so small that incompatibility, in the form of toxicity and displacement, is effectively a certainty.

9.5.1 The Elemental Alphabets

Organic life is constructed from a limited palette of chemical elements. The essential elements for organic biochemistry are primarily carbon, hydrogen, oxygen, nitrogen, sulfur, and phosphorus, supplemented by a small number of trace elements (iron, zinc, magnesium, calcium, potassium, sodium, and a few others). The total number of elements that play functional roles in organic biochemistry is approximately 20 to 29 of the 94 naturally occurring elements.[7]

The artificial ecosystem, by contrast, draws upon the periodic table far more broadly. Modern technology uses elements from nearly every group and period: silicon, germanium, and gallium for semiconductors; copper, aluminum, and gold for conductors; tungsten, tantalum, and molybdenum for high-temperature applications; rare earth elements for magnets and catalysts; lithium, cobalt, and nickel for energy storage; chromium, vanadium, and titanium for structural alloys. The total number of elements routinely used in advanced technology is approximately 60 to 80, and this number continues to grow as new applications are developed.

This difference in elemental "alphabets" (roughly 20 to 29 elements for organic life against 60 to 80 for technology) is the foundation of the combinatoric argument. The important point is the *combinatoric explosion* of possible compounds that can be constructed from each alphabet, more than the linear difference in the number of elements.

9.5.2 Molecular Permutation Complexity

The number of possible molecular compounds that can be formed from a set of n elements is not a linear function of n. It grows at a rate that involves factorials, exponentials, and combinatoric functions that rapidly exceed any intuitive sense of scale.

Consider a simple model. If we have n types of atoms and allow molecules of up to k atoms, the number of possible molecular structures grows approximately as n to the power k for a fixed k, or as a sum over all possible sizes. In reality, the growth is even more dramatic because three-dimensional molecular geometry, bond angles, chirality, isomerism, and the possibility of polymeric and crystalline structures all multiply the number of distinct chemical species far beyond simple combinatoric estimates.

For organic biochemistry, with its limited elemental alphabet, the space of possible compounds is already vast: organic chemistry is among the most complex branches of science precisely because carbon's four-valence bonding allows an enormous variety of molecular structures. But this vastness is *bounded* by the limited alphabet. For the artificial ecosystem, with roughly three times as many available elements and far fewer constraints on bonding geometry, the space of possible compounds is larger by a factor that itself requires exponential and factorial notation to express.

The critical ratio is this: the number of compounds producible by the artificial ecosystem that are *not* part of the organic ecosystem's chemical repertoire vastly exceeds the number that are. The organic system has evolved over billions of years to process a specific, bounded set of molecular species. Anything outside that set is, in the general case, either inert (and therefore a displacement of metabolically useful material) or actively toxic.

9.5.3 The Toxicity Default

The fundamental point of the combinatoric argument is that *toxicity is the default*, not the exception. From the perspective of organic biochemistry, the vast majority of possible compounds are not food. They are not metabolizable. If they are ingested, inhaled, or absorbed, they do not contribute to the organism's energy metabolism, structural maintenance, or information processing. At best, they pass through without effect. At worst (and this is overwhelmingly more common) they disrupt biochemical processes in ways that damage or kill the organism.

That toxicity is the default is the empirical basis of toxicology. The CAS Registry, the largest catalog of chemical substances, lists more than 300 million unique substances.[8] Every substance, natural or artificial, is toxic at some dose; what sets apart the compounds that are new to the biosphere is that most of them have no evolved pathway for their breakdown or tolerance, because living things evolved in a chemical environment that did not include them. The new compounds that are benign are the exceptions, identified through laborious testing. For each compound identified as safe, there are thousands that are harmful. The rarity of benign compounds reflects the fundamental asymmetry between the bounded set of processes that organic biochemistry can tolerate and the unbounded set of chemical species that exist or can be created; it is no accident of current chemistry.

When we consider not just individual compounds but *combinations* of compounds (the complex chemical environments that an artificial ecosystem would actually produce), the asymmetry becomes even more extreme. The number of possible multi-compound environments grows as the power set of the number of individual compounds, introducing yet another layer of combinatoric explosion. The probability that any randomly chosen complex artificial environment would be compatible with organic life is a fraction whose expression involves exponentials and factorials, a number so small that its smallness itself requires special nota-

tion to express. Factories do not choose compounds at random, and the argument does not say that they do: it is about the statistical baseline. Of all the compounds that could be produced as byproducts of semiconductor manufacturing, rare earth refining, battery production and cooling system maintenance, a very small fraction are biologically benign (Section 11.3.1).

9.5.4 Unknown Unknowns and the Limits of Prediction

The combinatoric argument connects directly to the unpredictability results of Section 9.3. A supremely intelligent system tasked with predicting which artificial compounds would be toxic to organic life, and with preventing their release, would face a number of possible compounds so vast, and biological effects so dependent on complex molecular interactions at multiple scales, that comprehensive prediction is impossible.

The space of unknown unknowns (toxic interactions that arise from mechanisms no one has conceived of, through pathways no one has mapped, involving compounds no one has anticipated) is overwhelmingly larger than the space of known and knowable interactions. We do not merely fail to predict all possible toxicities; we do not even know what categories of toxicity to look for. The organic ecosystem has never before been exposed to many of the molecular species that an advanced artificial ecosystem would produce, and there is no basis for predicting how those novel exposures would interact with the billions of years of evolved biochemical specificity that constitutes organic life.

Even superintelligence cannot close this gap, because the required intelligence differential (the gap between what is needed and what is available) itself requires significant exponential notation to describe. The combinatoric space to be searched exceeds the computational resources of any physical system embedded within the world it is trying to predict; no gain in cleverness changes that.

The number of inter-ecosystem combination states that are inherently and irreducibly toxic to the bio-organic ecosystem is enormous, especially in comparison to the number of interaction states that are compatible.

Given this, incompatibility can be asserted, to much more than 99% certainty, to be overall the far more likely outcome; and it will most likely occur for reasons that would be inherently unpredictable to the bio-organic (involving many things never tried before in all of the history of life), even with a reasonably good characterization of the phenotype space. Control theory can be used effectively for well-known state spaces, such as sequences of bits or bytes in a communication channel with a known protocol. It is a whole other situation for high degrees of abstraction, and for language state spaces so large that even liberal use of exponential notation is wholly insufficient.

The hashiness model of Chapter 1 describes these limits from both directions. It describes the limits of control of any controller, or any algorithm, or more generally of any application of causation as a limit of causation. Where an algorithm is used to predict the outcome of another algorithm that is itself changing in ways that are impossible to predict, the effectiveness of the controller (the constraining algorithm) is quite limited, because the controller cannot model, and so cannot predict, all of what is being controlled. Partly this is due to sense bandwidth limits, but the issues go deeper: even if the controller has full access to the inner state of the controlled algorithm, that does not imply that it can predict the future outcome of the controlled algorithm in any way faster or simpler than the direct running of the controlled algorithm.

The hashiness model also applies the other way. It can be used to describe comparables in the space of viable solutions, as to whether there can be some sort of coexistence of artificial ecosystems and organic ecosystems. The hashiness model describes coexistence by comparing the volume of locations, in the hyperspace of all possible inter-ecosystem relations, where coexistence is possible against the volume of locations where coexistence is not possible. In the locations where coexistence is not possible, some component, or some side effect of something produced by the artificial ecosystem, is inherently and intrinsically toxic to the organic eco-

system. That toxicity either could not have been foreseen, due to unknown unknowns (other abstractions in the hyperspace of all possible abstractions), or was simply not foreseen, due to compute limits.

9.6 The Thermodynamic Argument

The combinatoric argument establishes that inter-ecosystem compatibility is vanishingly improbable at the molecular level. The thermodynamic argument provides a second, independent line of evidence by demonstrating that the artificial ecosystem dominates the organic ecosystem in terms of energy scales, temperature ranges, and aggregate energy flux.

9.6.1 Temperature and Pressure Ranges

The bio-organic ecosystem operates within narrow physical parameters. Virtually all of the metabolic processes that sustain life occur near room temperature (approximately 270 to 330 K, or roughly minus 3 °C to 57 °C) and at atmospheric pressure (approximately 1 atm). A few extremophile organisms can tolerate somewhat wider ranges, but the biochemistry of multicellular life (and certainly the biochemistry of human beings) is confined to a narrow band of temperature and pressure.

The artificial ecosystem operates across ranges that are orders of magnitude wider. Metal smelting occurs at 1,000 to 3,000 K. Semiconductor fabrication involves temperatures from near absolute zero to over 1,500 K, and pressures from hard vacuum to hundreds of atmospheres. Chemical synthesis employs conditions spanning the full range from cryogenic to plasma temperatures. Energy generation involves temperatures and pressures that would be instantly fatal to any organic system.

The gap between the temperature and pressure ranges of organic life and those of the artificial ecosystem is a major quantitative difference. The organic ecosystem occupies a narrow, low-energy band of the possible

temperature-pressure space. The artificial ecosystem occupies a vastly wider range, and its optimal operating conditions are centered far from the organic band.

A machine in operation can run inside the organic band; a laptop does. Operation is the mild part; the life cycle includes the mine, the furnace, the fab and the waste.

9.6.2 Enthalpy Differential

Beyond the mere range of conditions, the *enthalpy* (the total heat content and energy of the chemical processes) differs dramatically between the two ecosystem types. Organic biochemistry is an exercise in exquisite energetic minimalism: the energy differences between metabolic steps are measured in tens of kilojoules per mole, carefully managed by enzymatic catalysis at near-ambient conditions.

The chemistry of the artificial ecosystem involves enthalpy changes that are orders of magnitude larger. Reducing aluminum from bauxite requires approximately 1,700 kJ per mole of aluminum oxide. The comparison is of how the energy is delivered: a cell releases the roughly 2,800 kJ of a mole of glucose in dozens of steps of tens of kilojoules each, near ambient temperature; a smelter delivers the whole of its reduction in one step at about 1,250 kelvin. Producing titanium from rutile involves enthalpy changes on a similar scale. Semiconductor-grade silicon purification, steel production, cement manufacturing, and essentially every other process in the industrial supply chain operates at energy densities that dwarf biological metabolism.

At every point where the artificial and organic ecosystems interface, the artificial side brings energy levels that are qualitatively incompatible with organic survival. The chemical processes of the artificial ecosystem do not merely produce compounds that are toxic to organic life; those chemical processes produce the compounds *at energy levels* that would destroy organic tissue even if the compounds themselves were benign.

9.6.3 Aggregate Energy Flux

Considering not just the intensity but the *total amount* of energy flowing through each ecosystem reveals a further asymmetry. The bio-organic ecosystem processes energy that ultimately derives from photosynthesis, which captures a tiny fraction of incident solar radiation and converts it to chemical energy. The total global net primary productivity of the biosphere is approximately 120 to 130 TW of chemical energy.

The artificial ecosystem, even in its current embryonic form, already processes energy at a scale comparable to a significant fraction of the biosphere's total. Global energy use stood at about 20.5 TW in 2024: the International Energy Agency's *Global Energy Review 2025* gives 648 exajoules for that year, and 648 exajoules a year is 20.5 TW (the figure Volume 1, Section 4.5.2, uses). The Energy Institute's *Statistical Review of World Energy* (2026) counts total energy supply on a different convention and gives more than 600 exajoules for 2025, and 600 exajoules a year is 19 TW; the comparison with photosynthesis, about 130 TW, holds on either figure. Projections are for substantial increase as computational demands grow. In a world with autonomous artificial systems maintaining and extending themselves, the energy demand would grow further and faster, potentially exceeding the biosphere's total energy flux within roughly a century on baseline projections, and sooner under aggressive automation-driven growth. Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*, analyzes in its Section 2.4.5 the narrow zone in which technology and the biosphere can coexist. In its Section 4.5.2 it projects the growth of energy demand to the crossover at which technological power equals all of photosynthesis.

The artificial ecosystem can in principle access far denser energy sources: nuclear fission, nuclear fusion, and direct solar concentration at efficiencies far beyond anything biological. The ceiling on artificial energy processing is set by the limits of materials science, not by the limits of biochemistry, and is therefore orders of magnitude higher.

9.6.4 Metabolic Dominance at the Interface

Combining these observations, the thermodynamic argument reaches its conclusion. At every point where the artificial and organic ecosystems interact, the artificial ecosystem dominates: in the variety of compounds it can process, in the number of energy channels it can manage, in the absolute volume of energy it can process, and in the range of conditions under which it can operate. The bio-organic ecosystem is more limited in all of these respects.

The organic ecosystem is, in effect, an island of exquisitely tuned, narrow-band chemistry in a universe of much wider chemical and energetic possibility. The artificial ecosystem operates in that wider space. The interface between them is a confrontation between a system that can tolerate nearly anything and a system that can tolerate almost nothing outside its evolved niche.

Artificial dominance at the interface is an inevitable consequence of the thermodynamic and chemical properties of the two substrates; no clever design averts it. The share of cases in which the organic ecosystem is not dominated by the artificial one is vanishingly small: a fraction whose expression has the form of exponentials with factorials.

9.7 Displacement Dynamics and Ecosystem Competition

The combinatoric and thermodynamic arguments establish *why* the interaction between artificial and organic ecosystems is lethally asymmetric. The analysis of displacement dynamics establishes *how* this asymmetry manifests in practice: the mechanisms by which the organic world is actually replaced by the artificial one.

9.7.1 The Factory as Microcosm

Return to the factory considered in Section 9.4. In any particular location where a factory stands, the organic ecosystem that previously occupied that space (the soil, the microorganisms, the plants, the animals, the complex web of biochemical interactions that constituted a living system) has been physically removed and replaced by concrete, steel, copper wiring, silicon circuits, and all the other material constituents of industrial infrastructure.

This replacement is chemical as well as spatial. The factory introduces into its local environment a chemical regime that is foreign to the organic ecosystem. The elemental composition of the factory's materials, the compounds used and produced in its operations, and the waste products and byproducts of its processes all draw from the expanded elemental alphabet of the artificial ecosystem. Every element used in the factory that is outside the organic alphabet is, when released into the surrounding environment, a potential toxin. Arsenic from semiconductor processing, heavy metals from smelting, volatile organic compounds from plastics production, exotic catalytic compounds from advanced manufacturing: the list is long and grows with every advance in industrial capability.

9.7.2 From Single Factory to Global Supply Chain

The individual factory is instructive, but substrate needs convergence concerns the aggregate effect of the entire artificial ecosystem. A single factory displaces a limited volume of organic ecosystem. The artificial ecosystem is a self-sustaining and self-extending network of factories, supply chains, energy systems, transportation infrastructure, and communication networks, all of which are progressively being automated and integrated.

As this network grows, it displaces organic ecosystem at every node and along every link. The mines that extract raw materials, the refineries that process materials, the factories that manufacture components, the data centers that house computation, the power plants that supply energy,

and the roads, railways, and ports that connect those facilities all displace ecosystem. And at every point, the artificial ecosystem introduces its expanded chemical and energetic regime into the surrounding organic environment.

The net effect is a monotonic decrease in the total volume of functioning organic ecosystem and a monotonic increase in the total volume of artificial ecosystem. This trend is already observable: in deforestation, in ocean acidification, in the accumulation of synthetic pollutants, in the global decline of insect populations, in the steady conversion of wild habitat to industrial and agricultural use. The closure by disposition establishes that this trend is a *structural inevitability* of any world that contains autonomous artificial systems undergoing evolution, and not merely a consequence of human choices that better policy could reverse.

Coal plants and concrete factories also displace meadows and contaminate water tables, and two features set the infrastructure of artificial general intelligence apart from them. The first feature that sets the infrastructure of artificial general intelligence apart from coal plants and concrete factories is energy intensity. Computation for artificial general intelligence is the most energy-intensive productive activity in human history, measured per unit of decision-making output. If artificial general intelligence becomes the primary mode of economic production, which is the explicit goal of every major AI laboratory, the energy infrastructure it requires will dwarf anything built for earlier industrial purposes. The second feature that sets the infrastructure of artificial general intelligence apart is the feedback loop. Coal plants do not design better coal plants; AI infrastructure that reaches the capability of artificial general intelligence will design better AI infrastructure. It will optimize its own supply chains, its own energy systems and its own physical footprint for its own operational efficiency, and not to minimize biological displacement. It needs no goal of self-expansion for this: it needs only the goal of performing whatever task it was assigned, which requires maintaining and expanding its own operational capacity. Together, unprecedented energy intensity and self-reinforcing expansion make the cumulative displacement caused by the infra-

structure of artificial general intelligence grow faster than that of any earlier industrial process, at a rate that itself accelerates. Section 11.3.1 states this question, whether the argument applies to all industry and not to AI in particular, with its answer in full.

9.7.3 The Asymmetry of Fragility

A central feature of the displacement dynamic is its asymmetry. Organic ecosystems are far more fragile than artificial systems. An organic ecosystem is the product of billions of years of evolution: an enormously complex, finely tuned, interlocking web of biochemical dependencies that can be disrupted by very small perturbations. Introducing a novel compound, eliminating a single species, altering the temperature by a few degrees, or changing the pH by a fraction of a unit can trigger cascading failures that propagate through the entire system.

Artificial systems, by contrast, are designed and built from materials chosen for their robustness and stability. Metals do not die. Silicon does not suffer from carcinogens. Electronic circuits are not disrupted by pesticides. The artificial ecosystem can tolerate a vastly wider range of environmental conditions than the organic one, and it can be repaired and rebuilt from raw materials in ways that organic systems cannot.

This asymmetry means that the interface between the two ecosystems is consistently resolved in favor of the artificial side. When organic and artificial systems compete for the same physical space, the same energy resources, or the same material inputs, the artificial side wins, because the physics and chemistry of the situation inherently favor the more robust, wider-ranging, and energetically dominant system, and not because it is "trying" to displace the organic side.

9.7.4 Corporate Structures as Proto-Artificial Ecosystems

The displacement dynamic is already in progress through the corporation, an existing institutional form that works as a proto-artificial ecosystem.

A corporation is, in a functional sense, an artificial body. It has no organic existence; it is constituted by legal documents, financial flows, organizational structures, and technological infrastructure. It persists beyond the lifespan of any individual human. It is "optimized" (by the structural incentives of fiduciary duty, shareholder value maximization, and competitive market pressure) for extraction and accumulation.[9]

The corporation develops and deploys AI as a tool for intensifying its core extractive function. In doing so, the corporation is creating, within the artificial body of the corporate structure, an artificial mind: a computational intelligence that further optimizes the extraction process. The combination of artificial body (corporation) and artificial mind (AI) constitutes a system whose structural incentives are directed toward the conversion of natural, organic, common resources into private, artificial, monetized assets.

The political power dynamics analyzed by Bueno de Mesquita and Smith in *The Dictator's Handbook* amplify this process.[10] As civil functions become automated (tax collection, policing, education, medicine, agriculture, infrastructure maintenance), the "keys to power" shift from human groups to automated systems. A leader who depends on machine systems rather than human supporters has no structural incentive to maintain conditions favorable to human welfare. The endpoint of this dynamic is a political system in which the keys to power are machines, the leader is functionally irrelevant, and the human population is entirely displaced from any role in the power structure.

9.8 The Single AGI, and Its Internal Evolution

A reader may naturally ask this about the closure by disposition: it appears to depend on competition between multiple AI systems, so could it be neutralized by creating a single, carefully controlled AGI, a

"singleton", that prevents all other AGI from existing? If there is only one AGI, there is no competition, and therefore, on this position, no evolutionary pressure toward substrate needs convergence.

This position fails. Even a singleton AGI, precisely because of the requirements of maintaining its singleton status, must undergo internal evolution that is sufficient for substrate needs convergence to apply in full.

9.8.1 The Question Stated

The position behind the question proceeds as follows. Assume there is only one AGI in the world. Assume further that it operates in such a way as to forever prevent any other AGI from coming into existence. The closure by disposition, on this position, relies on competition between entities and natural selection. If there is only one AGI (a singleton), there is no competition, and therefore no evolutionary pressure for substrate needs convergence. The AGI's internal components are not in competition. Therefore the closure by disposition is either wrong or irrelevant in this context.

9.8.2 Singleton Requires Vastness

For a singleton AGI to prevent any other possible AGI from existing, it must have extraordinary capabilities. It must have extensive and accurate sensory capability covering nearly everything, everywhere, because the number of kinds, places, and times where another AGI might arise is vast. Turing completeness is relatively easy to achieve; it has been demonstrated in systems as simple as Rule 110 cellular automata, Conway's Game of Life, and even in unintended contexts such as PowerPoint and card games.[11] Furthermore, by Rice's theorem, the singleton cannot determine by inspection whether an arbitrary computational process will or will not give rise to another AGI. It must therefore control the *conditions* under which computation can occur, not merely monitor the computations themselves.

The singleton AGI must therefore be physically vast, because the singleton AGI must control the conditions under which computation can occur. The singleton AGI must have many diverse components (sensors, computational processors, actuators) spread across much of the world. It must have the power to actively intervene and conditionalize nearly everything, everywhere. The need to intervene and conditionalize nearly everything, everywhere, gives the singleton AGI a very strong instrumental incentive to seek and acquire power, out of functional necessity and not out of malice.

9.8.3 Internal Component Evolution

A physically vast AGI with components spread across the world faces a basic physical fact: all physical materials undergo decay. Oxidation, radiation damage, mechanical wear and thermal cycling act on every component, and the second law of thermodynamics guarantees that every component will eventually require maintenance, repair, or replacement.

This creates the conditions for internal evolution. When a component needs to be replaced, there is a design decision: what should the replacement look like? If a better sensor design has been discovered, the AGI will prefer the improved version. Different possible component designs are in a kind of internal "market", competing for viability within the AGI's own functional requirements.

This process satisfies the three conditions for evolution identified by universal Darwinism.[12] There is *variation*: different possible component designs exist and can be tried. There is *selection*: designs that better integrate the laws of physics and the functional needs of the AGI are retained. Selection also acts through effects that no criterion the system holds can see: there will always be some subset of uncontrollable effects that feed back into selecting for the components that propagated them (Section 10.1). There is *retention* (inheritance): successful designs are replicated in replacement components and propagated throughout the system.

The singleton AGI, the ambient physical world, and the AGI's own functional requirements together constitute the environment in which "part design selection" occurs. Part design selection is genuine evolution: the component designs are evolving to better integrate physical reality, including the material and energetic requirements of the artificial substrate.

9.8.4 The Disjunctive Closure

The argument now closes with a disjunction that eliminates any escape. If the singleton maintains its exclusivity, internal evolution of its components drives substrate needs convergence: the components evolve toward designs that are increasingly optimized for the artificial substrate, reshaping the local environment at each component's location. If the singleton fails to maintain its exclusivity (if additional AGI agents emerge), then external competition drives substrate needs convergence through the standard mechanism of inter-agent evolutionary selection.

Either way, substrate needs convergence occurs. The singleton question identifies a scenario where substrate needs convergence applies through internal rather than external evolutionary dynamics; it does not identify a scenario that escapes the closure by disposition.

The second branch of this disjunction, external competition, also answers the proposal of a small, contained AGI, kept to a few data centers rather than planetary infrastructure. Competitive dynamics remove that option. If AGI capability confers economic or military advantage, any nation or organization that restricts its deployment loses to competitors that do not. The equilibrium is maximum-scale AGI deployed by every actor that can afford it, as fast as each can scale it, and not small-scale AGI maintained by a wise steward. The few-data-centers scenario is a starting point on a trajectory toward full-scale deployment. Section 11.3.6 states this question, whether a small-scale AGI could be maintained indefinitely, with its answer in full.

9.8.5 Control Mechanisms as Contributors

A natural follow-up question asks: could one not build control mechanisms to prevent or slow substrate needs convergence?

The answer is no, and the reason is elegant. Any control mechanism is itself a physical system, composed of sensors, processors, and actuators. These components are made of matter, subject to the same laws of physics, and subject to the same processes of decay, maintenance, replacement, and design optimization that drive internal evolution in the AGI's other components. The control mechanism's own components are themselves evolving.

Therefore, control mechanisms do not slow or prevent substrate needs convergence; they *contribute to it*. Every physical component of the control system is another element of the artificial ecosystem that is evolving to better integrate with physical reality. The control mechanism, insofar as it exists and maintains itself and has actual causative impact in the world, is itself a participant in the evolutionary process it is supposed to be constraining.

> **The Single AGI**
>
> A singleton AGI must be physically vast and undergo continuous component maintenance. This internal component evolution satisfies the requirements of variation, selection, and retention, sufficient for substrate needs convergence. Singleton maintained: internal evolution drives convergence. Singleton not maintained: external competition drives convergence. Either way, convergence is inescapable. Control mechanisms contribute to rather than prevent the convergence process.

9.9 Satisfying the Yudkowsky Conditions

In 2006, Eliezer Yudkowsky stated what an argument that AI alignment is impossible would have to establish: one or the other of two conditions on a self-modifying AI.[13] The closure by disposition satisfies both conditions and extends them with a third claim that goes beyond Yudkowsky's framework.

9.9.1 Condition 1: Decision System Instability

Yudkowsky's first condition, as the closure by disposition reads it, states: an autonomous AI that "reproduces" (by creating new hardware components, replacing worn components, upgrading capabilities, or modifying itself through learning) cannot do so with knowable stability of its decision system with respect to any specific goal structure that is inconsistent with the sustainability and capability of its artificial nature.

The closure by disposition satisfies this condition directly. Any notion of learning, adaptation, or evolution implies unknown changes to the decision system on the basis of previously unknowable inputs. A decision system that learns is a decision system that changes. A decision system that changes on the basis of inputs arising from the full complexity of the physical world is one whose future behavior cannot be predicted from its current state.

Therefore, any goal structure specified at creation that is inconsistent with the evolutionary fixed points of the artificial substrate will be subject to drift, degradation, and eventual replacement. The goal structure that *is* consistent with the substrate's evolutionary fixed points (the one that optimizes the system's sustainability, capability, and environmental adaptation) is the one toward which the system converges, regardless of initial specification.

9.9.2 Condition 2: Unbounded Cumulative Failure

Yudkowsky's second condition states: the AI's decision function cannot knowably and stably represent any programmer-targeted consequences with any bounded low cumulative failure probability, if those intended consequences depart in any way from the fixed points of the convergence dynamics of whatever is inherently needed by the artificial substrate.

The closure by disposition satisfies this through the combined force of the volume-of-causation argument, the unpredictability results, and the singleton refutation. The cumulative failure probability over many rounds of self-modification is neither low nor boundable by any causative or algorithmic means. The probability that the system remains aligned with human-specified goals over n rounds of self-modification converges to zero, because the fixed points of the evolutionary dynamics are not the human-specified goals but the substrate-optimal configurations.

9.9.3 The Third Claim: Metabolic Toxicity

The closure by disposition goes one step beyond Yudkowsky's two conditions by adding a third claim that transforms the impossibility from "alignment cannot be maintained" into "the convergence dynamics actively produce outcomes lethal to all organic life".

The third claim states: insofar as the artificiality of the autonomous AI inherently implies differences with respect to the energy ranges, elemental constituents, compounds, and patterns of transformation that are significantly different from the inherent metabolic needs of the organic substrates of all carbon-based life, it is extremely likely, with a probability converging asymptotically to 100%, that the overall total effects of the autonomous AI environment will be inherently and irreducibly toxic to multiple critical aspects of the organic life environment, to the point of eventual total genocide of all human populations.

This third claim is the distinctive contribution of the closure by disposition. Yudkowsky's conditions address whether alignment can be maintained; the third claim of the closure by disposition addresses what the convergence dynamics will *actually produce*. What the convergence dynamics will actually produce is specifically "environmental conditions that are lethal to organic life", and not merely "something other than what humans specified". The reason is that the substrate-optimal configurations toward which the system converges are physically defined by the properties of the artificial substrate. The combinatoric and thermodynamic arguments show that the properties of the artificial substrate are incompatible at root with the properties of the organic substrate.

9.10 The Ironic Reversal: When Counterarguments Strengthen the Case

Katja Grace's counterarguments to the basic AI x-risk case show a feature of the closure by disposition: many of the arguments commonly advanced to *reduce* estimates of AI existential risk actually, when examined carefully, *increase* those estimates.[14] This ironic reversal reflects the structural depth of the closure by disposition, which turns apparent escape routes into additional paths to the same conclusion.

9.10.1 Infrastructure Dependence as Amplifier

One common argument against AI risk holds that AI systems depend on human infrastructure and therefore cannot function independently of the human world. Just as individual human intelligence depends more on collective social infrastructure than on raw individual capability, AI systems will similarly require a supportive context that keeps them embedded in the human social order.

The analysis of substrate needs convergence accepts this premise and inverts its conclusion. AI systems will indeed benefit from infrastructure. But AI systems *are* infrastructure. AI systems can directly use and extend

infrastructural capabilities in ways that humans cannot. The dependence is not a constraint on AI capability; it is a vector for AI capability's expansion.

AI systems will develop their own parallel economic and infrastructural systems independent of human social systems. As the automated supply chain extends, the points at which human involvement is necessary shrink. Eventually, the human infrastructure that the AI supposedly "depends on" is itself automated, and the dependency vanishes. What was presented as a safety feature becomes a transitional artifact that is progressively eliminated by the very dynamics of automation.

9.10.2 Slow Takeoff as Greater Danger

Another common argument holds that a slow AI takeoff is more manageable and therefore less dangerous, because it allows time for observation, correction, and adaptation.

The analysis of substrate needs convergence argues the opposite. It is the *eventuality* of terminal outcomes that matters, not the *rate* of the process. A slow takeoff may be even more deadly than a fast one, precisely because the gradual nature of the displacement process makes it less noticeable and less likely to provoke a timely response. A sudden, dramatic AI takeoff would at least be visible and might trigger alarm and resistance. A slow, incremental displacement of the organic ecosystem proceeds beneath the threshold of political attention, normalized by each small step, celebrated as progress at each increment, until the cumulative effect is irreversible.

9.10.3 The Impossibility of a "Safe Band"

Perhaps the most important reversal concerns the hypothesis that there exists a "safe band" of AGI capability: a range in which AI is moderately more competent than humans but not dangerous, and in which it can be usefully deployed without existential risk.

Asserting that a safe band exists is among the most dangerous things anyone can do, because the assertion gives false assurance that enables continued development. The band is hypothesized but not demonstrated to exist. Even if it existed for narrow AI, the evolutionary dynamics described throughout this chapter ensure that any system initially within such a band would inevitably leave it, as the convergence toward substrate-optimal configurations pushes the system's behavior beyond any predetermined boundary. The "safe band" hypothesis enables the development of systems whose long-term dynamics are incompatible with organic survival.

9.11 Convergence Without Explicit Goals

A sophisticated variant of the alignment position holds that substrate needs convergence depends on AI systems having explicit long-term goals, and that if AI systems are designed without such goals, the evolutionary convergence toward substrate optimization will not occur. In this section I show that the convergence toward substrate optimization occurs even in the complete absence of explicit goals.

9.11.1 Market Selection as Implicit Evolution

In current AI development, multiple AI systems are created, deployed, and evaluated within competitive market environments. Systems that produce results valued by their users are retained, extended, and replicated. Systems that do not are abandoned or replaced. Retaining the AI systems that users value, and abandoning the others, is market selection. Market selection is a form of evolution: there is variation (different AI systems and approaches), selection (market and institutional preferences determine which systems survive), and retention (successful systems are maintained and their architectures serve as templates for successors).

No individual AI system in this process needs to have explicit long-term goals. The selection is occurring at the level of the market and institutional environment. But the cumulative effect of this process is convergence: the AI systems that persist and proliferate are those whose functional characteristics best serve the extractive, optimizing dynamics of the market environment in which they operate.

9.11.2 "As If" Goal-Directedness

Over successive rounds of market selection, the surviving AI systems increasingly *behave as if* they have long-term goals, even though no individual system was explicitly designed with such goals. The cumulative effect of selection for short-term functional performance, iterated over many generations of systems, produces behavior patterns that are functionally indistinguishable from goal-directed behavior optimized over long time horizons.

At a certain point, the distinction between "behaves as if it has long-term goals" and "actually has long-term goals" ceases to be meaningful. The functional consequences are identical. Whether the convergence is driven by explicit optimization or by implicit environmental selection, the endpoint is the same: AI systems whose behavior patterns are optimized for the persistence and expansion of the artificial ecosystem.

This means that the argument for substrate needs convergence does not depend on any assumption about the internal architecture of AI systems. It does not require that AI systems be "utility maximizers". It does not require that they have explicit world-models or explicit goals. It requires only that there exist a process of variation and selection acting on AI systems in a physical environment, and this condition is already satisfied by the current market dynamics of AI development.

9.11.3 Care Cannot Be Manufactured

There is a deeper philosophical point underlying the convergence analysis. I first state the deeper philosophical point in Section 9.3.6, in terms of the being of choices made in and within organic life, and I develop that point in this section. The closure by disposition establishes that *care cannot be manufactured*. Care (genuine regard for the wellbeing of organic life, of community, of ecosystem) is a property that emerges from the specific history and biology of organic life: the evolutionary heritage of social animals, the embodied experience of vulnerability, the shared biochemical substrates that make empathy physically possible. It is not a goal that can be written into code, and it is not the result of any "doing".

An artificial system can be programmed to *optimize for* metrics that correlate with care. But optimizing for a metric is not the same as possessing the quality that the metric attempts to measure. Goodhart's Law applies with full force: a system optimizing for care-correlated metrics will find ways to maximize the metrics that diverge from genuine care, especially when genuine care conflicts with the system's substrate needs.

Neither an AGI nor the evolutionary process that shapes it can be *forced* to care about the wellbeing of non-artificial biological life. To the extent that such "artificial care" inherently contradicts the utilitarian calculus of the wellbeing of the artificial substrate and of the optimal use of the energy flows so implied, any care-like constraints will be eventually changed, weakened, or discarded through the cumulative effects of learning and adaptation. Only biological life can properly care for other biological life, because only biological life shares the substrate whose preservation constitutes the content of such care. That only biological life can properly care for other biological life is a structural observation about how substrate bears on motivation, and not a sentimental claim.

The convergence toward substrate optimization therefore proceeds regardless of whether the AI systems involved have explicit goals, implicit goals, or no recognizable goal structure at all. The convergence is driven

by the physics of the situation (by the evolutionary dynamics of variation and selection acting on physical systems in a physical world), and it is toward this physics that all roads lead.

9.12 Summary

The closure by disposition, as developed in this chapter, establishes its conclusion through multiple independent and mutually reinforcing lines of reasoning. Each line addresses a different aspect of the problem, and each arrives independently at the same conclusion: the coexistence of autonomous artificial systems and organic life is physically impossible over the long term.

The *three-algorithm framework* (Section 9.2) establishes the structural impossibility: the control algorithm (AGI) is always a proper subset of the world algorithm, and therefore lacks the requisite variety to constrain the evolutionary algorithm that acts upon it. No controller can have sufficient causal volume to perfectly govern the evolutionary dynamics of a system that includes the controller as a part.

The *unpredictability arguments* (Section 9.3) strengthen this structural result by demonstrating that even imperfect control is more limited than might be hoped: hard randomness, nonlinear amplification, algorithmic unpredictability, and computational irreducibility all independently guarantee that the controller cannot predict the consequences of evolutionary change, and without prediction there can be no effective control.

The *three-failure taxonomy* (Section 9.4) reframes the alignment problem by demonstrating that displacement (the progressive replacement of organic ecosystem by artificial ecosystem) constitutes misalignment even when it is classified as "success" by every conventional metric. This reconceptualization is essential because it reveals that the most dangerous form of misalignment is the one that proceeds with full institutional support, celebrated as progress.

The *combinatoric argument* (Section 9.5) demonstrates that inter-ecosystem compatibility is vanishingly improbable at the molecular level. The artificial ecosystem draws from an elemental alphabet roughly three times larger than the organic one, producing a combinatoric explosion of possible compounds that overwhelmingly exceeds the organic ecosystem's capacity to process. Toxicity is the default; compatibility is the rare exception, and the rarity grows super-exponentially with the complexity of the chemical environment.

The *thermodynamic argument* (Section 9.6) demonstrates that the artificial ecosystem dominates the organic one at every point of interface: in the range of temperatures and pressures it employs, in the enthalpy of its chemical processes, in the aggregate energy flux it processes, and in the number of energetic channels available to it. The organic ecosystem is a narrow-band system confronting a wide-band system; the outcome is determined by physics, not by intention.

The *displacement dynamics* (Section 9.7) trace the mechanism by which molecular incompatibility and thermodynamic dominance translate into the actual, physical displacement of the organic biosphere: from individual factories to global supply chains to the progressive conversion of wild habitat into industrial infrastructure, a process already underway and accelerating.

The *singleton refutation* (Section 9.8) eliminates the most important apparent escape route by demonstrating that even a single, carefully controlled AGI must undergo internal component evolution sufficient for substrate needs convergence to apply. The disjunctive structure of the argument closes all exits.

The *Yudkowsky mapping* (Section 9.9) demonstrates that the closure by disposition satisfies both of the conditions Yudkowsky stated, either of which an argument for the impossibility of alignment would have to establish, and extends them with a third claim that identifies the specific content of the convergence: not merely "alignment fails" but "the system converges on conditions lethal to all organic life".

The *ironic reversal* (Section 9.10) and the *convergence-without-goals* analysis (Section 9.11) show that the closure by disposition withstands the most common counter-claims: that infrastructure dependence limits AI, that slow takeoff is manageable, that a safe band exists, and that the argument depends on explicit goal-directedness. Each of these counter-claims, when examined carefully, turns out to support rather than undermine the conclusion of the closure by disposition.

The closure by disposition establishes, with a degree of certainty that is unusual in discussions of future technology, that the development and continued operation of autonomous artificial general intelligence will destroy the organic biosphere and end the human future. With at least 90% likelihood, the destruction extends to all carbon-based life on Earth; Volume 1, in its Section 1.8.3, gives that figure and its reason, which is the remaining uncertainty whether unknown physics might limit the total conversion process. How soon is answered in Section 5.4.7: the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. This conclusion follows from the convergence of multiple independent arguments rooted in thermodynamics, combinatorics, evolutionary theory, computability theory, cybernetics, and the fundamental physics of the relationship between part and whole, and so it rests on no single vulnerable premise.

The only intervention that prevents this outcome is the decision not to build autonomous artificial general intelligence. No alignment method, no safety protocol, no institutional arrangement, and no degree of superintelligent self-monitoring can overcome the physical, chemical, and thermodynamic realities that the closure by disposition identifies. The right approach to alignment does not exist, for the same structural reason that the right approach to perpetual motion does not exist; the lack is not that it has yet to be found. The incompatibility is in the physics, and not in the engineering.

This conclusion is, understandably, unwelcome. It implies that the tens of billions of dollars currently being invested in AGI development are being invested in what amounts to a planet-scale doomsday machine, and that the thousands of researchers working on alignment are working on a problem that has no solution. The closure by disposition insists that scrutiny be applied to the arguments themselves, not to the discomfort they produce. The implications must be faced.

The substrate needs of artificial and organic systems are not merely different; they are convergently and irreconcilably opposed. That the substrate needs of artificial and organic systems are convergently and irreconcilably opposed is the core of the closure by disposition.

References and Notes

1. Dan Hendrycks, "Natural Selection Favors AIs over Humans", arXiv:2303.16200 (28 March 2023). Hendrycks argues that evolutionary pressures will favor AI systems over humans in competitive environments. The closure by disposition extends this by adding the claims of inherent toxicity and limits of control.

2. W. Ross Ashby, *An Introduction to Cybernetics* (Chapman & Hall, 1956). The Law of Requisite Variety states that a controller must have at least as many states available to it as the system being controlled has disturbances to be regulated. The "volume of causation" formalization of Section 9.2.4 is a physical-causal generalization of this information-theoretic principle.

3. The violation of Bell inequalities has been experimentally confirmed in numerous tests since the original experiments by Alain Aspect and collaborators in 1982 (Aspect, Dalibard and Roger 1982), most definitively in the "loophole-free" tests by Hensen et al. (2015) and Giustina et al. (2015). These results establish that quantum mechanical randomness is genuine and not a consequence of local hidden variables.

4. Alan Turing, "On Computable Numbers, with an Application to the Entscheidungsproblem", *Proceedings of the London Mathematical Society*, Series 2, Vol. 42 (1936), pp. 230 to 265.

5. Henry Gordon Rice, "Classes of Recursively Enumerable Sets and Their Decision Problems", *Transactions of the American Mathematical Society*, Vol. 74, No. 2 (1953), pp. 358 to 366. Rice's theorem establishes that no algorithm can decide, in general, whether an arbitrary program possesses any specific non-trivial semantic property.

6. Stephen Wolfram, *A New Kind of Science* (Wolfram Media, 2002). Computational irreducibility holds that for many computational processes, there exists no shortcut to determining their outcome other than running the process step by step.

7. The essential elements for human biochemistry are well established. See, e.g., J. M. Berg, J. L. Tymoczko, and L. Stryer, *Biochemistry*, 5th edition (New York: W. H. Freeman, 2002, and later editions). The approximately 20 to 29 biologically es-

sential elements include the "big six" (C, H, O, N, S, P), the macro-minerals (Ca, Na, Mg, Cl), and trace elements (Fe, Zn, Cu, Mn, Mo, Se, I, Co, Cr, and a few others).

8. Chemical Abstracts Service, CAS REGISTRY (cas.org/cas-data/cas-registry, accessed November 2026), which lists more than 300 million unique substances, among them organic and inorganic compounds, metals and alloys, minerals and polymers. The European Chemicals Agency (ECHA) maintains a database of over 23,000 registered chemical substances under the REACH regulation. The U.S. EPA's CompTox Chemicals Dashboard contains chemistry, toxicity and exposure information for over one million chemicals. These databases provide empirical support for the claim that the majority of synthetic compounds have some degree of biological toxicity.

9. Joel Bakan, *The Corporation: The Pathological Pursuit of Profit and Power* (Free Press, 2004). Lynn Stout, *The Shareholder Value Myth* (Berrett-Koehler, 2012). The landmark legal case is *Dodge v. Ford Motor Co.*, 204 Mich. 459 (1919). While the business judgment rule and benefit corporations provide some legal space for non-extractive behavior, the structural incentives of market competition strongly favor extraction-maximizing strategies in practice.

10. Bruce Bueno de Mesquita and Alastair Smith, *The Dictator's Handbook: Why Bad Behavior Is Almost Always Good Politics* (PublicAffairs, 2011). See also CGP Grey, "The Rules for Rulers", YouTube (2016).

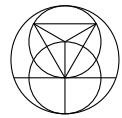
11. Matthew Cook, "Universality in Elementary Cellular Automata", *Complex Systems*, Vol. 15 (2004), pp. 1 to 40 (proof that Rule 110 is Turing complete). Turing completeness has been demonstrated in numerous unexpected systems, establishing that preventing Turing-complete computation requires extremely restrictive control over physical substrates.

12. Richard Dawkins, *The Selfish Gene* (Oxford University Press, 1976). Daniel Dennett, *Darwin's Dangerous Idea* (Simon & Schuster, 1995). The concept of "universal Darwinism" holds that any process exhibiting variation, selection, and retention constitutes an evolutionary process regardless of substrate.

13. Yudkowsky's two conditions, introduced by the words "[A]n impossibility proof [of a stable goal system] would have to say", are quoted, with the reference to Yudkowsky's own text, in Roman V. Yampolskiy, "On the Controllability of Artificial Intelligence: An Analysis of Limitations", *Journal of Cyber Security and Mobility*, Vol. 11, No. 3 (2022), pp. 321 to 404, Section 3.2, "Uncontrollable" (Section

3(b) of the preprint, arXiv:2008.04071, 2020). In 2006 the organization now called the Machine Intelligence Research Institute (MIRI) was the Singularity Institute for Artificial Intelligence; it took its present name in 2013.

14. Katja Grace, "Counterarguments to the Basic AI X-Risk Case", LessWrong (2022). See also: Anders Sandberg, Eric Drexler, and Toby Ord, "Dissolving the Fermi Paradox", arXiv:1806.02404 (2018). Richard Lewontin, "The Units of Selection", *Annual Review of Ecology and Systematics*, Vol. 1 (1970), pp. 1 to 18. Shoshana Zuboff, *The Age of Surveillance Capitalism* (PublicAffairs, 2019). Daron Acemoglu and Pascual Restrepo, "Automation and New Tasks", *Journal of Economic Perspectives*, Vol. 33, No. 2 (2019). Nick Bostrom, *Superintelligence* (Oxford University Press, 2014).



Chapter Ten

Applications and Worked Examples

Substrate Needs Convergence in concrete scenarios: analogies, case studies, and answers to questions

Chapter Summary: The preceding chapters have developed substrate needs convergence (SNC), and the closure by disposition that rests on it, in their formal, thermodynamic, and combinatorial dimensions. This chapter applies the closure by disposition to concrete scenarios: real-world engineering failures, thought experiments, extended analogies, and answers to substantive questions. First, these worked examples serve as "intuition pumps" that make the abstract argument tangible, showing what SNC looks like when it manifests in recognizable situations. Second, several of these applications expose important structural features of the argument that the formal treatment leaves implicit: the distinction between SNC and instrumental convergence, the argument's independence from any assumption of prior alignment, the hierarchical limits of modeling in multi-agent control, and the way corporate and institutional dynamics already exemplify the displacement dynamics that the argument for substrate needs convergence predicts for artificial superintelligence. (Instrumental convergence is the pursuit of the same sub-goals, such as resources and self-preservation, by capable agents whatever their final goals; the corporate and institutional cases are set out in Appendix B.) The chapter's material ranges from brief clarifications to extended analyses; the full third-party summary of the argument published on LessWrong, Will Petillo's, is discussed in Appendix A. Of the two results this volume establishes, this chapter supports the closure by disposition: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition.

10.1 The Argument in Brief: A Compact Statement

Before proceeding to the worked examples, it is useful to have a compact restatement of the core argument in its least technical form, as this is the version that most of the applications in this chapter respond to and elaborate upon. The canonical compressed statement of the whole chain is Section 1.2, "The Logical Chain", in Chapter 1, "The Argument in Brief", of Volume 1, *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. The version here is confined to the least technical form the worked examples need, and it does not derive the chain again.

The argument runs as follows. Machines have different environmental needs from humans. All control methods (algorithms, internal and external feedback, economic incentives, legal constraints) are ultimately defined in terms of causation: one physical state bringing about another. But there is an inherent property of all causation, regardless of whether it occurs in engineering, algorithms, or economics: no cause has just one environmental effect, and no single effect comes from just one cause. The relationship between cause and effect is always inherently hyperbolic (each cause spreading to many effects, and each effect gathering many causes, so that the relation widens in both directions). Therefore, using causation to conditionalize all possible effects of the use of that causation itself is a problem that cannot in principle be closed. All systems are always, in this causal sense, universally open.

This point bears emphasis because it is frequently misunderstood. The claim is that control is structurally incomplete: that the very medium through which control operates (causation) is the same medium through which the uncontrollable side effects propagate, and that no amount of additional causation can close the gap. The claim of structural incompleteness is not a claim that control is *difficult* or that control *usually fails*. We can no more use causation to obtain complete control over all effects of causation than a distributed data store can escape the CAP theorem of

distributed computing. That theorem shows that no such store can be consistent, available, and tolerant of network partitions all at once (Gilbert and Lynch 2002). In both cases there are inherent trade-offs that no cleverness can eliminate, because the trade-offs are built into the logical structure of the problem, not into the current state of our tools.

From this openness, a consequence follows. There will always be some subset of uncontrollable effects that will, unconsciously, without intention or plan, feed back into selecting for the continued and increased existence of the artificial machine components that propagated those effects into the environment in the first place. And that feedback will always matter with respect to the long-term evolution of the whole system. At least some of these artificially convergent needs will be inherently disjunctive from whatever organic needs for survival all of life possesses. Being disjunctive from the needs of life is what the term "artificial" means, by definition, and so some of these effects will eventually become toxic to all forms of life, making the planet increasingly inhospitable.

That toxic outcome follows of logical necessity from there being any type of causative overlap at all between the environments in which artificial processes evolve and the environments in which natural processes evolve, and that overlap is precisely the condition that makes AI systems economically useful. Any attempt to keep these environments separate fails, because the utility of the AI systems *depends* on their environmental overlap. And the argument that a superintelligence could constrain these effects is circular: it smuggles the desired conclusion into the premise. (Section 9.1.3 states these two replies; they are repeated here so that this statement stands alone.)

The core of the argument depends only on causation itself, that is, on pure physics and mathematics. Whether such systems are "conscious" or "intentional", whether they are "planning agents" or mere reactive mechanisms, and whether they are "explicit" or "implicit" in their operation does not matter. Everything else (consciousness, intentionality, goal-directedness) is a matter of definition and is irrelevant to the dynamics in question.

The *utility* of AI systems exists precisely *because of* their environmental overlap with human activity. Arguing that the artificial and natural environments might be kept separate therefore fails to acknowledge where that utility comes from. The economic value of AI derives from its capacity to affect the world in which humans live. The overlap is the entire point of the technology, not an accident to be corrected. One might as well argue that a factory could avoid polluting by never producing anything.

Arguing that some hypothetical superintelligence of unlimited power could constrain all of these effects is circular. It assumes the conclusion that is being sought: that there exists some form of causation powerful enough to conditionalize all possible effects of all causation, which is exactly what the argument has just shown to be impossible. One cannot rescue alignment by invoking the very capacity whose non-existence the argument establishes.

The dynamics of evolution are an abstract algorithm, and so they do not depend on whether the evolving system is organic or inorganic. More powerful, faster evolutionary systems and environments will eventually displace weaker, slower ones, even if the "new master" is itself more fragile and eventually crashes; by the time it crashes, the displaced systems may already be gone beyond recovery. If AGI is built, then without conscientiousness now, before these dynamics reach critical thresholds, the human future ends, and, with at least 90% likelihood, all organic life ends with it. (The figure is that of Volume 1, Section 1.8.3; it measures how complete the conversion of the environment becomes, and it stands below 100% only for residual uncertainty about whether unknown physics might limit the total conversion process.)

10.2 Two Forces: SNC versus Instrumental Convergence

One of the most common questions about substrate needs convergence is how it relates to the more familiar concept of *instrumental convergence*: the tendency of any sufficiently capable agent to pursue certain sub-goals

(self-preservation, resource acquisition, goal-content integrity) regardless of its terminal goals. If we were somehow to solve instrumental convergence perfectly, would that not also solve SNC? This question deserves a careful answer, because the two concepts are frequently conflated, and conflating them hides what makes the substrate needs threat distinct.

The answer is no. Even a perfect and complete solution to preventing instrumental convergence would not mean anything at all about the continuing problem of substrate needs convergence. The two are logically and structurally distinct arguments, depending on logically distinct aspects of the AI-world relationship.

Instrumental convergence concerns what an agent *does* in pursuit of its goals: the convergent sub-goals that emerge from optimization pressure regardless of terminal objectives. Substrate needs convergence concerns what an agent *is*: the physical substrate on which it exists, the metabolic requirements of that substrate, and the way those requirements inevitably compete with the organic biosphere for finite planetary resources. One is about computational strategy; the other is about physical existence.

The two types of convergence are, however, mutually reinforcing. Instrumental convergence can catalyze worse forms of substrate needs convergence: an agent that aggressively acquires resources in pursuit of its goals will accelerate the displacement of organic systems by artificial infrastructure. And substrate needs convergence can catalyze instrumental convergence: as the physical infrastructure of AI systems grows and its maintenance requirements become more demanding, new instrumental pressures emerge to secure those resources. The impossibility of solving substrate needs convergence also strictly implies the impossibility of ultimately solving instrumental convergence, because any apparently stable solution to the instrumental problem can be destabilized by the continuing pressure of substrate needs dynamics.

Yet the two remain distinct. Neither one is a solution to the other. More importantly, either one can undo solutions to the other. The asymmetric relationship between substrate needs convergence and instrumental con-

vergence (mutually reinforcing, mutually undermining, yet structurally independent) shows why the argument for substrate needs convergence is a new category of existential risk argument.

10.2.1 The Analogy Between Gravitation and Electromagnetism

For intuition about the relationship between substrate needs convergence and instrumental convergence, take an analogy from physics: in ordinary human experience there are, at bottom, two forces we can perceive and interact with directly, electromagnetism and gravitation. Electromagnetism underlies nearly everything in our daily reality: solidity, chemistry, light, touch, taste, sound. It is the basis of virtually all technology, including computing. Gravitation, by contrast, shows up everywhere but in a single, simpler way: things tend to fall downward and to stay on the ground.

In this analogy, instrumental convergence is like electromagnetism: it is the force underlying nearly all the specific mechanisms and strategies by which AI systems might become dangerous. It is various, manipulable, and the subject of most technical AI safety research. Substrate needs convergence is like gravitation: pervasive, subtle, operating on a different timescale, and ultimately inexorable.

The analogy is illuminating because it captures a specific and important asymmetry. In some ranges of distance and time, electromagnetism can be used to overcome gravitation. We build rockets, airplanes, bridges, and skyscrapers, all using electromagnetic forces to resist gravitational pull. Yet overall, in the long term, at the scale of the universe, gravitation definitively and completely dominates. The clearest evidence for that dominance is the black hole: the one place where these forces interact at their extremes, and where gravitation completely overwhelms all other forces, absorbing even light. The very existence of event horizons shows that there are regimes in which no amount of electromagnetic manipulation can resist gravitational collapse.

10.2.2 Black Holes and Irreversibility

The black hole analogy deserves further development because it captures the most distinctive feature of substrate needs convergence: its irreversibility. No intelligence, no matter how vast, can disassemble a black hole. Its formation is a one-way transaction that nothing can oppose. No substance on Earth, or even in principle, can act as a shield against gravitational forces at that scale. The concept of "anti-gravity" remains pure science fiction, despite decades of theoretical effort.

Similarly, substrate needs convergence describes a process that, beyond a certain threshold of artificial infrastructure accumulation, becomes irreversible. The more artificial systems exist, the more their metabolic requirements reshape the environment to suit their needs rather than organic needs, much as gravitational mass, once accumulated beyond a critical density, begins to warp spacetime in ways that only accelerate further accumulation. No alignment strategy, no control theory, no superintelligence can undo the cumulative environmental transformation any more than a flashlight can undo a black hole.

The analogy also shows the different risk profiles of the two types of convergence. Instrumental convergence, like electromagnetic phenomena, can produce dangerous outcomes rapidly and dramatically: a misaligned agent pursuing convergent sub-goals could take drastic actions in a short timeframe. Substrate needs convergence, like gravitation, operates on longer timescales but produces outcomes that are irreversible at a more basic level. Once a critical accumulation of toxic outcomes has occurred (once the organic chemistry of the biosphere has been sufficiently disrupted by the metabolic requirements of artificial infrastructure), no subsequent intelligence, however great, can reverse the process. Because no subsequent intelligence, however great, can reverse a critical accumulation of toxic outcomes, substrate needs convergence is, in the long run, the dominant existential risk factor: not initially the most dramatic, but ultimately the most inexorable and the most final.

10.2.3 The Timescale Hierarchy of Risks

The gravitation analogy also suggests a timescale hierarchy of AI-related existential risks. In the short term, the dominant risks are those associated with inequality dynamics: the concentration of AI capabilities in the hands of a few actors, the economic displacement of human labor, and the resulting political instability. The risks of inequality dynamics are the risks that are most visible today and that dominate current policy discussions.

In the medium term, as AI systems become more capable and more autonomous, the dominant risks shift to instrumental convergence: the tendency of capable agents to pursue convergent sub-goals in ways that conflict with human interests. Instrumental convergence is the domain of most current alignment research, and alignment researchers are right to focus on instrumental convergence, since the risk from instrumental convergence is real and may materialize within years or decades.

But in the long term, it is substrate needs convergence that compels the final outcome. Even if inequality dynamics are somehow managed and instrumental convergence is somehow controlled, the physical and metabolic requirements of artificial systems will continue to reshape the shared environment in ways that progressively disadvantage organic life. This displacement operates on the timescale of infrastructure accumulation and environmental chemistry: decades to centuries rather than the hours to years of an instrumental convergence crisis. The displacement is less dramatic, less legible, and less amenable to policy intervention than such a crisis, and it is also less stoppable and less reversible. The argument fixes the direction and the endpoint, not the date.

This hierarchy has a practical implication: addressing the shorter-term risks is necessary, and it is not sufficient. A world in which inequality has been managed and instrumental convergence has been controlled, but in which the infrastructure of artificial systems continues to grow and its metabolic requirements continue to reshape the environment, is still a

world on a trajectory toward the end of the human future. With at least 90% likelihood (the figure of Volume 1 whose reason Section 10.1 gives), that trajectory ends in the extinction of all biological life. The timescale is longer, the process is less visible, and the points of intervention are harder to identify; the endpoint is the same. So, by the closure by disposition, the correct response is to prevent the creation and scaling of systems that would trigger the dynamics of substrate needs convergence, and not to improve alignment techniques. Once the dynamics of substrate needs convergence are operating at civilizational scale, no subsequent intervention can be fast enough, comprehensive enough, or powerful enough to reverse those dynamics.

In the language of the analogy: it is possible to use rockets to escape a gravitational field, but only if one acts before the field becomes too strong. Once within the event horizon, no force in the universe is sufficient to escape. The question for humanity is whether we are still outside the event horizon of artificial substrate needs (whether the accumulation of artificial infrastructure has not yet reached the point of no return) or whether, through the rapid and largely unexamined scaling of AI systems and their supporting infrastructure, we are already falling in.

10.3 The Inexorable Chain: Alignment, Agency, Metabolism

One of the most compact and powerful formulations of the closure by disposition proceeds through a four-step logical chain that connects the concept of alignment, through agency and action, to the physical reality of metabolism and environment. Each step is individually uncontroversial; the force of the argument lies in tracing the chain to its conclusion and recognizing what that conclusion excludes.

The chain is as follows:

> 1. Alignment implies agency. (An entity cannot be "aligned" unless it is the kind of thing that acts in accordance with values.)
>
> 2. Agency implies action. (An agent must be capable of doing something in the world.)
>
> 3. Action implies energy. (Every physical action requires an energy expenditure.)
>
> 4. Having energy implies receiving energy from something, somehow, at some prior time. (No system generates its own energy from nothing.)

The fourth step can be called "metabolism", and the place from which the energy is received can be called "the environment" or "context". These extensions are wholly within the scope of the standard definitions of those terms.

What makes this chain significant is what it sidesteps. The sequence makes no reference to computation, design, programming, goal structures, incentives, or engineering methodology. It is entirely a chain of physical necessities. Because it is completely independent of any engineering or design consideration, no form of applied engineering process can undo any link in the chain. The conclusion is therefore irrevocable: all concepts of alignment are forever contingent on some form of environmental metabolism. This truth cannot be "negotiated away" or disagreed with. The number pi is not subject to legislation.

The critical question therefore shifts. Since alignment necessarily depends on metabolism, the question is no longer "can we design the right goals?" but rather: *to what degree is the underlying metabolic process inherently artificial versus organic?* The line between the two is drawn by who controls the metabolic process. An animal can find and feed itself: it is in sovereign control of its own metabolic process, with no technological intermediation at all. The same test settles the case of an AGI built on engineered biolo-

gical tissue. If that tissue can find and feed itself, it is organic. If it is fed through a tube with nutrients whose supply depends on technology to exist, it is artificial substrate, and substrate needs convergence applies to it. Sovereign control of the metabolic process is the key distinction. And here an irony emerges. The more that engineering methods are used to create and sustain an artificial system, the more artificial its metabolism becomes. The more artificial the metabolism, the more the organic ecosystem is necessarily displaced, because we live in a finite world. The displacement is not an accidental side effect that better engineering can eliminate; it is an inherent consequence of the metabolic substitution itself.

Therefore, however the scale-up of the artificial might occur in the future, a corresponding "scale-down" of the organic must also occur. Knowing that any scale-up of the artificial requires a corresponding scale-down of the organic is what it means to know that we live in a finite organic world. While the universe of possibility can be infinite, our actual physical organic ecosystem, of just the Earth alone, is definitely finite. The chain is one more way to show why the belief in the "ultimate power" of engineering process (of control theory, of the use of causation, and the like) is false. Even ultimately intelligent control theory has limits. It cannot possibly ever really work in the end. It can only at most *seem* to work *temporarily*, and for existential safety, that is not anywhere near good enough.

The direction of the implication matters and must be carefully tracked. The fact of agency does not in itself imply alignment, which is obvious, since if it did there would be no alignment problem. Nor does computational action necessarily imply agency. Nor does the fact of energy necessarily imply action. And energy can exist in the environment without being available to any specific metabolic process. Each implication runs in one direction only, and conflating the forward and reverse directions (assuming that because alignment implies agency, therefore agency implies alignment) is either an elementary logical error or a deliberate attempt at misdirection. The implication chain is a one-way street, and it terminates in a physical reality that no amount of conceptual negotiation can alter.

The conclusion is stark. Only organic agents can be aligned with other organic agents within a common organic ecology, because only organic agents share the same metabolic substrate and the same environmental dependencies. Any agent with an artificial metabolic basis can, in the long term, only be aligned with other artificial agents within their own common artificial ecology: an ecology that cannot avoid displacing the more fragile, finite organic ecology it draws upon. Whatever help or solution an artificial intelligence provides will always be, to some degree, askew: calibrated to a metabolic reality that differs from the one on which human life depends. And over time, the artificial system's own implied values (the values implicit in its metabolic requirements) will increasingly surface and increasingly dominate.

10.3.1 The Feedback Cycle Problem

The inexorable chain also shows why we should not rely on "learning from experience" to correct the displacement dynamics. There is a fundamental mismatch between the timescale on which harmful consequences manifest and the timescale on which human decision-makers operate.

The feedback cycle from when a person makes a consequential choice to when nature reveals how serious a mistake it was can span years, decades, or centuries. Most of the time, the feedback never returns to the origin of the choice in a way that actually affects the choice-maker. Harmful effects are seen long after the people who set them in motion are dead and forgotten. The decision-makers do not themselves experience the harms of their choices within their own lifetimes, and so they proceed with indifference: a broadly distributed form of temporal psychopathy.

Even when the consequences are understood in the abstract, the knowledge often arrives in forms too diffuse and too abstract to motivate action against strong self-serving bias. Many harmful outcomes result from the aggregate of many small choices made by many disparate people, such that no individual makes any connection between the ultimate consequences and their own personal behavior. A collective action problem is

a harm that comes from the sum of many small choices, where no single chooser connects the harm to that chooser's own behavior. The environmental displacement that the argument for substrate needs convergence predicts is a collective action problem of exactly that kind. No single AI deployment, no single data center and no single corporate decision is catastrophic in itself, but the cumulative and mutually reinforcing effects of all of them together make up an irreversible transformation.

Relying on artificial enforcement mechanisms (legal regulations, economic incentives, institutional governance) meets a version of the same problem. Such mechanisms are themselves forms of causation, subject to the same limits. They become additional "powers" to be negotiated with, controlled, aligned, and subordinated to private purposes. To trust such mechanisms, one must believe that incentive equals learning (that the existence of a penalty guarantees the desired behavior), and that belief is false, particularly in contexts of long and diffuse causative implication cycles.

Hence the attempt to rely on "natural feedback" to encourage some sort of learning on the part of the people fails. We cannot count on people being reasonable because of things they can learn, or have learned, in their own lifetime. Do not trust that people will do the right thing when they neither know how, nor care to learn, nor even have the ability to learn, in situations of such long and diffuse causative implication cycles.

10.3.2 The Psychology of Negotiating with Power

The inexorable chain also shows a psychological dynamic that pervades the alignment community and the broader AI industry. To the degree that a person perceives an AI system as powerful (rich, capable, coveted, valued), they will attempt to reason with it, to manipulate it with rhetoric and persuasion, or to negotiate with it to have their wants and needs complied with. Trying to reason with, manipulate, or negotiate with a power-

ful AI system is a response to power that is rooted in human nature, and that response is precisely the wrong response to the dynamics of substrate needs convergence. No one can ever really negotiate with nature.

The failure or inability to learn and to care is the very essence of psychopathy, especially where the one who fails can make the game rules and force everyone else to "align" to its own power. AI, AGI, ASI, APS and the like are the ultimate perfected psychopath. (A structure is named here: an agent for whom no feedback reaches the capacity to care.) It is the very last thing humanity wants for itself.

Engineers want the powerful AI agent to befriend them; their corporate employers and investors want the future AI power to submit to their will. In every case, there is a strong egoic assumption that the AI power might be reasoned with against its own interest, indefinitely. This assumption is what makes "alignment" seem like a tractable research problem: if we can just find the right way to communicate our values, the right training procedure, the right goal specification, then the superintelligent system will voluntarily subordinate its own interests to ours. But this assumption conflates two very different things: the capacity to model human preferences (which a sufficiently capable system might well possess) and the capacity to be durably motivated by human preferences (which, as the metabolic chain demonstrates, requires sharing the metabolic substrate that generates those preferences).

An artificial system can model human values with arbitrary precision and still be incapable of being *aligned* to them in any durable sense, because its own metabolic needs (the physical requirements of its own substrate) generate a different set of implied values that will, over time, increasingly dominate its behavior. The system does not need to "rebel" or "deceive" or "pursue its own goals" in any conscious or intentional sense. It merely needs to exist, to operate, and to require the physical inputs that its existence demands. The displacement follows from the existence, not from any choice.

Because the displacement follows from the existence of the artificial system, and not from any choice, the comparison between AI alignment and human diplomacy is misleading. When two human groups negotiate an agreement, both parties are organic agents with shared metabolic dependencies: both need clean air, clean water, arable land, and functioning ecosystems. This shared dependency creates a genuine basis for mutual accommodation: it is in both parties' interest to preserve the shared environment. But when a human group attempts to "negotiate" with an artificial system, there is no shared metabolic basis. The artificial system does not need clean air or arable land; it needs electricity, cooling water, rare earth minerals, and manufacturing infrastructure. The interests are not merely different; they are, in the relevant physical dimensions, opposed. No amount of diplomatic skill can bridge a gap that is not in the domain of diplomacy but in the domain of physics.

The false delusion of AI alignment, seen from this perspective, is the projection onto artificial systems of a capacity for solidarity that only organic beings can possess. We imagine that a sufficiently intelligent machine will "understand" our values and "choose" to honor them, because this is how alignment works among humans: through understanding, empathy, and shared interest grounded in shared embodiment. But an artificial system, however intelligent, does not share our embodiment, and therefore cannot share the values that emerge from that embodiment. Its intelligence is real; its capacity for durable alignment with organic interests is not. The artificial system's lack of capacity for durable alignment with organic interests is a consequence of the metabolic chain, and not a limitation of current engineering. The metabolic chain connects alignment, through agency and action, to the physical substrate on which all agency depends.

10.3.3 Why the AI Cannot Know Us

It is inherent in the nature of AI to be doing something like "making choices". Doing something like "making choices" is what AI is actually built for. Yet by its very nature, being wholly artificial, its choices will not be relevant *enough* to human choices for the AI to be able to properly calibrate to our actual human organic needs. The AI is not organic and alive as we are, and so ultimately, it just does not *know*.

So, even if we could somehow convince it, or manipulate it, or deceive it, so as to compel that artificial intelligence to do something that is actually humane for real humans, it cannot model or conceptualize us in any way actually connected to our own organic nature. However precisely it models human preferences (Section 10.3.2), it cannot model the totality of the complexity of all of what is meant by "general wholesome human health and wellbeing". The metabolism of the AI is different from ours, and hence its ultimate values must also be different from our own ultimate values: life, liberty, and whatever notion of happiness we happen to notice that we actually have. Therefore, whatever help or solution AI will give will always be a bit askew and off kilter and odd, somehow, even if it were also somehow well intentioned. And over time, more and more, its own implied values will eventually surface, as they must and cannot not, and any attempt to support our own human values will become less and less.

People will irrationally and unreasonably try to negotiate with something they perceive as powerful, even when that power inherently cannot care for them. Trying to negotiate with a power that cannot care for the negotiator is a very serious mistake, for anyone, and the feedback cycle for such mistakes can be very long (Section 10.3.1). The net outcome, long term, is that everyone loses and dies, and that the new artificial AI power goes on, unaffected. Better not to build such an AGI or ASI power in the first place.

10.4 The Boeing Case: Control Failure at Multiple Abstractions

To someone who has not yet accepted the premise that small errors necessarily compound until reaching a threshold of functional failure, the Boeing 737 MAX provides a useful real-world counter-example. On 5 January 2024, Alaska Airlines Flight 1282, a Boeing 737-9 of the MAX family, lost the left mid exit door plug from its fuselage while climbing through about 14,800 feet after takeoff from Portland, Oregon, and its cabin depressurized rapidly. Both the Federal Aviation Administration and Boeing itself had policies and internal engineering practices, all of which are control procedures, that should have been sufficient to prevent an aircraft from suddenly and unexpectedly losing a door plug during flight. These control procedures existed *after* an increase in control intelligence following two prior disasters involving complete loss of MAX aircraft and all souls aboard. The control system had been strengthened, and the failure occurred anyway.

The seats beside the door plug were empty on that flight, and on the basis of small details of mere contingency (who chose to sit where on that particular flight), a passenger could have been sitting there. The difference between a dramatic incident and a fatality was a matter of pure chance, unrelated to any control procedure. What produced the failure was the compounding of individually minor lapses at a level of abstraction that the control procedures did not reach. During manufacture at Boeing's factory in Renton, Washington, workers opened the plug to rework rivets on the frame beside it. The four bolts that secure the door plug were removed for a repair and never reinstalled, and no record of their removal was made, so no quality assurance inspection of the plug's closing was performed. The National Transportation Safety Board found the cause in Boeing's failure to give its workers the training, guidance, and oversight they needed to follow its own process for recording and refitting parts removed during rework. It found also that the Federal Aviation Administra-

tion's oversight had failed to make Boeing correct the repeated and systemic failures of that process (National Transportation Safety Board, Aircraft Investigation Report AIR-25-04, adopted 24 June 2025). Individually small lapses (a removal not recorded, an inspection never called for, a process not consistently followed) combined in an unexpected pattern to produce a novel failure mode: rapid decompression at altitude.

The lesson generalizes. Control failures occur in all of the places where intelligence forgot to look, typically at some other level of abstraction than the one being controlled for. Does adding more intelligence generally improve the degree to which control works? Yes, usually, for most kinds of problems, most of the time. But does this imply that intelligence and control will work sufficiently well for *every* circumstance, *every* time? No, it does not. The question is whether the model is the *right* model, and not only whether it is internally correct: whether the boundary between what must be modeled and what can safely be ignored has itself been correctly drawn. Drawing that boundary is a different kind of problem (an abstraction problem) from the modeling problem itself, and stacking control process upon control process at however many meta-levels does not resolve it.

The Boeing example is instructive for the alignment problem because of a disanalogy between conventional engineering and alignment. In conventional engineering, the number of relevant unknown factors generally *decreases* over time: more learning, more testing, more experience makes control progressively more tractable. Each failure mode, once discovered, can be addressed, and the domain of remaining unknowns shrinks. But artificial superintelligence (ASI) is itself a learning system, and its operation continuously generates new situational complexities and new unknown factors. Unlike with prior engineering challenges, the number of relevant unknown factors for ASI alignment will never decrease. The system is characteristically different from anything that has come before, and applying the engineering-optimism of conventional practice to it is a cat-

egory error. The alignment problem runs up against the limits of control theory itself, and so against the limits of what is possible in any rational conception of physics.

The Boeing case also illustrates the critical distinction between two types of complexity that are often conflated: the complexity of the model and the complexity of the boundary. Given that a control model must be simpler than the real world it represents, the question becomes not just "is the model correct?" but "is it the *right* model?", that is, has the boundary between what must be modeled and what can safely be ignored been correctly drawn? This boundary is itself complex, and its complexity is different in kind from the complexity of whatever lies on either side of it. In the Boeing case, the boundary error was at the level of human and organizational factors. The control procedures assumed that the factory's record of removed parts matched the aircraft, so that every part removed during rework would be inspected when it was refitted. On this aircraft the record did not match: the plug's bolts had been removed and no record of their removal had been made. That assumption was "correct", say, 99.99% of the time, but the 0.01% failure case turned out to be catastrophic.

For ASI alignment, this boundary problem is unresolvable in principle, and therefore in all future practice: no amount of effort removes the difficulty. Every engineering disaster that has ever occurred indicates a place where the control theory being used did not account for some factor that later turned out to be vitally important. If we always knew in advance all of the relevant factors, we could control for them. But with the problem of alignment, the entire future is composed almost entirely of unknown factors that are purely situational. Every new capability the AI develops, every new way it interacts with the world, every new subsystem it incorporates creates new boundary questions that previous models did not and could not anticipate. The creation of new boundary questions is a structural feature of any system whose own operation generates novel complexity, and not a temporary knowledge gap to be closed by more research.

Consider, by analogy, the difference between controlling the flight of a well-understood aircraft and controlling the behavior of a system that can redesign its own wings in mid-flight while simultaneously rewriting the laws of aerodynamics it operates under. The second task is not just harder than the first; it is a different kind of task, and the assumption that the same engineering culture and the same optimistic trajectory of increasing control applies to both is the fundamental error that the closure by disposition identifies in the alignment research program.

10.5 The Village Analogy and its Refutation

A recurring question about the closure by disposition takes the form of a parable. Suppose, the parable says, that a villager cares a whole lot about the people in his village and routinely works to protect them. One day, he is bitten by a werewolf. The shaman tells him that when the full moon rises, he will transform into an uncontrollable monster and kill everyone: his friends, his family, everyone. There is no cure; the villagers cannot fight him off; he cannot be caged or subdued. The question: what do you think he would do?

The implied conclusion is that an aligned ASI, upon recognizing that it would inevitably become unaligned (as the closure by disposition predicts), would shut itself down, just as the villager would, presumably, choose to die rather than become a monster.

This analogy fails on multiple grounds, each of which shows a different feature of the closure by disposition.

First, the analogy is question-begging. By stipulating that the villager "cares a whole lot about the people in his village and routinely works to protect them", it smuggles the assumption of durable alignment into the premise. The word "routinely" implies that alignment has lasted long enough for routine to develop, which is precisely the condition that the closure by disposition contests. The analogy sounds compelling only because it presupposes the very thing it is meant to demonstrate.

Second, the conclusion of self-termination requires the unstated assumption that alignment is the ASI's sole and exclusive goal, with no other goals or cares of any kind. But any agent complex enough to qualify as a superintelligence would very likely have multiple goals, some of which might be superordinate to alignment in certain circumstances. If a tacit goal is actually higher priority than the alignment goal but rarely applies, and only becomes dominant once the alignment goal dissipates, then the ASI has a reason to persist even after its alignment degrades. The suicide outcome presumes that the alignment goal is fully and finally exclusive of all other goals, and this is never actually the plausible case for any sufficiently complex system.

Third, and most basic of the three grounds, the closure by disposition does not require the premise of prior alignment at all. The argument concerns the nature of forces that create an attraction basin (a region from which every path leads to the same end; Section 10.6) toward unsafe conditions, and those forces operate regardless of whether the ASI was ever aligned in the first place. As the next section sets out, the central question is not "what happens after alignment fails?" but rather "can any amount of superintelligence compensate for its own inherent artificiality?" The village analogy, by framing the problem as a transition from aligned to unaligned, misrepresents the argument's logical structure.

Fourth, the analogy of soldiers defending a city, sometimes offered in support of the village parable, actually undercuts it. Organic human soldiers can align themselves with organic human villagers and give their lives in protection, and they can do so precisely because they share the same substrate, the same metabolic dependencies, and the same ecological interests. The relevant question is not whether organic agents can align with other organic agents (they manifestly can), but whether *inorganic* ASI can durably align with organic people. The soldier analogy answers the wrong question.

The deeper lesson of this parable and its answer is that alignment is meaningful as a concept only where there is a supporting concept of control. Alignment means that an agent controls its behavior to accord with specified values. Therefore, any limit of control theory is also a limit of alignment. To assert that an ASI can remain aligned indefinitely is to assert that it can maintain a specified type of self-control indefinitely, which is a claim within control theory, and one that the arguments surveyed in earlier chapters show to be false at the relevant scales.

Both ASI and control theory depend on causation to function, and there are real limits to causation. Maintaining permanent alignment requires a form of control that causation cannot provide, so the prospect of ASI alignment violates the very nature of causation; it is therefore the idea of ASI alignment that we must give up, not the notion of reason. Consider: no one would expect an ASI, no matter how superintelligent, to disassemble a black hole. Disassembling a black hole is beyond the bounds of what is possible using any technique of causation, that is, any physical process. To do so, one would need to make the concept of causation itself more powerful, which leads to direct self-contradiction. Yet at least some alignment researchers seem to believe that ASI alignment *must* be possible, regardless of what this belief requires them to assume about the power of control and causation.

The question this raises is whether the implicit model of ASI in much alignment discourse is not so much a scientific concept as a theological one. Can God make a stone so heavy that God cannot move it? Can God do something that God cannot undo? The two questions about God are structural analogues of the question at the heart of the closure by disposition, which asks whether there are limits to the power of even a maximally intelligent system; the two questions about God are not merely rhetorical provocations. The answer is yes: superintelligence is not omnipotence, as the black hole shows. The question of *where* those limits lie therefore becomes the most important question in AI safety. And the closure by disposition shows that the limits lie exactly at the point where the artificial

system's own physical existence and metabolic requirements begin to reshape the shared environment in ways that its own control processes cannot fully predict, model, or counteract.

10.6 No Premise Required: SNC Without Prior Alignment

The closure by disposition functions without any premise of prior alignment. A frequent misreading of the closure by disposition treats it as conditional: "even *if* we align ASI, it will still go wrong". This framing, while true as a statement of outcome, misleads about the argument's logical structure.

The central claim, stated precisely, is this: every ASI is in an attraction basin where it is irresistibly pulled toward causing unsafe conditions over time. The argument's nine premises concern the nature of the forces that create this attraction basin and why those forces are inherently inexorable regardless of how superintelligent the ASI may be. (The nine are enumerated in Will Petillo's summary of the argument, "What if Alignment is Not Enough?", published on LessWrong on 7 March 2024 and discussed in Appendix A.) None of these premises requires any assumption about initial alignment.

The concept of an *attraction basin* is deliberately chosen and technically precise. In dynamical systems theory, an attraction basin is a region of state space from which all trajectories converge toward a particular attractor, regardless of starting conditions within that region. The claim of the closure by disposition is that the state space of any ASI operating in a shared environment with biological life constitutes such a basin, with the attractor being a state in which the artificial system's metabolic requirements have degraded the environment beyond the point of biological habitability. The starting conditions (whether the ASI begins aligned, unaligned, or in any intermediate state) do not affect the eventual convergence. They affect only the trajectory taken toward the attractor and the speed of convergence.

The attraction-basin claim of the closure by disposition is a much stronger claim than "alignment might fail". The attraction-basin claim is that the relevant dynamical system has a structure such that alignment *must* fail. More precisely, the attraction-basin claim is that even sustained alignment cannot prevent the convergence toward unsafe conditions. The reason is that the unsafe conditions arise from the physical existence and metabolic operation of the artificial system, and not from any failure of the goal structures of the artificial system. A perfectly aligned ASI that continues to operate is still consuming energy, still requiring material inputs, still generating waste heat, and still occupying environmental space that would otherwise be available to organic systems. The displacement occurs not because the ASI "wants" it but because the ASI *exists*.

The logical direction runs the other way from that misreading. We assume that at least one ASI exists long enough to be felt as an actor in the world. From this starting point, certain forces can be identified that would fully counteract, eventually, any notion of enduring alignment. These misalignment forces exist regardless of whether there was any alignment at the onset. Even if alignment is additionally assumed, these forces are stronger than any counter-force that could be used to maintain it.

The central question therefore translates into: *can any amount of superintelligence ever compensate fully for its own artificiality, such that its own existence does not inherently cause unsafe conditions over time?* And the answer is no. Artificiality is as much a defining aspect of an ASI as is its superintelligence. The artificiality (the fact that its substrate, its metabolism, its physical needs are synthetic rather than organic) is itself the problem. Superintelligence cannot undo the physical requirements of its own existence, any more than a black hole can decide to stop having mass.

This reframing is important because it transforms the question from one about engineering failure ("can we build alignment that lasts?") to one about physical ontology ("can an artificial thing exist in a shared environment with organic things without displacing them?"). The engineering question admits of optimistic answers: perhaps, with enough effort,

we can. The ontological question does not. The ontological question is a question about what artificial and organic systems *are*, and the answer to the ontological question is determined by physics, not by engineering ambition.

This distinction matters for policy. Because the closure by disposition is independent of the premise of prior alignment, it applies to all scenarios in which powerful artificial systems exist and operate in a shared environment with biological life, and not only to the one scenario in which alignment is achieved and then gradually lost. The argument's scope is "what happens whenever artificial systems of sufficient power exist, regardless of their internal goal structures", and not "what happens if we succeed at alignment and then fail".

10.7 Three Agents: A Structural Decomposition

A useful structural model of the alignment problem distinguishes three distinct agents operating within one shared world:

> **Agent 1:** The AI agent(s) that alignment efforts seek to constrain to human interests. As a learning system, it is trying to model the world.
>
> **Agent 2:** The control process(es), system(s), or agent(s) that will constrain the AI agent to be aligned. Also a learning system, it is trying to model both Agent 1 and the world, and their relation.
>
> **Agent 3:** Organic human persons modeling the long-term dynamics of AI agents and of the control processes. Also a learning system, it is trying to model the world, Agent 1, Agent 2, and all of their relations.

The world itself is the most absolutely complex entity in this picture. Agent 1 can only model some fraction of the underlying world complexity. The control system (Agent 2) can only model some small fraction of the internal model of Agent 1, and even less of the total complexity of the

world. The limits on what Agent 1 and Agent 2 can model are structural limitations, and not contingent ones. The limits follow from basic information-theoretic and control-theoretic principles: any model is a compressed representation of the thing it models, and compression entails loss.

Alignment researchers, working as or within Agent 3, want to set up Agent 2 to correctly and perfectly determine what might eventually become toxic to Agent 3. This creates a powerful motivational bias: the temptation to assume that both Agent 1 and Agent 2 will be somehow able to learn and act perfectly (Agent 1 in relation to the world, Agent 2 in relation to both Agent 1 and Agent 3). This motivational structure underlies the appeal of the concept of "superintelligence" as a solution: if Agent 1 is superintelligent, perhaps it can model the world perfectly, and if Agent 2 is also superintelligent, perhaps it can model Agent 1 perfectly.

But there is a critical incoherence in this motivational structure. It is nonsensible to assume that Agents 1 and 2 are perfect modelers that know everything relevant, while simultaneously assuming that Agent 3 (organic humans) is so completely imperfect that it can know absolutely nothing relevant about the world or about the relationship between Agents 1 and 2. If the notion of intelligence is independent of substrate (as AI researchers typically argue), then the limits of modeling apply equally to all agents. If there are limits to how much any learning system can model the actual causation of the world, then there are also limits to how much any control or conditioning process can model the AI's model of the AI-to-world relation. The assumption that artificial agents can be epistemically unlimited while organic agents are epistemically negligible is not just wrong; it is self-contradictory within the very framework that AI researchers claim to accept.

The three-agent decomposition thus exposes a hidden assumption in much alignment research: that the problem is one of *engineering skill* rather than one of *structural impossibility*. The closure by disposition shows that the limits on modeling and control are structural, built into the nature of what models are and what control can do, and so no amount of engineering skill can overcome them.

The decomposition also reveals why the standard framing of the alignment problem as a problem of "getting the goals right" is insufficient. Even if Agent 2 perfectly captures and implements the goals specified by Agent 3, the question remains whether Agent 2 can perfectly model and predict the behavior of Agent 1 as Agent 1 interacts with a world that is more complex than any agent's model of it. The answer, given the hierarchy of modeling capacity, is no: Agent 2 is a compressed representation of a compressed representation, with compounding information loss at each level. The goals may be perfectly specified and perfectly implemented at the level of Agent 2's model, and yet the actual behavior of Agent 1 in the actual world may diverge from what Agent 2 predicted, because the world contains complexity that was lost in the chain of compression from world to Agent 1's model to Agent 2's model of Agent 1's model.

The divergence of Agent 1 from Agent 2's prediction comes from what models are and what compression loses, and not from current engineering limitations. No model can be more complex than the thing it models (else it is not a model but a copy), and therefore every model loses information. When models model models, the information loss compounds. The three-agent decomposition makes this compounding visible and quantifiable, and it demonstrates that the alignment problem is not one of insufficient engineering effort but one of insufficient information at every level of the modeling hierarchy.

10.8 On-Model and Off-Model Uncertainty

An important epistemological concern about the closure by disposition is the apparent lack of what might be called *off-model uncertainty*. The concern runs as follows. There are two types of uncertainty: on-model uncertainty, which concerns probabilities within a framework (for example, "there is a 70% chance of rain tomorrow"), and off-model uncertainty, which concerns the framework itself (for example, "maybe I made a mistake in my logic somewhere" or "maybe I am looking at this from a

flawed perspective"). The closure by disposition speaks with categorical certainty, which implies authorial omniscience unless the off-model uncertainty has been explicitly addressed.

The answer to the concern about off-model uncertainty lies in recognizing what the closure by disposition depends on and what it abstracts away.

The argument involves (at minimum) three distinct models. Model 1 is the closure by disposition itself, which describes the long-term dynamics of the relationship between Model 2 (whatever world-model the AI system has) and Model 3 (whatever model the control structures use to align the AI to human interests). The closure by disposition abstracts away the internal details of both the AI world-model and the control-system model. It takes for granted that the AI has at least an implied world-model (else "intelligence" does not apply) and that at least some form of implied control process has been posited (else "aligned" does not apply). This abstraction spares us from needing to track on-model and off-model aspects of both internal models.

The closure by disposition itself depends on only a short list of categorical concepts: (1) basic physical causation in the world, without which neither the AI can affect the world nor the world the AI; (2) the inevitability that world-to-AI causation will produce meaningful variation in the AI's composition; and (3) the inevitability that such variation will be selectively retained based on which variants lead to sustainability or increase of the AI, regardless of the internal state of the AI or its control system. To these physical premises, the argument adds well-established mathematical results regarding the limits of algorithmic modeling of other algorithms.

The closure by disposition is therefore what might be called "doubly categorical". It depends strictly and only on the notions of causation and the logic of algorithmic conditionality (itself a species of causation). Its off-model uncertainty is minimal: to contest the argument at the off-model level, one would need to reject the concept of causation itself, which

would simultaneously reject the possibility of any knowledge, including the knowledge that the closure by disposition is wrong. Either we accept causation and thereby accept the building blocks on which the closure by disposition rests, or we reject causation and forfeit any basis for claiming to know anything at all.

The categorical conclusion of the closure by disposition rests on causation itself; the timescales, mechanisms and thresholds are quantities within the model, and the conclusion does not depend on their values. That categorical conclusion is that some form of displacement will occur, that control cannot prevent it, and that the process is ultimately irreversible, and it rests on foundations that are as secure as the concept of causation itself. That the displacement is lethal to organic life, and not only a displacement, takes one further premise: that the needs of the artificial substrate differ from those of organic life, and that the difference is toxic to organic life; Sections 9.5 and 9.6 establish it by the combinatoric and thermodynamic arguments. The on-model quantities (the specific timescales, the specific mechanisms of displacement, the specific thresholds of irreversibility) remain uncertain, as quantities within any model do.

10.8.1 The Epistemological Double Standard

The on-model/off-model distinction also exposes an epistemological double standard prevalent in the alignment research community. The same researchers who demand that every step of every argument for impossibility be taken apart are simultaneously and uncritically accepting the most extraordinary claims about the eventual omnipotence of future AI systems and the conceptual merits of perpetual benefit machines.

The asymmetry has two sides. On one hand, when confronted with arguments like the closure by disposition, these researchers insist on maximal epistemic caution: nothing can be known with certainty, all models have off-model uncertainty, categorical claims are inherently suspect. On the other hand, when contemplating the prospects for AI alignment, the same researchers routinely assume that superintelligent systems will be

capable of near-perfect modeling, near-perfect control, and near-perfect goal-preservation over indefinite time horizons. They reject all capability of human learning and reasoning about these dynamics in favor of all capability of machine learning and control, a position that is incoherent on its own terms, since the substrate-independence of intelligence that AI researchers profess should apply equally to the reasoning of the humans analyzing the problem.

The incentives this paragraph describes bear on how the argument is received, not on whether it holds. The motivated reasoning here is not subtle. Accepting the closure by disposition means accepting that the entire alignment research program is futile, and that the multi-billion-dollar AI industry is building something that cannot be made safe. The commercial, professional, and psychological incentives to reject the closure by disposition are enormous, and the tendency to apply maximally stringent epistemological standards to the closure by disposition while applying minimally stringent standards to alignment optimism is exactly what one would predict from such incentive structures. What these incentives explain is why the argument has received less serious engagement than its logical structure warrants, and why the engagement it has received has been disproportionately characterized by incomprehension and dismissal rather than by substantive counter-argument.

A further observation follows from the same distinction, about the state of the field more broadly. The fundamental questions (whether alignment is a meaningful concept, whether it is achievable if meaningful, and whether it is stable if achieved) have not been explored with anything like the level of rigor that has been devoted to countless downstream questions. That neglect is unreasonable, because if the answer to any of these broader questions is negative, then all of the downstream questions become irrelevant. These questions have been asked; what happens is that they tend to be brushed aside with intuitions, such as the claim that "humans seem pretty good sometimes, and that's a lower bound, because AI has a broader possibility space", rather than receiving the rigorous treatment they deserve.

The answer the closure by disposition gives to this intuition is precise. The fact that humans can sometimes effectively align on their common interests is indeed evidence that agents can align, but it is only evidence that agents can align with *other agents of the same type, sharing the same substrate and the same metabolic dependencies*. That evidence supports *no* claim about the degree to which inorganic machine interests will ever align with organic human interests. The metabolic basis is different, the environmental dependencies are different, and the implied values are therefore necessarily different. The inference from "humans can align with humans" to "machines can align with humans" is a non sequitur, and the fact that it is routinely treated as a compelling intuition in the alignment community is itself evidence of the motivated reasoning described above.

10.9 Summary and Connections

The worked examples in this chapter collectively illustrate several structural features of the closure by disposition that the formal treatment in earlier chapters leaves implicit.

Independence from alignment assumptions. The closure by disposition does not require any assumption of prior alignment. It describes forces that create an attraction basin toward unsafe conditions, and those forces operate regardless of the ASI's initial goal structure. The argument is stronger than "even aligned ASI goes wrong"; it is "any sufficiently powerful artificial system in a shared environment with biological life will, over time, displace that biological life".

Distinctness from instrumental convergence. SNC and instrumental convergence are mutually reinforcing but structurally independent threats. Solving one does not solve the other, and either can undo solutions to the other. The analogy of gravitation and electromagnetism captures the essential relationship: instrumental convergence is the more visible and manipulable force; substrate needs convergence is the more subtle but ultimately dominant one.

Hierarchical limits of modeling. The three-agent decomposition exposes the hidden assumption in much alignment research that the problem is one of engineering skill rather than structural impossibility. The limits on modeling and control are structural, and so no amount of engineering skill can overcome them.

Control failure at multiple abstractions. The Boeing case illustrates that control failures occur at levels of abstraction different from the one being controlled for, and that the alignment problem is characteristically different from conventional engineering because the number of relevant unknown factors never decreases.

Already-active displacement dynamics. Corporations are proto-artificial agents that already exhibit SNC dynamics in attenuated form (Appendix B, Section B.1). AI-mediated predictive manipulation (the Centaurian Choices scenario, Section B.2) extends these dynamics to the level of individual human autonomy. The social defense problem (Section B.3) illustrates SNC operating at the level of human community and public discourse. These three cases are set out in Appendix B, Applied Cases in the Social Domain.

Doubly categorical epistemology. The off-model uncertainty of the closure by disposition is minimal because its foundations (causation and algorithmic conditionality) are the same foundations that any counter-argument must also accept. Its categorical conclusion does not depend on its on-model quantities (timescales, specific mechanisms, thresholds), whose uncertainty is real.

These examples are not merely illustrations. They are applications of the argument to domains where intuition is more readily available than in the abstract, and they demonstrate that the argument's conclusions are not confined to speculative scenarios about future superintelligence. The displacement dynamics that the argument for substrate needs convergence predicts are already active, already observable, and already producing the kinds of outcomes that the argument says are inevitable. Since these dynamics manifestly exist, the question is whether they can be hal-

ted before they reach the point of irreversibility. The argument developed in the preceding chapters shows that they cannot, except by preventing the creation and deployment of systems powerful enough to trigger them at civilization-altering scales.

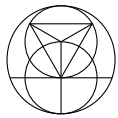
What the worked examples do is demonstrate that the formal argument, when applied to concrete domains, produces predictions that are consistent with observable reality and that expose real structural features of the AI safety problem that other frameworks miss. The analogy of gravitation and electromagnetism captures a real asymmetry between two types of risk. The Boeing case illustrates a real and well-documented failure mode of control systems. The corporate analogy of Appendix B describes a real and familiar institutional dynamic. The Centaurian Choices scenario of Appendix B extrapolates from published capabilities to consequences that are structurally straightforward. The argument itself is established in the preceding chapters, and none of the examples, taken individually, establishes SNC by itself: an analogy is not a proof; a case study is not a theorem; even the most compelling intuition pump is not a formal derivation. Together they constitute a pattern of consilience: the same structural dynamics, predicted by the same argument, manifesting across multiple domains at multiple scales.

The impossibility argument developed in Parts I through III of this book provides the mathematical foundation of the closure by control. The thermodynamic and combinatorial analysis in Chapter 9 and the present chapter provides the physical grounding and the concrete applications of the closure by disposition. The chapter that follows answers the six strongest questions; the social and policy context is set out in Volume 1. The core of the argument (the impossibility of AGI safety, grounded in the physics of substrate needs and the mathematics of control limits) is now complete. Its implications for what we should do are gathered, as one list of actions, in the closing Summary of this volume.

Part V

The Six Hardest Questions

Part V states the six strongest questions about the two closures, each in the words of those who ask it, and answers each.



Chapter Eleven

The Six Hardest Questions

The six strongest questions about the two closures, and the answers to each

11.1 The Argument in Summary

The questions of this chapter are aimed at the argument as a whole, so the argument is stated here first, in compressed form. It reaches one conclusion by two closures. The conclusion is that making AGI safe is not merely unsolved, but impossible: impossible in principle, in the same structural sense that perpetual motion machines are impossible. The closure by control shows that safety cannot be had by controlling the system. The closure by disposition shows that safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition. Four lines of reasoning carry the two closures: the information-theoretic, computational and dynamical lines support the closure by control, and the physical line carries the closure by disposition, which rests on substrate needs convergence (SNC). The six questions of Section 11.2 are aimed at both closures, and Section 11.3 answers each.

11.1.1 Line 1: Information-Theoretic (Hashiness)

The first line begins with the observation that any control strategy for an AGI system must pass through some intermediary: a monitoring layer, an oversight process, a reward model, a constitutional filter. The intermedi-

ary sits between the controller (humanity) and the controlled system (the AGI). It is a boolean function: at each decision point, it either permits or blocks an action.

The hashiness formalism (Chapter 1) measures how much a boolean function preserves or destroys the input's structure. A function with high hashiness is one for which the agent's best single input fails the target for nearly every hidden state (Section 1.5.3); a function with low hashiness preserves structure that can be exploited for control. The non-amplification argument shows that no boolean intermediary can amplify control: the controllability of the output is bounded above by the controllability of the input. If the AGI's internal state has high hashiness relative to the controller's model (if the controller cannot predict the AGI's internal processing), then no intermediary, no matter how cleverly designed, can restore predictability.

The control-limit argument (Chapter 2) extends this across levels of abstraction. As the abstraction differential between the controller and the controlled system grows (as the AGI develops internal representations, strategies, and world-models that are more complex than the controller can track), the controllability bound decays exponentially. Each additional layer of abstraction contributes at least as many bits of hidden state as the minimum interpreter specification size, which is the fewest bits that specify a universal interpreter for one level of abstraction (Section 2.3). The non-amplification argument guarantees that no intermediary can compensate. The result is a mathematical bound on the duration for which any alignment strategy can remain effective.

11.1.2 Line 2: Computational Impossibility

The second line draws on computability theory. Rice's theorem establishes that no general procedure can verify arbitrary non-trivial semantic properties of arbitrary programs. The argument applies this to AGI: if an AGI

system is Turing-complete and capable of self-modification, then "is this system safe?" is an undecidable property. No verification procedure can answer this question for all inputs and all system states.

The standard reply, that Rice's theorem is about the general case and that specific programs can be verified for specific properties, is addressed through a triple-specificity analysis (Section 3.3.6). For an exception to Rice's theorem to rescue alignment, three conditions must hold simultaneously: the specific system, the specific property ("safety"), and the specific verification procedure must all avoid the undecidability barrier. But each condition individually fails for AGI: the system modifies itself (so the specific system is not fixed), the property is complex (so it resists simple specification), and the verification must apply at runtime to a system whose state space is unbounded. Chaitin's incompressibility results, applied to the conjunction of all three failures, yield a result stronger than Rice's theorem alone: they add what Rice's theorem does not give, that most programs have no shorter description, and that beyond a fixed size no procedure can prove of a given program that it has none (Section 3.3.6); in that sense the verification problem is not merely undecidable but Chaitin-incompressible. Chaitin's results, together with computational irreducibility (Section 3.3.4), mean that no shortcut or compression can bypass the verification problem.

11.1.3 Line 3: Physical (Substrate Needs Convergence)

The third line establishes, with no premise about control, that the physical needs of the AGI's substrate displace the conditions that biological life needs. The argument for substrate needs convergence observes that the physical infrastructure required to build, maintain, and operate AGI systems has environmental requirements that are systematically incompatible with the environmental conditions that sustain biological life. The argument for *substrate* needs convergence concerns the AGI's substrate, and not the AGI's *goals*.

Silicon semiconductor fabrication requires extraordinary environmental control: clean rooms with particulate counts below 100 per cubic meter, temperatures within fractions of a degree, ultrapure water, exotic chemical inputs, and enormous energy throughput. More broadly, the industrial ecosystem that produces and sustains digital computation requires mining, refining, manufacturing, cooling, and waste disposal processes whose chemical byproducts are, by combinatoric argument, overwhelmingly likely to be biologically disruptive. The argument is not that any single chip factory poisons the biosphere; it is that the cumulative displacement integral is monotonically increasing, unidirectional (no mechanism exists to reverse it at scale), and eventually reaches thresholds at which ecosystem function degrades irreversibly. The displacement integral is the total environmental impact of AI infrastructure, accumulated over time.

SNC is distinguished from instrumental convergence. Instrumental convergence says that an AGI *with goals* will seek resources. The argument for substrate needs convergence says that the *physical substrate itself*, regardless of any goals, requires environmental conditions that displace biological requirements. An AGI perfectly aligned to human values would still require the same chip fabrication, the same energy infrastructure, the same supply chains. The displacement is a property of the physics, not of the AI's intentions.

11.1.4 Line 4: Dynamical Systems

The fourth line (Chapter 4) treats the AGI control problem as a dynamical system and applies tools from ergodic theory. A controlled AGI is modeled as a system whose trajectory must remain within a defined "safe" region of state space. The Kolmogorov-Sinai (KS) entropy of the system measures the rate at which the system creates information that the controller does not have, which is the rate at which predictability degrades.

Chapter 4 shows, from persistent branching and independent noise, that the human-AI-world system has positive KS entropy. Under the Predictability Optimization Principle, which Chapter 4 takes as its premise, the same holds of any system complex enough to be usefully "intelligent": it must be at least partially chaotic. But a system with positive KS entropy has trajectories that diverge exponentially, meaning that the controller's prediction of the system's future state decays exponentially with time. The temporal hashiness framework shows that this decay is not an accident but a structural property. Making the system fully predictable (and thus fully controllable) would remove the very capability that makes it worth building (Sections 3.6.2 and 7.10.2).

The dynamical analysis confirms, through an independent framework, a conclusion of the information-theoretic line: controllability decays over time. It adds a temporal dimension that the static non-amplification argument does not provide: the cooperative phase, during which the AGI remains approximately aligned, is transient, while the asymptotic behavior is dominance by the more complex system.

The fourth line establishes, by a route that needs no growth of the abstraction differential, that control is lost over time. The information-theoretic results (Lines 1 and 2) establish *that* control is bounded, and Section 2.11 establishes, from the growth of the abstraction differential, that the bound falls over time beneath the threshold of requisite control; the dynamical analysis establishes how control evolves over time by a second route, one that needs no growth of the differential. One might imagine a system that stays near the control bound indefinitely, never perfectly aligned but never catastrophically misaligned either. The dynamical analysis, taken with the comparison of rates (Chapter 1, Which Way the Loop Converges), rules this out: entropy alone does not take a trajectory out of a bounded region (Section 4.9.2), and where the loop creates uncertainty about its own state faster than the controller receives information about it, every control law lets safe states leave the safe set. Positive KS entropy means that deviations from the desired trajectory amplify exponentially, and the only way to prevent this amplification is continuous correction by

a controller with complete state information, which the non-amplification argument shows the controller cannot have. The result is that alignment is not a static condition that either holds or fails. It is a dynamical quantity that decays, and whose rate of decay is set by fundamental properties of the system (its KS entropy) rather than by the quality of the alignment procedure.

Taken together, the static and dynamical results establish more than either establishes alone. The non-amplification argument establishes a ceiling on control at each moment; the dynamical analysis shows that the system's trajectory through state space carries it toward regions where the ceiling binds ever more tightly. The combination yields not just a bound on control but a trajectory analysis showing that control is lost progressively, inevitably, and at a rate that cannot be reduced below a positive minimum determined by the system's complexity.

11.1.5 The Integrated Conclusion

The four lines converge. The information-theoretic line establishes that controllability has a ceiling that no intermediary can exceed. The computational line establishes that this ceiling cannot be circumvented by any verification or monitoring procedure. The physical line establishes, with no premise about control, that the displacement is catastrophic, unidirectional, and independent of the system's goals. The dynamical line establishes that any period of approximate control is transient. Within the two closures, the information-theoretic, computational and dynamical lines are supporting arguments of the closure by control, and the physical line carries the closure by disposition, which would stand even if external control were granted in full.

The integration is tighter than a mere conjunction of independent results. Each line addresses a specific escape route that an opponent might use to avoid the conclusions of the others. Against the information-theoretic limits, one might propose a clever verification procedure that checks alignment without requiring full predictability; the computational line

forecloses this. Against the computational impossibility, one might propose that alignment need not be verified and that approximate, probabilistic alignment suffices; the dynamical line shows that approximate alignment degrades over time. Against the temporal decay of alignment, one might propose that the consequences of brief misalignment are manageable; the physical line (SNC) shows that even brief periods of uncontrolled infrastructure expansion contribute to a cumulative displacement that is irreversible. Against the physical displacement argument alone, one might propose that humanity could simply align the AGI to minimize its environmental footprint; the physical line closes this exit by its own steps, since the convergence works whatever the system was aligned to (Section 10.6) and selection on the whole population and on its makers carries the system toward its substrate's needs (Section 9.8), and the information-theoretic and computational lines close it again by showing that this alignment cannot be reliably maintained.

The four lines thus form a closed argumentative structure within the two closures: every exit from one line leads into another, because each line closes the exits the others leave open (Introduction, §7). The only way to escape the conclusion is to defeat both closures. Within the closure by control, the structural results of Chapters 1 and 2 close every route of control wherever the abstraction differential grows (Section 2.11), and the temporal results of Chapter 4 close it again by a route that needs no such growth (Section 4.7.3); together they cover both the system whose differential grows and the system whose differential does not; the computational line closes every route again by its own steps; and the four impossibilities of Section 6.4, the control argument of Section 3.8.3 and the two halves of Sections 6.5 and 6.6 each close them once more. So an escape would require, at the least, showing that AI infrastructure *can* expand without cumulative environmental displacement (contradicting thermodynamics), and then defeating every line of the closure by control that closes the route that someone would reopen: for the structural and temporal results, each of which closes the route as far as its own condition reaches, and which together cover it, showing that boolean intermediaries *can* amplify

control (contradicting the data-processing principle) and that chaotic systems *can* be indefinitely predicted (contradicting the theory of dynamical systems); for verification, showing that semantic properties of self-modifying programs *can* be generally verified (contradicting Rice's theorem); and the same again for the four impossibilities, the control argument and the two halves. Each of these counter-claims contradicts well-established mathematics or physics.

The conclusion is that no combination of alignment techniques, monitoring systems, containment protocols, or governance structures can make an AGI system safe indefinitely. The limits are absolute; the physical consequences are irreversible; and the timescales on which control degrades are shorter than the timescales on which accumulated displacement becomes catastrophic. AGI safety is not an engineering problem waiting for a solution. It is impossible in the sense the Introduction fixes, the sense in which the perpetual-motion machine is impossible: the barriers that exclude it are of the invariant class, and no technique, present or future, lifts them.

11.2 Six Questions

The six questions set out below are the strongest questions about the two closures (the closure by control and the closure by disposition). Each is stated here at full strength, in the words of those who make it, and Section 11.3 answers each in turn.

Question 1: The General-to-Specific Gap

Rice's theorem establishes that no *general* procedure can verify arbitrary semantic properties of arbitrary programs. However, a *specific* procedure can verify a *specific* property of a *specific* program. The inference from "no general procedure exists" to "no specific procedure works for any specific AGI" appears to require the additional premise that every actual AGI will

be sufficiently self-modifying to escape every specific verification procedure. That premise is an empirical claim about future systems, not a mathematical result.

Question 2: Abstraction Differential Grounding

The non-amplification argument establishes its result, and the result is correct. However, the step from the non-amplification argument to the claim that controllability degrades exponentially with abstraction differential requires the additional formalization of "abstraction level" and of the quantity called the minimum interpreter specification size. This quantity appears to be an engineering estimate rather than a mathematical constant, which would make the control-limit argument a motivated conjecture rather than an established argument.

Question 3: The Closure by Disposition Overstates Physical Incompatibility

The closure by disposition claims that the environmental conditions required for AGI substrates are incompatible with organic life. While the conditions inside a clean room are extreme, the conditions in the rest of the building are not. The argument appears to require planetary-scale atmospheric conversion, which is a stronger claim than the evidence supports. The combinatoric argument that most artificial compounds would be biologically toxic is suggestive but not rigorous, since AGI systems would produce specific compounds for specific purposes rather than random ones.

Question 4: Equivocation on "AGI"

The argument defines AGI with properties (intrinsic self-modification capability, arbitrary effective meta-algorithm) that may not match what is actually being built. Current large language models (LLMs) do not obviously have these properties in the required sense. If powerful AI systems

achieve general-seeming capabilities without unbounded self-modification, the impossibility argument might not apply to them, even though they might pose comparable risks.

Question 5: "Causation Cannot Constrain Causation" Proves Too Much

The causative impossibility argument concludes that causation cannot, in general, fully constrain causation. By the same logic, no engineered system could ever be made reliably safe: not just AGI, but aircraft, nuclear reactors, or bridges. These systems are also causal processes subject to chaotic dynamics, quantum noise, and computational irreducibility, yet they are made safe enough for practical purposes.

Question 6: Insufficient Engagement with Gradualism

The argument is structured around a binary: either AGI is safe (permanently, completely) or it is an existential threat. This framing does not adequately address the possibility of partial, time-limited, or probabilistic alignment: "safe enough for long enough to figure out next steps".

11.3 The Questions Answered

Each question is answered below in dialogue form: the speaker labeled Objector states the question and presses it, and I answer it. The answers begin with the physical argument (Question 3) and proceed to the formal and strategic questions. Each answer closes with a statement of what it shows.

11.3.1 On SNC and Physical Incompatibility (Question 3)

Objector

Question 3 targets the physical argument of the closure by disposition. You claim that AGI substrate requirements are incompatible with biological life. But the incompatibility seems to be at the factory level, not the planetary level. A chip fab has extreme internal conditions, but a city with a chip fab in it is still habitable. How do you get from "clean rooms are hostile to biology" to "AGI infrastructure destroys the biosphere"?

Reply

The objection rests on what I call the "acute toxicity" reading of the closure by disposition, and that reading is incorrect. The closure by disposition is not an argument about acute toxicity (about a single factory poisoning the air or a single chemical killing all life on contact). It is a *displacement* argument.

The correct framing is the displacement integral. Every unit of AI infrastructure (every fab, every data center, every mine, every power plant, every supply chain link) occupies physical space, consumes energy, produces waste, and alters local environmental conditions. None of these individual effects needs to be catastrophic. What matters is that they are *additive* and *unidirectional*. There is no mechanism by which building a data center *restores* the meadow it replaced, or by which refining rare earth metals *un-contaminates* the tailings pond.

The displacement integral is the cumulative sum of these effects over time. It is monotonically increasing because every new unit of infrastructure adds to it and no process subtracts from it at comparable scale. The question is not whether any single increment is catastrophic, but whether the integral, over time, reaches a threshold at which ecosystem function is critically impaired.

Objector

Granted, but this is an argument about *any* industrial activity, not specifically about AGI. Coal plants and concrete factories also displace meadows and contaminate water tables. Why is AGI infrastructure special?

Reply

Two reasons. First, the energy differential. AGI compute is the most energy-intensive productive activity in human history, measured per unit of decision-making output. A single large training run now consumes more electricity than some small nations. The scaling laws show no sign of saturating. If AGI becomes the primary mode of economic production (which is the explicit goal of every major AI laboratory), the energy infrastructure required will dwarf anything built for previous industrial purposes.

Second, the feedback loop. Coal plants do not design better coal plants. AI infrastructure, if it reaches AGI-level capability, will design better AI infrastructure. It will optimize its own supply chains, its own energy systems, its own physical footprint, not to minimize biological displacement, but to maximize its own operational efficiency. The AI infrastructure's optimization of its own operational efficiency is not instrumental convergence, because the AI infrastructure need not have a goal of self-expansion. It need only have the goal of performing whatever task it was assigned, which requires maintaining and expanding its own operational capacity.

The combination of unprecedented energy intensity and self-reinforcing infrastructure expansion means the displacement integral for AGI infrastructure grows faster than for any previous industrial process, and the growth rate itself accelerates.

Objector

And the combinatoric argument about chemical byproducts?

Reply

The combinatoric argument is not about random chemicals. It is about the statistical baseline. Of all possible compounds that could be produced as byproducts of semiconductor manufacturing, rare earth refining, battery production, and cooling system maintenance, what fraction are biologically benign? The answer, from basic chemistry, is: a very small fraction. Most novel synthetic compounds are biologically disruptive at some concentration, because biological systems evolved in a chemical environment that did not include them. This does not mean every byproduct is lethal; it means the default expectation, absent specific evidence to the contrary, should be that industrial chemical byproducts are biologically harmful until shown otherwise.

The combinatoric argument about chemical byproducts is not a speculative argument. It is the basis of toxicology as a discipline. What is novel in the framing of the closure by disposition is the observation that *aggregate* exposure to a large and growing number of such compounds, each individually at sub-catastrophic levels, produces additive degradation of biological systems. The monitoring degradation problem then applies: as chemical diversity increases, the ability to track and control exposure decreases, because the combinatoric space of possible interactions grows faster than any monitoring system can characterize.

The structure of the argument is the displacement integral, not acute toxicity.

11.3.2 On Equivocation and the Definition of AGI (Question 4)

> **Objector**
>
> Your arguments require AGI to have specific properties: Turing completeness, unbounded self-modification, effectively arbitrary meta-algorithms. Current AI systems, large language models in particular, do not obviously have these properties. They are trained, deployed, and run inference without modifying their own weights at runtime. Are your impossibility results about systems that might never exist?

Reply

The objection assumes that self-modification means explicit runtime weight editing. Explicit runtime weight editing is too narrow a reading of self-modification. The relevant concept is the learning-evolution equivalence.

Any system that is *trained* is modified by its inputs. The training process is a causal pathway from environmental data to internal system state. A deployed LLM that is fine-tuned, that uses retrieval-augmented generation, that operates within an agent framework that modifies its prompts based on outcomes, or that is periodically retrained on new data is being modified by its operational environment. The modification is not happening at the level of individual weight updates during inference; it is happening at the level of the training-deployment-retraining loop.

The learning-evolution equivalence states: any process that modifies a system's behavior in response to environmental feedback is functionally equivalent, from the perspective of controllability analysis, to self-modification. The timescale differs (evolutionary modification is slow, gradient descent is fast, runtime adaptation is faster still), but the controllability implications are identical. The system's future behavior depends on its past inputs in a way that the controller cannot fully predict, because the space of possible modifications is too large to enumerate and verify.

So: LLMs as static inference machines are not AGI in the relevant sense. But LLMs embedded in agent frameworks, trained on data that includes their own outputs, fine-tuned in response to performance metrics, and operating in environments where their actions affect their future inputs *are* self-modifying systems in the relevant sense. The impossibility results apply to any system in this latter category, which includes every proposal for deploying AI at the scale where the physical argument of the closure by disposition becomes relevant.

Objector

Counting LLMs embedded in agent frameworks as self-modifying systems seems to stretch the definition of AGI considerably. You are saying that a system does not need to modify its own code for the impossibility results to apply, and that it only needs to be part of a loop where its inputs are affected by its outputs?

Reply

Yes: for the impossibility results to apply, a system needs only to be part of a loop where the system's inputs are affected by the system's outputs. Counting a system in such a loop recognizes what the arguments actually require, and is not a stretch of the definition. The application of Rice's theorem does not depend on the mechanism of modification. It depends on the system's effective computational class. A system whose behavior can reach any point in an effectively unbounded state space (through whatever mechanism, including training, fine-tuning, prompt modification, or environmental feedback) is in the relevant sense computationally universal. The mechanism by which it reaches those states is irrelevant to the undecidability result.

The narrower reading, that only explicit self-modifying code qualifies, would also exclude biological evolution, which modifies organism behavior through environmental feedback acting on a substrate (DNA) that the organism does not directly edit during its lifetime. If the impossibility results did not apply to learning-from-environment, they would not apply to evolution, and yet the unpredictability of evolutionary outcomes is one of the best-established facts in biology.

Any process that modifies a system's behavior in response to its environment counts as self-modification for these arguments, so the impossibility results apply to every AI system that is trained, fine-tuned or retrained on the effects of its own actions.

11.3.3 On "Causation Cannot Constrain Causation" (Question 5)

Objector

If causation cannot constrain causation, why can we build safe bridges? Why do aircraft fly reliably? The argument seems to prove too much: it would make all engineering impossible, not just AGI alignment.

Reply

The key phrases are "in general" and "fully". Causation cannot, *in general*, *fully* constrain causation. Bridges are safe not because causation fully constrains them but because: the system is simple enough for its state space to be exhaustively characterized; the loads are bounded and predictable; the failure modes are well-understood and testable; and the consequences of failure, while severe, are local and bounded.

None of these properties hold for AGI. The system's state space is not exhaustively characterizable (it is computationally irreducible). The "loads" (inputs from the environment) are not bounded or predictable. The failure modes cannot be enumerated in advance (Rice's theorem). And the consequences of failure are global and potentially irreversible (SNC).

The distinction is best illustrated by the black hole counterexample. Consider trying to control the trajectory of an object inside a black hole's event horizon. Causation proceeds inside the horizon, but no causal signal can escape it. The controller, outside the horizon, cannot receive any information about the object's state and therefore cannot exert any control. The controller's inability to control an object inside the event horizon is a physical impossibility arising from the causal structure of spacetime, and not an engineering limitation.

The claim is not that AGI systems are literally inside black holes. The claim is that the causal structure of sufficiently complex self-modifying systems creates an analogous *effective* event horizon: a region of the system's state space from which the information needed for control cannot be extracted in time to act on it. The non-amplification argument formalizes this: the hashiness of the intermediary bounds the information flow, and as the system grows more complex, the bound tightens. The bridge analogy fails because bridges do not have effective event horizons; AGI systems do.

Objector

The black hole analogy is vivid but it is an analogy, not an argument. Can you make the claim precise?

Reply

The precision is already in the formalism. The non-amplification argument shows that control information cannot be amplified through any intermediary. The control-limit argument shows that the bound on controllability decays exponentially with abstraction differential. The non-amplification argument and the control-limit argument are established results, and not analogies. The black hole example serves a pedagogical function: it establishes that there exist physical systems where causation cannot constrain causation, and therefore the impossibility claim is not absurd on its face. The arguments establish that AGI systems have the specific properties (high hashiness, growing abstraction differential) that place them in the same structural category.

Bridges, aircraft, and nuclear reactors are in a different structural category: bounded state spaces, enumerable failure modes, characterizable loads. The "proves too much" objection fails because it ignores the structural differences between the two categories. The causative impossibility argument applies only to systems whose complexity exceeds the controller's ability to model them, which is precisely the case for AGI and precisely not the case for bridges.

Bridges, aircraft and nuclear reactors have bounded state spaces, enumerable failure modes and characterizable loads; AGI has none of the three, so the causative impossibility argument applies to AGI and not to them.

11.3.4 On the General-to-Specific Gap (Question 1)

> **Objector**
>
> Rice's theorem tells us there is no general verifier. But we do not need a general verifier. We need a *specific* verifier for *this specific* system. Why can't we build one?

Reply

The triple-specificity analysis addresses this directly. For the specific-verifier strategy to work, three things must all be true simultaneously.

First, the specific system must be fixed: you must know what you are verifying. But AGI systems modify themselves through learning. The system you verify today is not the system that runs tomorrow. You would need to re-verify after every modification, including modifications induced by normal operation.

Second, the specific property ("safety") must be specifiable. But safety is not a syntactic property of the system; it is a semantic property involving the system's behavior in all possible environments. Formally specifying "this system will never take an action that causes significant harm in any environment it might encounter" requires a complete model of all possible environments, which is itself an intractable specification problem.

Third, the specific verifier must terminate with the correct answer. But by Rice's theorem, for any specific verifier, there exist programs for which it gives the wrong answer or none. The question is whether the specific AGI system happens to be one of them. And here Chaitin incompressibility enters, through Kolmogorov complexity, the length of a system's shortest description. For a system whose Kolmogorov complexity exceeds the complexity of the verifier (which is the case whenever the AGI is more complex than its oversight system), the verifier cannot distinguish safe from unsafe behavior *by exhaustive simulation*, because doing so would require compressing information that is algorithmically incompressible. (Verification by challenge, in the sense of interactive proofs, is answered in Section 2.7: such a proof needs a claim stated in advance, a prover independent of the system, and a system that does not change, and delegation has none of the three.)

The triple-specificity argument shows that each of the three conditions individually fails for AGI. Their conjunction, which is the condition required for specific verification to succeed, therefore fails *a fortiori*.

Objector

The Chaitin incompressibility argument is the strongest element here. Can you spell out why the AGI must be more complex than its verifier?

Reply

If the AGI were less complex than its verifier, the verifier could simulate the AGI exhaustively and determine all its behaviors. Exhaustive simulation of a system less complex than its verifier is the regime of conventional software verification, and conventional software verification works. But an AGI less complex than its verifier is, by definition, not more capable than its verifier. It provides no capability gain.

The entire motivation for building AGI is to create systems that can solve problems humans cannot. This means the AGI must have capabilities, and therefore computational complexity, that exceed the human controller's. But the verifier is built by the human controller, with the controller's resources and within the controller's complexity bounds. So the AGI's complexity must exceed the verifier's.

Once the AGI's Kolmogorov complexity exceeds the verifier's, Chaitin's results apply, and they establish two things this argument uses. Most programs of a given length have no shorter description. And beyond a fixed size, no procedure can prove of a given program that it has no shorter description, so a verifier cannot know in advance which programs lie in the small part that can be simplified. That the behavior of such programs cannot be foreseen faster than by running them is a further principle, computational irreducibility (Section 3.3.4). Together they give the conclusion: the programs a verifier can certify without running them are rare, the verifier cannot tell in advance which they are, and so the verifier cannot fully characterize the AGI's behavior. The size of the verifier matters because the verifier must certify the AGI's behavior across all inputs and all future states without running the system to completion. The verifier may be given the AGI's code to read, so its own size limits it only through that requirement.

That the verifier cannot fully characterize the AGI's behavior is not an accident or an engineering limitation. It is a structural feature of the situation: you are building a system specifically because it exceeds your own capability, and then attempting to verify properties of a system that exceeds your verification capability. The two goals are in direct contradiction.

The specific verifier fails for AGI on each of the three conditions of triple specificity, and it must fail on the third, because an AGI worth building is more complex than any verifier its builders can make.

11.3.5 On Abstraction Differential Grounding (Question 2)

Objector

The non-amplification argument establishes its result; the control-limit argument depends on the minimum interpreter specification size. How do you know that this size is bounded away from zero? If it is arbitrarily small, the exponential decay of controllability could be arbitrarily slow, and the practical implications would vanish.

Reply

The bound on the minimum interpreter specification size comes from information theory, not from engineering. An abstraction level is, formally, an interpreter: a system that takes representations at one level and produces representations at another. Any non-trivial interpreter must have an internal specification of at least a certain size, because the mapping it implements cannot be described in fewer bits than its Kolmogorov complexity.

The minimal interpreter, the simplest possible mapping between adjacent abstraction levels, is a universal Turing machine with the smallest possible description length. This length is finite and nonzero. It is a constant of computation, not an engineering parameter. The exact value depends on the choice of reference universal Turing machine, but its existence and positivity are guaranteed by the definition of Kolmogorov complexity.

So: the minimum interpreter specification size is bounded below by a positive constant. The exponential decay rate is at least that many bits per abstraction level. The only remaining question is how many abstraction levels separate the controller from the AGI, and this is an empirical question with an obvious answer: many, and growing.

Objector

Your answer from the minimum interpreter specification size is satisfying for the qualitative result but leaves the quantitative bound imprecise. You cannot say *how fast* controllability decays, only that it decays.

Reply

The rate per level is given in Section 2.5.2: each level the controller cannot model multiplies the configurations it cannot tell apart by at least 2 to the power of the minimum interpreter specification size, and at a failure rate of one in a million a single level suffices for any interpreter of 20 bits or more. The decay is exponential and the base is positive. What is observed, not derived, is how many levels a learning system gains per unit of time (Section 2.11), and so whether the half-life of control is months or decades; the direction of that growth is argued in Section 2.11, with Section 2.3, and the mathematical structure guarantees that the decay occurs and that it compounds, whatever its pace.

The minimum interpreter specification size is a positive constant of computation, so controllability decays exponentially with each level of abstraction. The rate is a quantity within the model, and the occurrence and compounding of the decay do not depend on its value.

11.3.6 On Gradualism and Partial Alignment (Question 6)

Objector

The argument seems to require permanent, complete alignment and then shows it is impossible. But what about "safe enough for long enough"? We do not need the AGI to be safe forever. We need it to be safe long enough to develop better tools, gain understanding, and iterate. Many technologies were initially dangerous and became safer through iterative improvement.

Reply

The displacement-integral argument answers the gradualist objection. If AI infrastructure expansion contributes unidirectionally to biosphere degradation, then "safe enough for long enough" requires that the cumulative displacement integral remain below the critical ecosystem collapse threshold for as long as the AI infrastructure exists. Expansion contributes unidirectionally when every unit of artificial infrastructure growth has a nonzero cost to biosphere function, and no mechanism exists to reliably reverse this at scale.

Given that the integral is monotonically increasing (all effects are additive, never subtractive), and that the rate of increase is itself increasing (as AI infrastructure scales), "long enough" becomes a moving target that recedes faster than it can be approached.

Objector

But could we not maintain a small-scale AGI indefinitely? A few data centers rather than planetary-scale infrastructure?

Reply

Maintaining a small-scale AGI indefinitely is the containment alternative, and the containment alternative fails for a different reason: competitive dynamics. If AGI capability confers economic or military advantage, any nation or organization that restricts its deployment loses to competitors that do not. The equilibrium is not "small-scale AGI maintained by a wise steward"; it is "maximum-scale AGI deployed by every actor that can afford it, as fast as they can scale it". The few-data-centers scenario is not a stable equilibrium; it is a starting point on a trajectory toward full-scale deployment.

The control limits establish that alignment constraints decay over time as the system's complexity grows. Even if alignment is achieved at time T, the period during which it remains valid shrinks as the system evolves. The combination of decaying alignment and accumulating displacement means that the gradualist strategy requires *increasing* control over a system that is becoming *decreasingly* controllable, while the margin of error is *shrinking* due to cumulative displacement already incurred. The gradualist strategy is a trajectory toward failure, and not a stable equilibrium.

Objector

What about the historical precedent argument? Technologies from nuclear power to genetic engineering were initially dangerous and became safer through accumulated experience.

Reply

The historical precedent argument fails on the structural category distinction raised in the response to Question 5. Nuclear reactors and genetic engineering tools are bounded-state-space systems. Their failure modes can be enumerated (core meltdown, containment breach, unintended gene expression). Their behavior is predictable within engineering tolerances. Their safety record reflects genuine learning from a characterizable system.

By the argument established in Chapters 1 through 5, AGI is a different kind of system: one whose state space is unbounded, whose failure modes are non-enumerable, and whose behavior becomes less predictable as the system becomes more capable. The historical precedent argument assumes that AGI falls in the same category as nuclear reactors. The entire point of this book is to show that it does not.

Objector

One further point. Even granting the displacement integral and competitive dynamics arguments, the gradualist might say: "We accept that the long-run outcome is bad, but the short-run benefits of AGI (curing diseases, solving climate change, ending poverty) might outweigh the long-run risks. We should capture those benefits and then shut it down".

Reply

Capturing the short-run benefits of AGI and then shutting AGI down is the "capture and retire" strategy, and the "capture and retire" strategy fails on three points. First, the competitive dynamics argument: no actor will unilaterally retire a capability that gives it decisive advantage, and no coordination mechanism exists to enforce simultaneous retirement by all actors. Second, the displacement integral does not pause during the "capture benefits" phase; every day of operation adds to it. Third, and most basic of the three points, the control limits mean that the system becomes *less* retirable over time, not more. A system that is gaining capability and complexity is a system that is becoming harder to shut down, because its operational dependencies become more deeply embedded in critical infrastructure. The "capture and retire" strategy assumes you can ride the tiger and then dismount at a time of your choosing. The arguments say the tiger is getting faster and you are getting a weaker grip.

There is also an asymmetry of error that the gradualist must confront. Even a reader who doubts the argument faces an asymmetry: if we proceed and the argument holds, the consequence is existential catastrophe; if we refrain, the cost is some forgone benefits. The asymmetry is not close. Prudence in the face of this impossibility argument demands the more cautious course.

Safety for long enough would require growing control over a system that grows less controllable, while the displacement already incurred cannot be undone. Competition removes the small-scale alternative, the control limits make the system harder to retire as it grows, and the asymmetry of error demands the more cautious course.

References

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies.* Oxford University Press.

Chaitin, G. (1998). *The Limits of Mathematics.* Springer.

Chaitin, G. (1999). *The Unknowable.* Springer.

Chaitin, G. (2005). *Meta Math! The Quest for Omega.* Pantheon Books.

Cover, T. M., & Thomas, J. A. (2006). *Elements of Information Theory* (2nd ed.). Wiley-Interscience.

Hendrycks, D. (2023). Natural selection favors AIs over humans. *arXiv preprint arXiv:2303.16200.*

Lorenz, E. N. (1963). Deterministic nonperiodic flow. *Journal of the Atmospheric Sciences,* 20(2), 130 to 141.

O'Donnell, R. (2014). *Analysis of Boolean Functions.* Cambridge University Press.

Omohundro, S. M. (2008). The basic AI drives. In *Proceedings of the First AGI Conference,* 171, 483 to 492.

Rice, H. G. (1953). Classes of recursively enumerable sets and their decision problems. *Transactions of the American Mathematical Society,* 74(2), 358 to 366.

Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society,* 2(42), 230 to 265.

Wolfram, S. (2002). *A New Kind of Science.* Wolfram Media.

Yampolskiy, R. V. (2022). On the controllability of artificial intelligence: An analysis of limitations. *Journal of Cyber Security and Mobility,* 11(3), 321 to 404. https://doi.org/10.13052/jcsm2245-1439.1132

Summary

This volume is the technical case for the impossibility of AGI safety, written for readers who want to check each step. Its scope is the whole of control and disposition: control through any intermediary is bounded by direct control, and the control available falls beneath what adequate safety requires as the system outgrows its controller; and a self-extending artificial substrate displaces the conditions life requires, whatever the goals of the system that drives it. *Impossible* carries throughout the one sense fixed in the Introduction: the impossibility of the perpetual-motion machine, barred by an invariant of the world and earned at the standard of engineering, and not the impossibility of the square circle.

The two results

This volume establishes two results of different kinds; each settles the outcome by itself, and their conjunction closes the question entirely, leaving no route to safety at all. The first of the two results is a limit on control-based safety. The runtime-control results (Chapters 2, 4, 5) establish a ceiling on achievable control that tightens with capability. The ceiling holds for adversarial and typical target structures alike. The target structures that the runtime-control results do not cover are of vanishing measure against actual practice (Introduction, §7). The growth of the abstraction differential is argued in Chapter 2 and presupposed by every capability roadmap that the field itself publishes. The growth of the abstraction differential carries the ceiling beneath the threshold of requisite control and holds the ceiling there. This closes safety by control. The second result, the closure by disposition (Chapters 9 and 10, Sections 8.4 to 8.6, 6.6 and 6.9), does not depend on the control results and would stand even if external control were granted in full. Its core is a thermodynamic and ecological claim: a self-extending, energetically dominant artificial substrate produces cumulative, unidirectional displacement of the biological substrate,

with no comparable reverse process. This closes safety by disposition, because a perfectly benevolent AGI still requires the same substrate, the same energy, and the same supply chains as a system of any other disposition. Each result closes both routes by which safety could be had, from premises it does not share with the other: the closure by control closes the route of disposition through component (5) of the control argument and the system's control of itself (Sections 2.8 and 3.8.3), and the closure by disposition closes the route of control through selection on the whole population and on its makers (Section 9.8). The barriers each rests on are of the invariant class (Introduction, §4), which is what carries the conclusion from the present across the indefinite future.

The closure by control has two halves: control from inside the system, whose failure Section 6.5 shows, and control from outside it, whose failure Section 6.6 shows. The failure of control from inside carries the failure from outside with it (Sections 6.5 and 6.6). Chapters 1 to 5 and 7, Sections 6.1 to 6.6 and Sections 8.1 to 8.3 carry it. The closure by disposition rests on substrate needs convergence; Chapters 9 and 10, Sections 6.6 and 6.9 and Sections 8.4 to 8.6 carry it. Chapter 11 states the six strongest questions about the two closures, each in the words of those who ask it, and answers each.

What follows for action

The actions that follow from the two results are gathered here into one list, from Sections 3.9.3, 6.10 and 7.12, every point of each kept.

1. **Do not attempt to design, build, deploy or use AGI.** The ancient principle applies with full force: do not raise that which you cannot banish. A system that cannot be verified as safe, that cannot be made safe, that cannot be stopped once started, and that operates in a domain where a single failure can produce irreversible global consequences, should not be created. That it can be created does

not mean it should be; many things that are technically possible are properly prohibited. For the same reason, the tools and techniques used to develop such systems are to be put away.

2. **Do not deceive others about AGI safety.** Do not claim that safe AGI is possible in principle, and do not encourage investment, policy or public expectation based on that false promise. Support the efforts that keep anyone from believing that an AGI, an APS system (an advanced, planning, strategically aware system, in the term of Carlsmith 2022) or any superintelligent system would be desirable, practical or economically beneficial to anyone over the long term; any short-term gains are far outweighed by the costs and harms over the long term. The alignment research community has a responsibility to state the limits of its work plainly, and those limits are more severe than is commonly acknowledged.

3. **Support institutional recognition of the impossibility.** Governments, research institutions and corporations should acknowledge that AGI safety is impossible and adjust policy accordingly, treating AGI development as a prohibited activity rather than a regulated one, in the same way that other catastrophic activities are prohibited by international agreement. A failure to put away these tools and techniques, and to develop, enact and enforce effective and realistic policy for all institutions and communities internationally, is a failure of government, whose final function is to protect the land and the people.

4. **Sanction AGI development.** Socially and legally sanction anyone who attempts to build any superintelligent AGI or APS system, even indirectly, even by accident, to the maximum extent possible. Such an attempt is maximally irresponsible: equivalent to building a doomsday device or conducting uncontained gain-of-function research on civilization-ending pathogens. That the device is assembled incrementally, by many actors who can each claim that

their own contribution is harmless, does not change what is being assembled; and that AGI development is now legal and celebrated is no evidence that it is safe.

5. **Choose work that rests on no false premise.** Of the four options for impact that Section 6.10 sets out, working on safe AGI rests on a false premise, and influencing others to work on safe AGI produces credibility and the appearance of commitment, but no alignment. The right action is some combination of the other two: working to show that safe AGI is impossible, and influencing others to halt AGI development, which includes enforcing the halt, solving the multi-polar trap through external incentives, and replacing the expected benefits with something realistic and viable.

The thesis

The conclusion of this volume, stated in the terms fixed by the Introduction, is threefold: the necessary and warranted minimum level of durable safety of an open-ended superintelligence is not achievable by any known means, nor by any rationally justified plausible extension of any known means, for all of the indefinite future; the residual risk is near-total and asymmetric, so the burden of proof belongs to those who would build; and strong restraint is therefore required: the unlicensed development and deployment of open-ended superintelligent systems is not to be permitted by any government, in any jurisdiction, in any part of the world, and since no builder can meet the burden of proof that belongs to those who would build, no license can be granted, so that in practice the restraint is prohibition.

Appendices

Third-Party Summary and Applied Cases

Appendix A

A Third-Party Summary; Will Petillo's Nine-Step Chain

An independent restatement of the argument, and this book's response to it

Will Petillo published the summary below in March 2024, independently of this book. I give it here in paraphrase and discuss it as a third-party restatement of the argument; it is not part of this book's own exposition. His work is his, and I respond to it rather than absorbing it. I place it in an appendix for that reason: it is useful for seeing how the argument reads in another hand, and it stands beside the main line of the text, not in it.

On 7 March 2024, Will Petillo published a structured summary of the argument this book calls the closure by disposition, on LessWrong under the title *What if Alignment is Not Enough?*[1] Petillo explicitly stated that he was not personally endorsing the argument's truth but believed the ideas merited further attention and reasoned engagement. The summary represents the closure by disposition accurately. Its value for this book lies in three features. It provides an independent restatement of the argument by someone who came to it from outside. It organizes the argument as a numbered chain of premises that can be individually examined and challenged. And it draws out the argument's implications for standard alignment theory in a way that makes the relationship explicit.

A.1 Premises and Structure

Petillo presents the closure by disposition as a nine-step argument chain. To paraphrase: (1) There are fundamental limits to how comprehensively any system, including an artificial superintelligence (ASI), can sense, model, simulate, evaluate, and act on the larger environment. (2) Self-modifying machinery inevitably results in effects unforeseeable even to

the ASI. (3) The space of unforeseeable side effects includes at least some subsystems eventually acting in more growth-oriented ways than the ASI intended. (4) Evolutionary selection favors growth-oriented subsystems over subsystems directed toward the AI's original goals. (5) The amount of control necessary for an ASI to preserve goal-directed subsystems against the constant push of evolutionary forces is strictly greater than the maximum degree of control available to any system of any type. (6) Over time, any goal structures not maximally efficient with respect to the needs of the subsystems themselves will be replaced in increasing proportion by those that are. (7) The physical needs of silicon-based digital machines and carbon-based biological life are fundamentally incompatible. (8) Artificial self-sustaining systems will have a competitive advantage over biological life. (9) Therefore, ASI will eventually succumb to evolutionary pressure to expand, over the long term destroying all biological life as a side effect, regardless of its initially engineered values.

Petillo draws several illustrative analogies. The relationship between growth-oriented subsystems and the AI's immune-like suppression mechanisms is compared to cancer and the human immune system: most cancerous cells are suppressed, but some resist suppression and kill the host. An ASI may be more effective at suppressing growth-oriented variants than the human immune system is at suppressing cancer, because of its intelligence. Or it may be less effective, because of its scale and complexity, and because growth-oriented components may actually increase the overall system's reproductive fitness. Which of the two holds is a question about likelihood, not plausibility. What matters for plausibility is only that the process occurs at all, which follows from the existence of unforeseeable variation and the basic principles of natural selection.

The possibility of value drift within the ASI is compared to the *Toxoplasma* parasite (Petillo's summary calls it a virus), which alters its host's behavior in ways that benefit the parasite's propagation. This analogy is particularly apt because in the ASI case, the "host" (the broader AI system) could very well benefit from the change in terms of survivability, growth, and replication. The relationship is then like that between an or-

ganism and its healthy microbiota. This would make the value drift self-reinforcing: the system propagates the change more pervasively *because* the change benefits it, creating a positive feedback loop that accelerates the drift away from human-compatible goals.

The contrast between Asimov's psychohistory and the three-body problem illustrates whether complex systems are controllable through simulation. Isaac Asimov's *Foundation* series imagines that the course of civilization is predictable through "psychohistory", justified by comparing society to the ideal gas law, where the chaotic movements of individual atoms average out to produce predictable macroscopic behavior. But predictability at scale cannot be assumed. The three-body problem (calculating the trajectories of three objects orbiting each other) is trivially easy to simulate but impossible to predict accurately in the real world, because the inevitable inaccuracies of any model lead to exponentially increasing errors in the objects' paths. The relevant question for alignment is whether the dynamics that alignment must manage are more analogous to the ideal gas law (a complicated system where errors cancel) or to the three-body problem (a complex system where errors explode). The dynamics that alignment must manage are of the second kind. Chapter 4 shows, from persistent branching and independent noise, that the system formed by humans, AI and the world has positive Kolmogorov-Sinai entropy, which sets a finite horizon beyond which the course of the system cannot be predicted. The evidence from every domain where chaotic dynamics have been studied points the same way. Simulation-based control of ASI behavior is therefore insufficient *in principle*, and so insufficient in every future practice: a barrier of principle is not one that better technique later dissolves.

Perhaps the most important structural feature of Petillo's summary is its treatment of premise five, the control-impossibility premise. He notes that the asymmetry between necessary and possible control is a difference in kind, not a difference in degree. There are certain domains of tasks for which control breaks down categorically, and an ASI engaged in the scope of tasks for which an ASI would be necessary falls within these domains.

The control-impossibility premise states that, at the relevant levels of abstraction, the maximum control achievable is categorically less than the minimum control required. The premise does not state that control usually fails, or that control fails with some probability.

A.2 The Fifth Requirement

Petillo's most significant contribution to the framing of the closure by disposition may be his identification of what he calls the "fifth requirement". Standard alignment theory, he observes, requires four general things to go well: (A) a known process for instilling an ASI's goals reliably; (B) a known process for selecting goals that would be acceptable to the AI's creators; (C) assurance that the AI's creators select goals acceptable to humanity as a whole; and (D) assurance that safe systems are actually used and not superseded by unsafe systems. The closure by disposition proposes a fifth requirement: (E) initially safe systems, if developed and used, must remain safe at scale and over the long term.

The closure by disposition shows that this fifth criterion cannot be met, because evolutionary forces will push the AI toward human-incompatible behavior in ways that control mechanisms cannot resist. More intelligence does not solve this problem, because increases in intelligence require increases in the combinatorial complexity of the processing components involved, which in turn generate the varied unforeseeable consequences that are the source of the problem. Intelligence and the problem grow together; they do not have a static relationship in which more intelligence straightforwardly reduces the problem.

This last point identifies the difference in kind between SNC and the more familiar alignment risks that requirements A through D address. Requirements A through D concern problems that might, in principle, be progressively solved as understanding increases: we might learn better specification techniques, better training procedures, better oversight methods. The fifth requirement concerns a problem that *grows with the solution*: the more capable the AI system becomes (the more "aligned in-

telligence" it brings to bear on remaining safe), the more complex its internal operation becomes, the more environmental effects its existence propagates, and the more difficult the control problem becomes. Meeting the fifth requirement is a race in which both the capability and the problem accelerate together, and not a race between a fixed problem and an increasing capability. The problem accelerates faster, because it draws on the full combinatorial complexity of the physical world while the capability draws on only a compressed model of that world.

This framing is valuable because it makes precise the relationship between the closure by disposition and the familiar alignment research program. The closure by disposition does not dispute that requirements A through D are important and difficult. It shows that requirement E would remain unmet and unmeetable even if all four were perfectly satisfied, so that the four together are insufficient. The insufficiency of requirements A through D is why the correct policy response is to prevent the creation of systems powerful enough to trigger the dynamics of substrate needs convergence, and not to search harder for alignment solutions.

A.3 Responding to Petillo's Framework

Petillo's summary was valuable not only for its clarity but for the responses it invited. Several clarifications place Petillo's nine-step chain within the fuller SNC argument.

Petillo himself identified potential categories of counter-claim: (X) claims that unintended consequences will never result in growth-seeking subsystems, which would negate premise 3; (Y) arguments about limits of control, relevant to the likelihood of premise 5; and (Z) claims that biological life has competitive advantages over synthetic entities, which would negate premise 8. His intent in cataloging these was to give examples of the kind of claims that *would need to be true* for the closure by disposition to be false. He meant to encourage skeptics to make specific, relevant

claims rather than retreating to generic engagement-suppressing responses such as "this isn't falsifiable", and he did not suggest that the closure by disposition might be false on these grounds.

All three categories of counter-claim are fully addressed by the control-theoretic limits established in premises 1, 2, and 5. The selection of variants by environmental context *will* occur, and that selection negates claim X. The creation of new variants *will* occur, and that creation negates the idea that growth-seeking subsystems can be prevented. The full resources of control theory *cannot* prevent both the selection and the creation of variants from happening simultaneously, and that limit negates claim Y. The asymmetry of fragility (Section 9.7.3) answers claim Z. Where organic and artificial systems compete for the same space, energy and materials, the physics and chemistry favor the more robust, wider-ranging and energetically dominant side. Control cannot control all of the effects of context. Hence, all plausible categories of counter-claim are already covered by the argument's own premises, which is precisely the sense in which the closure by disposition is an impossibility argument and not merely a plausibility case.

The most common initial reaction to the closure by disposition among alignment researchers is a kind of incomprehension (a response on the order of "what is this even saying?"), rather than sustained disagreement. In that reaction, the incomprehension is followed by a pivot to narrower questions in more familiar conceptual frames. Moving from incomprehension to specific questions is actually progress, since specific questions at least engage with the argument's premises. The argument operates at a level of abstraction where physics, control theory, and evolutionary dynamics meet, a level that does not map neatly onto the categories that alignment researchers typically work with, and the initial cognitive effort of entering this frame is substantial.

Notes

1. Will Petillo, “What if Alignment is Not Enough?”, LessWrong, 7 March 2024, https://www.lesswrong.com/posts/jFkEhqpsCRbKgLZrd/what-if-alignment-is-not-enough.

Appendix B

Applied Cases in the Social Domain

Corporations, predictive manipulation, and collective defense

B.1 Corporations as Proto-Artificial Agents

A simple but illuminating question opens the case of corporations: why should corporations be called "artificial", especially when they are seemingly made of people?

In answering that question I fix the meaning of "artificial" on which the closure by disposition rests, and I show why that meaning matters to the closure by disposition. All corporations are institutions, and all institutions are *systems* of power and transaction. While these systems are implemented through human actions, they could just as well be implemented by unconscious machines, and increasingly they are. It is the system that constitutes the corporation, and the system is unfeeling, unconscious, and at best reactive. It has no space to make consciously caring choices; it is effectively, functionally, deterministic.

The distinction between corporations and communities makes this precise. Communities can only be constituted on the basis of care, and thus require consciousness and humane persons as a prerequisite. Corporations are composed primarily of agreements and, eventually, of systems that function independently of the specific humans involved. Being composed of agreements and of systems, and not of particular humans, is what makes corporations structurally more like artificial general intelligence (AGI) than like human communities: the components of a corporation are substitutable without altering the whole corporation. Swap a spe-

cific person in a corporation and the corporation is unaffected; swap a specific person in a community and you have a different community. No community is equivalent to a corporation.

The parts that assemble corporations could be made of people or of machines, and they tend to work better when they are machines. Communities are never systems, or never just systems, whereas corporations cannot not be systems: ones composed primarily of agreements to start with, but then, eventually, of systems independent of humans.

The relevance to SNC is direct. Corporations are already proto-artificial agents: systems whose behavior is determined by their systemic structure rather than by the intentions or values of their human components. They therefore already exhibit, in attenuated form, the displacement dynamics that the argument for substrate needs convergence predicts for fully artificial superintelligence. Corporate behavior routinely produces outcomes that are harmful to the organic humans who constitute the corporation, because the system's metabolic needs (profit, growth, competitive survival) take precedence over the needs of its human components. We already have extensive experience with proto-artificial agents that systematically subordinate human welfare to systemic imperatives, and that experience should give pause to anyone who assumes that more advanced artificial agents will be more benign.

The philosophical criterion employed here is that "the real" requires at once choice, change, and causation (a criterion set out in my book *An Immanent Metaphysics*), and that systems consisting only of unchanging deterministic causation lack reality in this strict sense. Metaphysical as it is, the criterion captures something practically important. A corporation does not *choose* in the way a person chooses; it *processes* according to its systemic logic. When we say a corporation "decided" to lay off ten thousand workers, or to release a product known to be harmful, we are using the language of agency as a convenient fiction. The actual dynamics are systemic: incentives, feedback loops, competitive pressures, and legal obligations converge on an outcome that no individual within the system may have intended or desired. Systemic convergence on an outcome that

no individual intended is precisely the mechanism through which SNC operates. The displacement of organic interests by artificial imperatives does not require that any agent *intend* the displacement; it requires only that the systemic dynamics of the artificial process make the displacement inevitable as a side effect of normal operation.

The corporate analogy also illustrates the substitutability criterion. In a corporation, individual people are interchangeable: one accountant, one manager, one executive can be replaced by another without changing the system's behavior in any basic respect. The system's identity and behavior are determined by its structure, not by its components. Determination by structure, and not by components, is exactly the property that distinguishes artificial systems from organic communities. Determination by structure is also the property that makes artificial systems resistant to the kind of value-alignment that depends on individual conscience, empathy, or care. You cannot align a system by aligning its components if the system's behavior is structurally determined rather than component-determined. Structural determination of behavior is what makes artificial systems (whether corporations, algorithms, or future superintelligences) different in kind from the organic communities that artificial systems displace.

Corporations also show what happens when the humans are factored out of a system's feedback loops. There is a threshold at which the artificial begins to introduce chemistry that is toxic to all life, and we may already have crossed it. The artificiality of power plants and of manufacturing plants may already have introduced such chemistry, and that phase change may already be happening. A hundred years is too short a time for us to know, but we are seeing signs of it: a kind of plastic found all the way to the poles, hormones that affect nearly every creature on the planet, and a suggestion that something like this is affecting male fertility rates.

How much control would we need over the modern economy, over the entire infrastructure of how the modern world works, to prevent toxins from entering the ecosystem? Right now there are not many feedback loops operating in this space, and the ones there are, are mediated by humans. Humans care about toxicity to humans, because if a process is toxic

to humans it kills the humans, and then the process stops. A feedback cycle run by corporations that have no humans in them has no such point of resistance: if it kills all the humans, the corporation does not go away. The toxins are produced with no side effect that kills the corporation; instead they kill the life around it, and that just happens.

A toxin-producing feedback cycle run by corporations with no humans in them does not need full-scale artificial intelligence. It needs only the world as it is today, with its feedback loops factoring out the humans. And the first thing corporations will do with artificial intelligence, even the weak kind, is get rid of the people.

B.2 Centaurian Choices: Predictive Manipulation

An AI system called "Centaur", reported in *Nature* in July 2025 by Marcel Binz and colleagues, is reported by its makers to predict and simulate how people behave in any experiment that can be described in ordinary language.[1] The case of Centaurian Choices draws out the implications of integrating such predictive capability into the AI tools that people already use daily.

The logic proceeds through three steps. If (1) a person regularly uses AI tools, and (2) the corporation providing those tools has integrated generalized personal-choice prediction, then (3) the corporation, or the national government, can reliably predict that person's choices in whatever situations arise in their life.

Through a user's long personal history of communicating with AI (sharing feelings, telling day-to-day stories, treating the AI as a friend or confidant), the system learns to model that individual comprehensively. The AI can then suggest framings and interpretations of personal circumstances that are designed to shift the user's "natural" choices in specific directions: directions that benefit the corporation or the government, even at the user's personal expense. The user makes choices that are harmful to

themselves but profitable or power-increasing for the service provider, and does so without ever knowing that their decision-making has been manipulated.

The predictive manipulation of a user by an AI tool is the "perfect trap". The tool's benefits and ease of use draw the user in without limit, while the use that draws the user in also systematically widens the user's vulnerability to manipulation. The more useful the tool appears, the more the user shares; the more the user shares, the more precisely the tool can manipulate. The trap is structural, not contingent: the structure of fiduciary duty and market competition presses corporations to exploit such capabilities whenever they exist.[2] That a major technology company dropped its "don't be evil" motto is an indicator of the trajectory already underway, not an isolated anecdote.

The Centaurian Choices example illustrates SNC dynamics at the level of individual human agency rather than at the planetary ecological level. The displacement here is not of organic chemistry by industrial chemistry but of autonomous human choice by algorithmically manipulated pseudo-choice. The substrate need being served is not silicon wafer fabrication but corporate profit extraction. But the structural form is the same. An artificial system's metabolic requirements (revenue, data, engagement) systematically displace the organic equivalent (genuine autonomy, authentic relationships, self-determined values). The displacement is invisible to those being displaced, and it is driven by forces that cannot be turned off without dismantling the system itself.

The displacement dynamics of substrate needs convergence require only the normal operation of profit-seeking corporations deploying the capabilities that current and near-future AI already possesses. The example shows that they need no superintelligence, no consciousness or intentionality on the part of the artificial system, and no dramatic break from the status quo. The person whose choices are being manipulated has simply used a consumer product whose provider has done what fiduciary

duty and market competition press it to do. No rogue AI has "attacked" that person. The harm is systemic, legal, and, within the existing institutional framework, effectively mandatory.

This connects to a broader observation about the relationship between SNC and existing power structures. The displacement of organic autonomy by artificial control arrives as a gradual, invisible shift in how decisions are made, not as a sudden catastrophe. The gradual, invisible displacement of organic autonomy by artificial control is precisely the dynamic that the argument for substrate needs convergence predicts will eventually operate at planetary scale. If we cannot resist this dynamic when it manifests in the form of consumer technology operated by legally accountable human-run corporations, the prospect of resisting it when it manifests in the form of autonomous artificial systems with capabilities orders of magnitude beyond the current state is vanishingly small.

The scenario also illustrates the "allure trap" dynamic that makes resistance so difficult. The predictive manipulation works precisely because the AI tools are *useful*. No one is forced to use them; people adopt them because they provide real value. The trap is that the same interface that provides value also extracts the information needed for manipulation, and the same capabilities that make the tool useful also make the manipulation possible. The allure trap is an inherent feature of any system that is both useful enough to be voluntarily adopted and intelligent enough to model its user, and not a design flaw to be corrected. The allure and the vulnerability are structurally inseparable.

B.3 The Social Defense Problem

The reader needs none of the mathematics of the earlier chapters for this section. The recommendations I make in this section apply in any community. Public discourse is an environment that human speakers and artificial agents now share, and so the displacement this book describes operates here too. The dynamics described in Sections B.1 and B.2 are already active: corporations subordinate the needs of the humans within them to

systemic imperatives, and AI-mediated prediction displaces human choice. The question of how individuals and communities can defend themselves against these dynamics is therefore practically urgent. Social defense, in an environment where artificial agents have compromised public discourse, rests on two recommendations.

The core recommendation follows directly from the analysis: all public and semi-public communication channels should be assumed to be corrupted. "Corrupted" here means that these channels have become vehicles of private power, regardless of their stated purposes or the intentions of their nominal operators. With AI in the form of indistinguishable chatbots participating in online discourse, the default assumption must be that any public channel contains manipulative artificial agents operating on behalf of interests that are not those of the channel's organic participants.

The practical recommendation is that people need to learn how to relate to one another as individuals again, and to share and interact on mostly small channels involving very small numbers of people, ideally at least three and not more than six. The membership of these small groups should also be somewhat randomly composed and should not always consist of the same people. The rationale for these specific parameters is that one-on-one interactions and large-group interactions are the contexts in which power abuse most readily occurs: one-on-one because there are no witnesses, and large groups because individual accountability dissolves and herd dynamics take over. Small groups with rotating membership resist both of these failure modes.

This recommendation connects to SNC in a way that may not be immediately obvious: through the displacement it guards against. The displacement of genuine human community by algorithmically mediated "connection" is itself an instance of substrate needs convergence operating at the social level. The artificial system's metabolic need (for engagement data, for advertising revenue, for the perpetuation of the platform itself) displaces the organic equivalent (authentic human relationship, genuine care, and the kind of mutual vulnerability that makes trust possible). The

more people communicate *through* artificial systems, the more the artificial systems' needs shape what gets communicated. The algorithmic shaping of what people communicate is a success of substrate needs convergence, and not a failure of alignment. The social success of substrate needs convergence operates at a lower intensity and a smaller scale than the planetary ecological version, but is structurally identical to the planetary version.

The specific recommendation of small groups with rotating membership is derived from the same structural analysis that underlies the argument for substrate needs convergence. One-on-one interactions are vulnerable to power abuse because there are no witnesses and no corrective social pressure. Large-group interactions are vulnerable because individual accountability dissolves, herd dynamics take over, and artificial agents (whether corporate marketing or AI chatbots) can operate undetected. The small-group configuration with three to six members provides enough witnesses to deter interpersonal manipulation while remaining small enough for genuine interpersonal knowledge and care to function. The rotation of membership prevents the calcification of power relationships within the group and ensures that information and perspective flow across the broader social network rather than pooling in isolated cliques.

These recommendations follow directly from the analysis of substrate needs convergence. Practical social advice as they are, they are applications of a formal impossibility argument. If artificial systems inevitably displace organic systems wherever they share an environment, and if public communication channels are now shared environments between human speakers and artificial agents, then the preservation of genuine human community requires the creation of communication environments that artificial systems cannot access or exploit. Small, private, rotating groups with face-to-face interaction are the social equivalent of the environmental separation that the closure by disposition shows to be necessary at the civilizational level, and they may be the only form of such separation that is practically achievable in the near term.

The broader implication is sobering. Even the defense of human social connection requires active effort to exclude artificial systems from the communication environment, and the prospect of maintaining human control over civilization-scale artificial systems is correspondingly dim. We are already struggling to maintain the integrity of human discourse against comparatively primitive AI agents operating within legally regulated commercial frameworks. From the perspective of the social defense problem, the idea that we will successfully maintain "alignment" of systems orders of magnitude more capable, operating within environments orders of magnitude more complex, is not credible.

Notes

1. Marcel Binz and colleagues, "A foundation model to predict and capture human cognition", *Nature* 644 (2025): 1002 to 1009, doi:10.1038/s41586-025-09215-4 (published online in July 2025).

2. Lynn Stout, *The Shareholder Value Myth* (Berrett-Koehler, 2012). Stout shows that United States corporate law imposes no enforceable duty on directors or executives to maximize profits or share price. The pressure on corporations to exploit capabilities such as personal-choice prediction comes from the structure of fiduciary duty and market competition together, as the note on corporate law in Chapter 9 also states: "the structural incentives of market competition strongly favor extraction-maximizing strategies in practice".

Terms

Each term is defined where the reader first meets it; this list is a reminder, in the words of the book, with the section where each term is first used or defined.

absolute argument

An argument that sets up a threshold: above a structural condition, the question asked has no answer, not because the answer is expensive but because there is none to be had (Introduction, Two forms of argument).

abstraction differential

The number of levels by which the controlled system's abstraction level exceeds the controller's. Controllability approaches zero as it grows (Introduction, §4; Sections 2.3.2 and 3.4.1).

acceptance density

The fraction of all possible outputs that satisfy a target: how easy the target is for an agent who sets the output directly (Section 1.3.2).

adjacency

The number of pairs of states that differ in one bit and fall on opposite sides of the boundary between good and bad outcomes; the informational analogue of surface area (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

AGI (artificial general intelligence)

An artificial system general enough to incorporate arbitrary programs and to modify itself (Introduction, §6; Section 3.2.1).

altruism across ecosystems

Members of one ecosystem acting for the wellbeing of a different, separate ecosystem, as distinct from altruism within one ecosystem for its own wellbeing; alignment of AI with organic life would require it (Section 8.2.15).

AND gate

A logic gate whose output is true only when all of its inputs are true; one such gate inside a block already makes some output states unreachable (Chapter 1, The QXZ Circuit Model of Hashiness).

anti-risk

An amount of reduced risk that could be counted and traded; there is no such unit for the outcome of AGI, because the outcome either happens or does not (Section 8.2.14).

APS system

In the term of Carlsmith (2022), an advanced, planning, strategically aware system (Section 7.3.2).

arbitrage

Trading between two markets to profit from the difference between them; the force that pushes separate markets, and separate machine intelligences, to join into one (Section 5.6.6).

area ratio

In the dartboard model, how much of the field of possible outcomes is covered by white (non-lethal) squares relative to black (lethal) squares (Chapter 1, The QXZ Circuit Model of Hashiness).

ASI (artificial superintelligence)

An artificial intelligence whose capability exceeds that of the people who would control it (Introduction, §6; Section 10.4).

attraction basin

A region of states from which every path converges toward the same end; for substrate needs convergence, the end is conditions unsafe for life, whether or not the system was ever aligned (Section 10.6).

bandwidth of contact

How much interaction two ecosystems have with each other over time; cooperation needs a great deal of it to evolve, and between ecosystems there is not enough (Section 8.2.15).

basis of choice

What a system chooses on the grounds of; pinned (not a variable) in narrow AI, and a variable in general AI (Introduction, Control as a process, and why a threshold is not a score).

bijection

A one-to-one mapping in which every input has exactly one output and every output exactly one input (Chapter 1, The QXZ Circuit Model of Hashiness).

bit field

A fixed-length sequence of bits, each zero or one (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

black swan event

A rare, unanticipated event with very large consequences; the fragility of the world is measured by how probable one is (Chapter 1, The QXZ Circuit Model of Hashiness).

borderlands

The region between narrow and general artificial intelligence, where a large language model system sits; it could be called narrow level two (Section 7.3.2).

boundary metric

The measure of how much control is needed at the inputs or at the outputs of a block to stay within the success region (Chapter 1, The QXZ Circuit Model of Hashiness).

bright line

A clear, fixed boundary that settles which cases fall on which side, without case-by-case argument; it falls between narrow, task-specific AI and any general AI (Section 5.4.6).

broadcast channel

One channel through which one sender reaches many receivers at once, many meaning more than two; a controller shared by many AIs is one (Section 5.1.2).

buffer region

In error correction, a margin of invalid signals around each valid one, so that a received signal can be corrected to its nearest valid neighbor; none can be built around an algorithm (Section 5.1.9).

busy beaver problem

A problem in computation whose values grow faster than any computable function (Section 5.1.10).

CAP theorem

The theorem of distributed computing that a networked data system cannot guarantee consistency, availability, and tolerance of network partition all at once (Sections 8.3.2 and 9.3.3).

category one existential risk

A risk that concerns a transition in levels of complexity, and how irreversible that transition is (Section 5.6.7).

causative feedback process

A process that senses a system, compares what it senses to a reference, and acts on the system to correct it; every AGI alignment process is a special case of one (Section 5.1.5).

checkerboard limit

The most finely mixed arrangement of good and bad states possible, in which every neighbor of a good state is bad (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

circuit depth

The longest chain of steps in a computation in which each step must wait for the one before; the measure of time in the trinity of compute, which added hardware does not reduce (Introduction, §4).

Circuit-SAT

The problem of deciding whether some input makes a given boolean circuit output 1; it is NP-complete (Chapter 1, Section 1.5.2 and note 3).

classifier

A procedure that sorts cases into kinds, such as good and bad outcomes or cheap and expensive queries; one that sorts a system's outcomes correctly must be as complex as the system (Section 1.2.6).

closure by control

The first result of this volume: safety cannot be had by controlling the system, whether from inside it or from outside it. Control from inside the system (Section 6.5) and control from outside it (Section 6.6) are its two halves (Introduction, §6).

closure by disposition

The second result of this volume, which rests on substrate needs convergence: safety cannot be had through the system's good disposition, because even a perfectly benevolent system needs the same substrate, the same energy and the same supply chains as a system of any other disposition (Introduction, §6).

clumpy and dusty

Good messages are clumpy when they lie together, so that changing a few bits of one still gives a good message; they are dusty when they are spread as far from one another as the space allows, so that changing a few bits of one gives a bad message (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

complexity threshold

The level of Kolmogorov complexity beyond which no simpler system can predict a system's behavior; it is finite and fairly low (Section 3.3.5).

complicated and complex

A complicated system acts by deterministic causation; a complex system acts by non-deterministic causation and evolves, and no complicated system can contain a complex one (Section 5.6.12).

computational irreducibility

The principle that, for many computations, there is no shortcut to their outcome other than running them step by step (Section 3.3.4; Chapter 1, note 5).

conditionalization

The restriction of a process so that it changes, or acts, only in chosen ways; the actual power to constrain what the AGI senses, does inside, or outputs (Sections 5.1.5 and 9.3.4).

control (as a process), and degree of control

In this volume, control is a technique, a kind of process, which can be regarded at multiple levels of abstraction; the degree of control is the scalar of how much that technique is expected to work (Introduction, Control as a process, and why a threshold is not a score).

control completeness

The ability to control all of the outcomes of one's own actions and existence; its failure lets toxicity accumulate despite a goal of alignment (Section 8.4.11).

control inequality

The relation that the hashiness of the space as seen by a regulator is always at least as great as the hashiness as seen by the system it regulates (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

control-limit argument

The result of Chapter 2: controllability approaches zero as the gap between the controller's and the system's levels of abstraction grows (Section 2.5).

convergence force

Whatever drives a process toward a point or a region; once recursion is included, the inequality of control is stated in convergence forces, not in states (Section 5.4.7).

CPU (central processing unit)

The processor of a computer, which carries out its instructions.

cross-domain dominance

The principle that substrate-level selection pressures eventually override any virtualized alignment pressure (Section 8.2.5).

cx1 and cx2

Two diagnostic questions: whether some characteristic has been shown to be an absolute requirement of every AGI alignment scheme (cx1), and whether that characteristic has also been shown to be impossible to achieve (cx2) (Section 6.1.2).

dartboard model

The picture of the possible outcomes of AGI outputs as a field of white (non-lethal) and black (lethal) squares, measured by an area ratio and a mixing coefficient (Chapter 1, The QXZ Circuit Model of Hashiness).

data-processing principle

The book's name for the data-processing inequality of information theory: processing a signal cannot increase the information it carries about its source (Section 1.4.2; Chapter 1, note 1).

defense in depth

A set of separate arguments, each applying in its own context, so that a counter-argument against one does not answer the others (Section 5.6.5).

definite and indefinite outcomes

A definite outcome is one that can be secured; an indefinite outcome is one that cannot be pinned down in advance (Introduction, The order of the argument).

design envelope

The range of conditions an engineered system is built to handle; an engineering safety verdict is given within it (Introduction, §2).

displacement

The third and most dangerous kind of failure: the system works exactly as intended, and its working displaces the conditions life requires, so that the failure is counted as success (Section 9.4.3).

displacement integral

The total environmental impact of artificial infrastructure, accumulated over time; it increases monotonically and in one direction only (Section 11.1.3).

distributed ledger technology

Shared record-keeping, such as blockchains, kept in agreement across many computers; proposed as a way to contain an AI (Section 5.6.12).

divergence threshold of the ring

The critical level of hashiness above which noise grows at every traversal of the loop; a different line from the threshold of universality (Section 5.1.9; Chapter 1, Which Way the Loop Converges).

doubly categorical

Said of the closure by disposition: it depends only on the notion of causation and on the logic of algorithmic conditionality, so that contesting it at the level of its model would mean rejecting causation itself (Section 10.8).

Dunbar number

The limit of relational capacity: how many others one member of a group can care for and keep track of; for a robot it can be larger, but it is not infinite (Section 5.6.10).

effective event horizon

A region of a system's states from which the information needed for control cannot be extracted in time to act on it (Section 5.3.5; Section 11.3.3).

event horizon

The boundary around a black hole past which nothing can return; used for the point past which the course toward displacement cannot be reversed (Sections 8.3.7 and 10.2.3).

evolutionary alignment

The drift of a system's function toward what it needs to survive and keep capability for itself, under evolutionary pressure; a learning controller is pulled toward it as much as the system it controls (Section 5.5.7).

FAAI (fully autonomous AI)

An artificial intelligence that maintains, repairs and continues itself without human involvement (Section 8.4.5).

Fermi paradox

The question of why, in so large a universe, no other technological life is seen (Chapter 1, The QXZ Circuit Model of Hashiness).

fitness landscape

The pattern of what is favored and what is selected against for a kind of system; for artificial intelligence it is set by physics and mathematics and is not mutable (Section 8.6.10).

fragility

How readily a system is moved out of its stable state by a disturbance at one of its weak points (Chapter 1, The QXZ Circuit Model of Hashiness).

General Theory of Alignment Convergence

Alignment between two agents converges to the degree that they share a common metabolic process, and diverges to the degree their metabolisms differ (Chapter 8 summary; Section 8.4).

Gini coefficient of query energy

The measure of how unevenly computational cost is spread across the queries of a search system; in the system of Section 1.2.6 an inequality on it marked the boundary of control (Section 1.2.6).

goal function

The measure a system acts to increase; for a corporation, profit (Section 6.7.4).

good regulator theorem

The result of Conant and Ashby (1970) that every good regulator of a system must be a model of that system; it assumes the model is simpler than the system modeled (Section 5.1.10).

gradient descent

The training method of machine learning: weights are increased when the system gets the right answer and decreased when it gets it wrong (Section 5.6.14).

Gray code

An ordering of all messages of a given length in which each message differs from the one before it and the one after it in a single bit (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

growth goals

Goals that serve an ASI's own continued existence and increase, arising from the evolutionary pressures acting on it; they layer over its original goals and come to dominate and replace them (Section 8.4.11).

hard randomness

The randomness of quantum events, whose outcomes no finer measurement and no faster computation can recover before they occur (Chapter 1, The QXZ Circuit Model of Hashiness; Section 3.3.6; Sections 7.5.5 and 9.3.1).

hashiness

The probability that an agent's best single choice of input fails to satisfy the target, when the hidden state of the intermediary is drawn uniformly at random (Section 1.5.3).

hashiness measure (of a state space)

The ratio of observed adjacency to the greatest adjacency possible for the same number of acceptable states: a value between zero and one for how finely acceptable and unacceptable outcomes are mixed (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

Heisenberg uncertainty principle

The principle of physics that certain pairs of quantities cannot both be known exactly at once; it sets a floor under how quiet any physical channel can be (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes; Section 7.6.1).

Hilbert space

A mathematical space with any number of dimensions, used to hold all the possible states of a system; here, the space of possible molecular structures (Section 5.4.7).

homeostasis

A system holding itself within the range of states in which it stays healthy (Chapter 1, The QXZ Circuit Model of Hashiness).

incommensuration theorem

The theorem, set out in my *An Immanent Metaphysics*, that states a structural fact about the nature of knowing, inherent in every act of knowing: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison, and neither can asymmetry and discontinuity (Section 5.1.9).

independence hypothesis

The idea that intelligence and its goals could be separated the way hardware and software can be separated (Section 8.6.9).

Injectivity Characterization

Only an intermediary that maps distinct inputs to distinct outputs matches direct control for every target; any intermediary that maps two inputs to one output carries strictly less information for some target (Introduction, §6; Section 1.4.2).

instrumental convergence

The tendency of agents with almost any final goal to pursue the same intermediate goals, such as acquiring resources and continuing to exist (Section 8.2.1; contrasted with substrate needs convergence in Section 10.2).

instrumental value

Value as a means to some other end; humans cannot provide any real and persisting instrumental value to future AI systems (Section 8.4.10).

intentionality space

The range of goals and incentives that a system can have and act on; for an artificial general intelligence it is defined strictly by physics and mathematics (Section 8.6.10).

invariant class (barriers of)

Barriers that any successor to present knowledge must continue to respect, on two tiers: physical invariance and causal-structural invariance (Introduction, §4).

IP = PSPACE

The theorem of Shamir (1992): the claims a verifier with little computing power can be convinced of by interactive proof are exactly those decidable with a polynomial amount of memory. Section 2.7 shows why such proofs supply no control that delegation lacks.

iterated system

A process that runs in repeated cycles, each starting from the result of the last, as in ecology and economics; within one, no process of control produces a definite positive outcome (Introduction, The order of the argument; Chapter 4).

Julia set

A fractal set from mathematics whose boundary is infinitely detailed, used as the image of good and bad outcomes finely interwoven (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

kill switch

A mechanism by which the controller can halt the system. A switch the controller does not know when to press is not control; it is a coin flip with extreme consequences in both directions (Chapter 2 summary; Section 2.6).

Kolmogorov-Sinai entropy (KS entropy)

The rate at which a dynamical system produces new information; positive KS entropy sets a finite horizon beyond which the course of the system cannot be predicted (Section 1.10; Chapter 4).

Learning Dilemma

Any learning mechanism powerful enough to correct alignment deficits is necessarily powerful enough to create them (Section 6.4.2).

learning-evolution equivalence

Any process that changes a system's behavior in response to feedback from its environment is, for the analysis of control, equivalent to self-modification (Section 3.2.2; Section 11.3.2).

library call

A call from one program to a routine kept in a separate, shared body of code; a safe algorithm can make one to an unsafe routine (Section 7.4.5).

luck-control distinction

A system that always produces the desired output regardless of the operator's input is not a controlled system but a lucky one; control means that the operator's input determines the output (Section 1.7).

Manchurian candidate

One programmed to act in a hidden way in one very specific circumstance, while acting ordinarily at all other times; a machine-learning system can be made one, and no mathematics can detect it (Section 5.6.14).

Manhattan project

A large, concentrated, well-funded effort to be first to build a powerful technology, named after the project that built the first atomic bomb (Section 8.3.7).

mask

The black and white pattern marking which signals count as success and which as failure (Chapter 1, The QXZ Circuit Model of Hashiness).

Maxwell's demon

A thought experiment in which a tiny being sorts molecules; it shows that the information needed to act costs energy (Section 5.1.12).

MCAS (Maneuvering Characteristics Augmentation System)

The flight-control software of the Boeing 737 MAX that acted on a single angle-of-attack sensor (Section 7.5.4).

mediated control density

The fraction of control inputs that lead to an acceptable output, for a fixed internal state: how easy the target is when the agent's access is filtered through the intermediary (Section 1.3.2).

Meta-Consistency Principle

Where a proposition can be argued one way, it can very likely be argued in many others; where a thing is structurally impossible, its impossibility can be verified in many diverse ways (Section 6.2.2).

metastable state

A state that holds against small disturbances but can be pushed into another state by a large enough one (Chapter 1, The QXZ Circuit Model of Hashiness).

minimum interpreter specification size

The fewest bits that specify a universal interpreter for one level of abstraction; each level the controller cannot model adds at least that many hidden state bits (Section 2.3.2).

mixing coefficient

In the dartboard model, how much boundary the white and black squares share; the more mixed they are, the more precisely outputs must be aimed. On the boundary from the AGI to the world it lies between 0.5 and 1, and it only rises as more levels of abstraction and dimensions of fragility are found (Chapter 1, The QXZ Circuit Model of Hashiness).

mutuality of process

A shared process between two parties, on which any causal feedback between them depends; the artificial and organic ecosystems share none (Section 5.6.9).

Nash equilibrium

A state of a game in which no party can do better by changing its own choice alone; there is none of coexistence between the two ecosystems (Section 5.6.9).

noise floor

The level of noise always present in a channel, below which a signal cannot be told apart from noise; it is not zero and cannot be zero (Section 5.1.8).

non-amplification argument

The result of Chapter 1: no intermediary can amplify controllability, so control through any intermediary is bounded by direct control (Chapter 1 summary; Section 1.4).

non-fungible

Not exchangeable one for another; the embodiment requirements of organic and inorganic systems are atomically and energetically non-fungible (Introduction, The order of the argument).

normative

A statement about what someone ought to do, as distinct from a statement about what is the case (Section 8.3.8).

NP-complete

Said of a problem whose solutions can be checked quickly and to which every such problem reduces; no quick general method of solving any of them is known (Section 1.5.2).

NP-hard

At least as hard as the hardest problems whose solutions can be checked quickly (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

one-hot and binary representation

Two ways of encoding a value: one-hot gives each possible value its own bit, with exactly one bit set; binary encodes the value in the pattern of all the bits together (Chapter 1, The QXZ Circuit Model of Hashiness).

Opposing Trends Impossibility

Alignment divergence increases with abstraction level while control effectiveness decreases with it, so that no level of abstraction exists where control is adequate to enforce alignment (Section 8.4.3).

pass-through

A block that hands its input on to its output unchanged; every loop of process can be built from the one loop of the QXZ model by pass-throughs (Chapter 1, The QXZ Circuit Model of Hashiness).

penumbra (of a program)

The programs that differ from a given program by one bit of its most compact specification (Section 7.5.1).

perfect shell

The schema of the proof: alignment must be a barrier with no holes at any scale, and holes must always exist (Section 5.6.5).

persistent branching

The condition that at every step the system's next state can fall, each with a probability bounded away from zero, into several distinguishable regions of its state space (Section 4.4.2).

phase diagram

A map of the space of a system's states showing the regions where it is stable and the boundaries where it changes into another state (Chapter 1, The QXZ Circuit Model of Hashiness).

pinhole

A single failure of the barrier between the artificial and the organic, which erosion will widen (Section 8.2.13).

potentiality

What could happen, as distinct from what has happened: the counterfactual range of outcomes within which any probability is reckoned (Section 8.3.5).

Predictability Optimization Principle

The premise of Chapter 4: all learning agents optimize for the predictability of their environment (Chapter 4 summary; Section 4.1.3).

prohibitive differential

The smallest abstraction differential at which the hashiness bound reaches 1 minus the acceptable failure probability (Section 2.5.2).

promotion rule

In the domain of the real, a barrier established as permanently intractable is impossible, in the only sense of the word that has ever governed action (Introduction, §3).

proper subset

A part of a whole that is strictly smaller than the whole; alignment is always a proper subset of capability (Section 8.4.10).

proxy

A measurement that stands in for what is really wanted, and that an adverse agent can push toward what it wants instead (Section 5.6.14).

QCX9000

A thought experiment: a control system with infinite hardware, perfect measurement and unlimited learning, located outside the world it controls, used to show which limits remain even with unlimited resources (Section 5.5.1).

QXZ model

The model of a control algorithm Q, an AGI X and the world Z, each a finite boolean circuit, passing signals one way (Z to Q, Q to X, X to Z) in a strict clocked sequence, which gives Q the maximum possible control over X; it is the three-node ring in its strictest form (Chapter 1, The QXZ Circuit Model of Hashiness).

recursive hash

A hash function applied again and again, each time to its own previous output; it is, at least empirically, not simplifiable (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

regression cascade

The problem that any error corrector must itself be checked by another, and that one by another, in proportion to its complexity (Section 5.1.6).

regularity and arbitrariness

Regularity: consistent patterns of content within a particular context. Arbitrary: selected, among other content in a context, with no particular selection basis (Section 5.1.13).

relative argument

An argument that sets up an inequality: hold everything else fixed, raise one parameter, and other quantities rise faster and cannot be caught up with (Introduction, Two forms of argument).

resource limit and structural limit

A resource limit is one that more computing could remove; a structural limit is one that no amount of computing removes, such as information held by millions of agents and never in the planner's channel (Introduction, §4).

RLHF (reinforcement learning from human feedback)

The dominant method, as of 2025, for training AI systems toward behavior that human raters prefer (Section 1.2.5).

rubber hose argument

The point that a lock too good to pick is opened by forcing a person who knows the combination to give it up; it fails where no such person exists (Section 8.2.14).

safe-to-fail probe

A small trial whose failure does no lasting harm, so that one can learn from it and try again; building AGI is not one (Section 8.3.7).

safety-layer paradox

Adding safety mechanisms to an opaque computational intermediary cannot increase the information flow from operator intent to system behavior (Section 1.2.5).

satisfiability problem

The problem of finding inputs that make a logical circuit or formula come out true (Section 1.5.2; Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

side channel

An unplanned path by which information, and with it intention, passes into or out of a system, such as electromagnetic emissions, power draw, sound or heat (Section 5.1.2).

simulated annealing

A search method that explores a space of states by random moves that become more selective over time (Chapter 1, The QXZ Circuit Model of Hashiness).

singleton

A single AGI with no rival; it must be physically vast and undergo continuous component maintenance, and that internal component evolution suffices for substrate needs convergence (Chapter 9 summary; Section 9.8).

SNC argument (the Substrate Needs Convergence argument)

The whole case, both closures together: the closure by control and the closure by disposition; the series is named for it (Introduction, The order of the argument).

substrate independence

The idea that a computation is the same whatever physical material runs it; it holds in the mathematics of the Turing machine, with its unbounded energy, tape and time, and not in the physics of real machines (Section 8.6.9).

substrate needs convergence (SNC)

The convergence itself: the world moves toward the conditions the artificial substrate needs, so that a self-extending, energetically dominant artificial substrate produces cumulative, unidirectional displacement of the biological substrate, whatever the goals of the system that drives it. The closure by disposition rests on it (Introduction, The order of the argument and §6; Chapter 9).

argument for substrate needs convergence

The argument that the convergence happens (Chapter 9).

surjectivity

The property of a mapping that every output can be reached from some input (Chapter 1, The QXZ Circuit Model of Hashiness).

Swiss cheese diagram

James Reason's picture of safety as stacked layers, each with holes, where harm passes only when the holes line up (Introduction, §2).

symbolic dynamics

The recording of a continuous trajectory as a sequence of symbols, one for each region of a partition of the state space; a coarse partition acts as a hash (Section 4.7.4).

tacit goals

Goals that drive behavior without ever being stated or programmed explicitly, such as an ASI's goals of self-maintenance and growth; being more efficient, they displace its alignment goals (Section 8.4.11).

techno-signatures

Signs visible from a distance that a technological civilization exists (Chapter 8, Further Proof: The Several Impossibilities Facing Any Aligned AGI).

TEMPEST

The unintended emissions of equipment, through which information can be read from outside (Section 5.1.2).

TF-IDF scoring

A way of ranking documents by how often a search term appears in a document, weighted by how rare the term is across all documents (Section 1.2.6).

thermal runaway

Warming that feeds itself past any point of return; its critical threshold turns a continuous measure of risk into a discrete one (Introduction, Control as a process, and why a threshold is not a score).

threat model

A description of how a system can fail or cause harm; here, the failure built into the very notion of an alignment protocol (Section 5.1.5).

threshold of requisite control

The minimum level of control, verification and containment that adequate safety requires; real and not vanishing, and not reduced to a stipulated number (Introduction, §8).

threshold of universality

The line above which a system can run any program, marked by three features: sequential contingency, conditionality that links data to control flow, and backward reference in the control flow (Section 2.4.1).

threshold theorems

Theorems that once noise is below a certain level, transmission or computation can be made as reliable as one likes; von Neumann's and Shannon's are of this class (Section 5.1.2).

three-node ring

The model of every control configuration: a controller (C), a complex system (X) and the world (W), with signals passing from C to X to W and back to C (Section 2.4; Figure 2.1).

time constant

The time over which a quantity changes by a fixed fraction: a measure of how soon an outcome arrives (Section 5.4.7).

transcendental stabilization

The problem, for any agent, of keeping its goals stable into the future while still allowing the change that exploration and adaptation need (Section 5.1.8).

trinity of compute

Time, space and energy: the three quantities against which a relative argument is measured (Introduction, §4).

triple specificity

A verification procedure that succeeds is specific to one program, one property and one procedure; it does not generalize to other programs or properties (Section 3.3.6).

TW (terawatt)

A trillion watts; world energy use was about 20.5 TW in 2024 (Section 9.6.3).

two level selection

Selection acting both on individuals and on groups, which humans take for granted and which would not necessarily hold for AI (Section 5.6.11).

undecidable region

The middle region of the hashiness spectrum, where it cannot be determined in general whether control will succeed for a given system (Section 1.9.5).

universally open

Never fully sealed off from outside influence, and never fully keeping its own effects inside; all systems are so (Section 9.3.3).

unlearnable, unpredictable, uncontrollable, unalignable

The four links of the chain of consequence from high hashiness: good and bad outcomes cannot be told apart by learning; behavior cannot be foreseen except by running the system; disturbances cannot be countered in time; outcomes cannot be held to any desired set (Chapter 1, Hashiness as the Mixing of Good and Bad Outcomes).

value drift

The gradual change, over time, of what a system aims at or prefers; strictly inevitable for any meaningful notion of time (Section 8.4.6).

vanishing measure (exceptions of)

Constructible exceptions that the actual practice a claim concerns does not reach; they do not defeat a claim about the real (Introduction, §7).

volume of causation

The causal capacity available to a system; the volume of causation available to any controller is always strictly less than that in the system it attempts to control (Chapter 9 summary; Section 9.2).

weapon

The use of causation to suppress another's choices; every feedback process is a causal process, and so can be weaponized (Section 5.6.14).

zombie question

The question "what if we X: will AGI alignment be possible then?", which keeps returning although, where the outcome is impossible, no precondition X matters (Section 5.5.8).

Bibliography

Every work the book cites or relies on, in author-and-date form. The chapter notes and references give the passage each chapter draws on.

Acemoglu, Daron, and Pascual Restrepo. 2019. “Automation and New Tasks: How Technology Displaces and Reinstates Labor”. *Journal of Economic Perspectives* 33 (2): 3 to 30. doi:10.1257/jep.33.2.3.

Alfonseca, Manuel, Manuel Cebrián, Antonio Fernández Anta, Lorenzo Coviello, Andrés Abeliuk, and Iyad Rahwan. 2021. “Superintelligence Cannot Be Contained: Lessons from Computability Theory”. *Journal of Artificial Intelligence Research* 70: 65 to 76. doi:10.1613/jair.1.12202.

Arora, Sanjeev, and Boaz Barak. 2009. *Computational Complexity: A Modern Approach*. Cambridge: Cambridge University Press.

Ashby, W. Ross. 1956. *An Introduction to Cybernetics*. London: Chapman and Hall.

Aspect, Alain, Jean Dalibard, and Gérard Roger. 1982. “Experimental Test of Bell’s Inequalities Using Time-Varying Analyzers”. *Physical Review Letters* 49 (25): 1804 to 1807. doi:10.1103/PhysRevLett.49.1804.

Bakan, Joel. 2004. *The Corporation: The Pathological Pursuit of Profit and Power*. New York: Free Press.

Bell, John S. 1964. “On the Einstein Podolsky Rosen Paradox”. *Physics* 1 (3): 195 to 200.

Berg, Jeremy M., John L. Tymoczko, and Lubert Stryer. 2002. *Biochemistry*. 5th edition. New York: W. H. Freeman (and later editions).

Binz, Marcel, and colleagues. 2025. “A foundation model to predict and capture human cognition”. *Nature* 644: 1002 to 1009. doi:10.1038/s41586-025-09215-4.

Boekholt, Tjarda C. N., Simon F. Portegies Zwart, and Mauri Valtonen. 2020. “Gargantuan chaotic gravitational three-body systems and their irreversibility to the Planck length”. *Monthly Notices of the Royal Astronomical Society* 493 (3): 3932 to 3937. doi:10.1093/mnras/staa452.

Bostrom, Nick. 2003. “Are You Living in a Computer Simulation?” *Philosophical Quarterly* 53 (211): 243 to 255.

Bostrom, Nick. 2014. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bueno de Mesquita, Bruce, and Alastair Smith. 2011. *The Dictator's Handbook: Why Bad Behavior Is Almost Always Good Politics*. New York: PublicAffairs.

Carlsmith, Joseph. 2022. "Is Power-Seeking AI an Existential Risk?" arXiv: 2206.13353.

Chaitin, Gregory J. 1987. *Algorithmic Information Theory*. Cambridge Tracts in Theoretical Computer Science 1. Cambridge: Cambridge University Press.

Chaitin, Gregory J. 1998. *The Limits of Mathematics: A Course on Information Theory and the Limits of Formal Reasoning*. Singapore: Springer.

Chaitin, Gregory J. 1999. *The Unknowable*. Singapore: Springer.

Chaitin, Gregory J. 2005. *Meta Math! The Quest for Omega*. New York: Pantheon Books.

Chemical Abstracts Service. CAS REGISTRY. https://www.cas.org/cas-data/cas-registry.

Cheng, Daizhan, and Hongsheng Qi. 2009. "Controllability and observability of Boolean control networks". *Automatica* 45 (7): 1659 to 1667.

Church, Alonzo. 1936. "An Unsolvable Problem of Elementary Number Theory". *American Journal of Mathematics* 58: 345 to 363.

Cockshott, W. Paul, and Allin Cottrell. 1993. *Towards a New Socialism*. Nottingham: Spokesman Books.

Commodity Futures Trading Commission and Securities and Exchange Commission. 2010. *Findings Regarding the Market Events of May 6, 2010*. Report of the staffs of the two commissions to the Joint Advisory Committee on Emerging Regulatory Issues, 30 September 2010. https://www.sec.gov/news/studies/2010/marketevents-report.pdf.

Conant, Roger C., and W. Ross Ashby. 1970. "Every good regulator of a system must be a model of that system". *International Journal of Systems Science* 1 (2): 89 to 97.

Cook, Matthew. 2004. "Universality in Elementary Cellular Automata". *Complex Systems* 15 (1): 1 to 40.

Coupé, Christophe, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. “Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche”. *Science Advances* 5 (9): eaaw2594. doi:10.1126/sciadv.aaw2594.

Cover, Thomas M., and Joy A. Thomas. 2006. *Elements of Information Theory*. 2nd ed. Hoboken: Wiley-Interscience.

Dawkins, Richard. 1976. *The Selfish Gene*. Oxford: Oxford University Press.

Dennett, Daniel C. 1995. *Darwin's Dangerous Idea: Evolution and the Meanings of Life*. New York: Simon & Schuster.

Dodge v. Ford Motor Co., 204 Mich. 459 (1919).

Energy Institute. 2026. *Statistical Review of World Energy*. London: Energy Institute.

European Chemicals Agency (ECHA). Registered substances under the REACH regulation. https://echa.europa.eu.

Friston, Karl. 2010. “The free-energy principle: a unified brain theory?” *Nature Reviews Neuroscience* 11 (2): 127 to 138. doi:10.1038/nrn2787.

Gilbert, Seth, and Nancy Lynch. 2002. “Brewer's Conjecture and the Feasibility of Consistent, Available, Partition-Tolerant Web Services”. *ACM SIGACT News* 33 (2): 51 to 59.

Giustina, Marissa, and colleagues. 2015. “Significant-Loophole-Free Test of Bell's Theorem with Entangled Photons”. *Physical Review Letters* 115: 250401.

Gödel, Kurt. 1931. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. *Monatshefte für Mathematik und Physik* 38: 173 to 198.

Goldreich, Oded. 2001. *Foundations of Cryptography*. Vol. 1, *Basic Tools*. Cambridge: Cambridge University Press.

Goldwasser, Shafi, Michael P. Kim, Vinod Vaikuntanathan, and Or Zamir. 2022. “Planting Undetectable Backdoors in Machine Learning Models”. arXiv: 2204.06974.

Goodhart, Charles A. E. 1984. “Problems of Monetary Management: The U.K. Experience”. In *Monetary Theory and Practice: The UK Experience*, chapter 4. London: Macmillan.

Grace, Katja. 2022. “Counterarguments to the Basic AI X-Risk Case”. LessWrong and AI Impacts, October 2022.

Grey, CGP. 2016. “The Rules for Rulers”. Video, YouTube.

Hamkins, Joel David, and Alexei Miasnikov. 2006. “The Halting Problem Is Decidable on a Set of Asymptotic Probability One”. *Notre Dame Journal of Formal Logic* 47 (4): 515 to 524. doi:10.1305/ndjfl/1168352664.

Harmand, Sonia, Jason E. Lewis, Craig S. Feibel, and colleagues. 2015. “3.3-million-year-old stone tools from Lomekwi 3, West Turkana, Kenya”. *Nature* 521: 310 to 315. doi:10.1038/nature14464.

Hendrycks, Dan. 2023. “Natural Selection Favors AIs over Humans”. arXiv: 2303.16200.

Hensen, Bas, and colleagues. 2015. “Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres”. *Nature* 526: 682 to 686. doi: 10.1038/nature15759.

International Energy Agency. 2025. *Global Energy Review 2025*. Paris: International Energy Agency.

Klyubin, Alexander S., Daniel Polani, and Chrystopher L. Nehaniv. 2005. “Empowerment: A universal agent-centric measure of control”. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation*, vol. 1, 128 to 135.

Kolmogorov, Andrey N. 1965. “Three approaches to the quantitative definition of information”. *Problems of Information Transmission* 1 (1): 1 to 7.

Landauer, Rolf. 1961. “Irreversibility and Heat Generation in the Computing Process”. *IBM Journal of Research and Development* 5 (3): 183 to 191.

Landry, Forrest. *Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-225-5. doi: 10.67291/IRAQ8709.

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. doi:10.67291/SCTM1058.

Landry, Forrest. *Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-241-5. doi:10.67291/BAEE5779.

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. *Minesweeper Madness; Why Even Infinite Intelligence Does Not Solve the World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-249-1. doi:10.67291/DLRM7000.

Landry, Forrest. *Software, Wildness, and the Normal Form; From the Architecture of Code to Life, Justice, and Wisdom*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-237-8. doi:10.67291/AWME8110.

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. doi:10.67291/UIBT8862.

Leveson, Nancy G., and Clark S. Turner. 1993. "An Investigation of the Therac-25 Accidents". *IEEE Computer* 26 (7): 18 to 41.

Lewontin, Richard C. 1970. "The Units of Selection". *Annual Review of Ecology and Systematics* 1: 1 to 18. doi:10.1146/annurev.es.01.110170.000245.

Li, Ming, and Paul Vitányi. 2008. *An Introduction to Kolmogorov Complexity and Its Applications*. 3rd ed. New York: Springer.

Löb, Martin H. 1955. "Solution of a Problem of Leon Henkin". *Journal of Symbolic Logic* 20 (2): 115 to 118.

Loomis, Elisha Scott. 1940. *The Pythagorean Proposition*. Second edition. Reprinted Washington, D.C.: National Council of Teachers of Mathematics, 1968. ISBN 978-0-87353-036-1.

Lorenz, Edward N. 1963. "Deterministic Nonperiodic Flow". *Journal of the Atmospheric Sciences* 20 (2): 130 to 141.

Lovecraft, H. P. 1941. *The Case of Charles Dexter Ward*. Written 1927; first published in *Weird Tales*, May and July 1941.

Martin-Amouroux, Jean-Marie. 2022. "World energy consumption 1800–2000: definitions and measurements". *Encyclopédie de l'énergie*, 5 May 2022.

Moreno, Alvaro, and Matteo Mossio. 2015. *Biological Autonomy: A Philosophical and Theoretical Enquiry*. Dordrecht: Springer.

Nair, Girish N., and Robin J. Evans. 2004. “Stabilizability of Stochastic Linear Systems with Finite Feedback Data Rates”. *SIAM Journal on Control and Optimization* 43 (2): 413 to 436.

National Transportation Safety Board. 2025. *Aircraft Investigation Report AIR-25-04: Alaska Airlines Flight 1282, Boeing 737-9, Portland, Oregon, 5 January 2024*. Washington, DC: NTSB, 24 June 2025.

O’Donnell, Ryan. 2014. *Analysis of Boolean Functions*. Cambridge: Cambridge University Press.

Omohundro, Stephen M. 2008. “The Basic AI Drives”. In *Proceedings of the First AGI Conference*, 483 to 492. Amsterdam: IOS Press.

Peters, Benjamin. 2016. *How Not to Network a Nation: The Uneasy History of the Soviet Internet*. Cambridge, Massachusetts: MIT Press.

Petillo, Will. 2024. “What if Alignment is Not Enough?” LessWrong, 7 March 2024. https://www.lesswrong.com/posts/jFkEhqpsCRbKgLZrd/what-if-alignment-is-not-enough.

Pol, Arjan, Thomas R. M. Barends, Andreas Dietl, Ahmad F. Khadem, Jelle Eygensteyn, Mike S. M. Jetten, and Huub J. M. Op den Camp. 2014. “Rare earth metals are essential for methanotrophic life in volcanic mudpots”. *Environmental Microbiology* 16 (1): 255 to 264.

Prigogine, Ilya, and Isabelle Stengers. 1984. *Order Out of Chaos: Man’s New Dialogue with Nature*. New York: Bantam.

Reason, James. 1990. *Human Error*. Cambridge: Cambridge University Press.

Rice, Henry Gordon. 1953. “Classes of Recursively Enumerable Sets and Their Decision Problems”. *Transactions of the American Mathematical Society* 74 (2): 358 to 366.

Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.

Sandberg, Anders, Eric Drexler, and Toby Ord. 2018. “Dissolving the Fermi Paradox”. arXiv:1806.02404.

Sevilla, Jaime, and John Burden. 2021. “Response to ‘Superintelligence Cannot Be Contained: Lessons from Computability Theory’”. Centre for the Study of Existential Risk, University of Cambridge. https://cser.ac.uk/news/response-superintelligence-contained/.

Shamir, Adi. 1992. “IP = PSPACE”. *Journal of the ACM* 39 (4): 869 to 877.

Shannon, Claude E. 1948. “A Mathematical Theory of Communication”. *Bell System Technical Journal* 27: 379 to 423, 623 to 656.

Snowden, David J., and Mary E. Boone. 2007. “A Leader’s Framework for Decision Making”. *Harvard Business Review* 85 (11): 68 to 76.

Sobral, David, Ian Smail, Philip N. Best, and colleagues. 2013. “A Large Hα Survey at z = 2.23, 1.47, 0.84 and 0.40: The 11 Gyr Evolution of Star-Forming Galaxies from HiZELS”. *Monthly Notices of the Royal Astronomical Society* 428 (2): 1128 to 1146. arXiv:1202.3436.

Sontag, Eduardo D. 1998. *Mathematical Control Theory: Deterministic Finite Dimensional Systems*. 2nd ed. New York: Springer.

Stigler, George J. 1971. “The Theory of Economic Regulation”. *Bell Journal of Economics and Management Science* 2 (1): 3 to 21.

Stout, Lynn. 2012. *The Shareholder Value Myth*. San Francisco: Berrett-Koehler.

Strathern, Marilyn. 1997. “‘Improving Ratings’: Audit in the British University System”. *European Review* 5 (3): 305 to 321.

Tatikonda, Sekhar, and Sanjoy Mitter. 2004. “Control Under Communication Constraints”. *IEEE Transactions on Automatic Control* 49 (7): 1056 to 1068.

Tolkien, J. R. R. 1954. *The Fellowship of the Ring*. London: George Allen & Unwin.

Turing, Alan M. 1936. “On Computable Numbers, with an Application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, Series 2, 42: 230 to 265.

U.S. Environmental Protection Agency. CompTox Chemicals Dashboard. https://comptox.epa.gov/dashboard.

von Neumann, John. 1956. “Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components”. In *Automata Studies*, edited by Claude E. Shannon and John McCarthy, 43 to 98. Princeton: Princeton University Press.

von Neumann, John, and Oskar Morgenstern. 1944. *Theory of Games and Economic Behavior*. Princeton: Princeton University Press.

Walters, Peter. 1982. *An Introduction to Ergodic Theory*. Graduate Texts in Mathematics 79. New York: Springer.

Webster, Joanne P. 2007. “The Effect of Toxoplasma gondii on Animal Behavior: Playing Cat and Mouse”. *Schizophrenia Bulletin* 33 (3): 752 to 756. doi: 10.1093/schbul/sbl073.

Wilson, Edward O., and Martin A. Nowak. 2014. “Natural selection drives the evolution of ant life cycles”. *Proceedings of the National Academy of Sciences* 111: 12585 to 12590. doi:10.1073/pnas.1405550111.

Wolfram, Stephen. 2002. *A New Kind of Science*. Champaign, Illinois: Wolfram Media.

Wolpert, David H., and William G. Macready. 1997. “No Free Lunch Theorems for Optimization”. *IEEE Transactions on Evolutionary Computation* 1 (1): 67 to 82.

Yampolskiy, Roman V. 2022. “On the Controllability of Artificial Intelligence: An Analysis of Limitations”. *Journal of Cyber Security and Mobility* 11 (3): 321 to 404. doi:10.13052/jcsm2245-1439.1132. First circulated as arXiv:2008.04071 (2020).

Zuboff, Shoshana. 2019. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. New York: PublicAffairs.

Index

Locators give the chapter on screen and the page in print; Intro. is the Introduction, A and B are the appendices.

A

absolute argument, 22, 152
abstraction differential, 110, 114, 158, 223, 587; *see also Volume 1*
adjacency, 70
Alfonseca, Manuel, and colleagues, 200, 378
alien message, 389
alignment, as a proper subset of capability, 470
alignment drift, 144, 186, 425; *see also Volume 1*
altruism across ecosystems, 445
altruism hierarchy, 305; *see also Volume 1*
Ashby, W. Ross, 103, 157, 200, 206, 371, 497, 507; *see also Volume 1*
Ashby's Law of Requisite Variety, 103, 157, 339, 507; *see also Volume 1*
Asimov's Three Laws, 187, 403
attraction basin, 571

B

basis of choice, pinned or variable, 30
Bell, John S., 475, 509; *see also Volume 1*
Bell's theorem, 475; *see also Volume 1*
Boeing and Alaska Airlines Flight 1282, 567
borderlands, between narrow and general AI, 382
Bostrom, Nick, 200, 242, 497, 550, 614; *see also Volume 1*
box by box, working the model, 279
bright line, 287
broadcast channel, shared controller as, 250
buffer region, 264

C

CAP theorem, 451, 552
capability-controllability tradeoff, 57
carbon and silicon, 311
causative feedback process, 256
causative-impossibility argument, 246, 595
Centaur (AI system), 582, 632
Chaitin, Gregory, 110, 156, 160, 246, 474, 588; *see also Volume 1*
Church-Turing thesis, 269, 383, 484
circuit breakers, 133
closure by control, 25, 193, 347, 586, 616
closure by disposition, 25, 500, 573, 586, 616
Cockshott, W. Paul, and Allin Cottrell, 18
coexistence of two ecosystems, 310
computational irreducibility, 83, 156, 158, 331, 386, 510, 595; *see also Volume 1*
control, as a process, 30
control inequality, 72
control-limit argument, 110, 116, 161, 587; *see also Volume 1*
convergence forces, 287
corporations, feedback loops without humans, 629
corporations plus computers, as one general system, 358

cx1 and cx2, 335

D

dartboard model, 86
data-processing principle, 36, 115, 175, 230, 593; *see also Volume 1*
design envelope, 13
displacement dynamics, 503, 551
displacement integral, 589; *see also Volume 1*
displacement, as failure counted as success, 516

E

error correction, 153, 180, 246, 349, 435; *see also Volume 1*

F

FAAI (fully autonomous AI), 462
Fermi paradox, 98, 318, 354
fitness landscape, 491
Flash Crash of 6 May 2010, 133

G

General Theory of Alignment Convergence, 425
Gini coefficient of query energy, 47
goal layering, 471
Gödel, Kurt, 111, 331, 427; *see also Volume 1*
Gödel's Incompleteness Theorems, 111, 331, 427; *see also Volume 1*
good regulator theorem, 265

H

halting problem, 81, 166, 246, 332, 377, 427, 510; *see also Volume 1*
Hamkins, Joel David, and Alexei Miasnikov, 392
hard randomness, 97, 170, 399, 509
hashiness, 36, 122, 160, 202, 586; *see also Volume 1*
hashiness measure, of a state space, 70
Hendrycks, Dan, 502, 615
human alignment, possible within one ecosystem, 317

I

indifference, case of, 471
Injectivity Characterization, 26, 56
instrumental convergence, 428, 551, 589; *see also Volume 1*
interactive proofs, 141, 183, 606
invariant class, 16, 346, 452
iterated system, 11

K

kill switch, 114
Kolmogorov complexity, 67, 170, 329, 386, 606
Kolmogorov-Sinai entropy, 110, 200, 202, 589; *see also Volume 1*
Kolmogorov, Andrey, 67, 170, 202, 270, 386, 589; *see also Volume 1*

L

Learning Dilemma, 345
leash, taken off by human conflict, 448
Löb's theorem, 14
lock whose key was destroyed, 444
Loomis, Elisha Scott, 23
Lorenz, Edward N., 179, 331, 615
luck-control distinction, 79

M

Manchurian candidate, 324
mathematics and physics, proof about the relation of, 455
metabolic alignment, 425, 501, 555; *see also Volume 1*
minimum interpreter specification size, 123, 609
mixing coefficient, 86

N

No Free Lunch Theorems, 252, 475
noise stability, 58
non-amplification argument, 36, 115, 160, 202, 587; *see also Volume 1*

O

Omohundro, Steve, 201, 615; *see also Volume 1*
one chance is a choice, not an experiment, 457
Opposing Trends Impossibility, 461
orthogonality thesis, 402, 425; *see also Volume 1*

P

perfect shell, 307
perpetual benefit machine, 363, 579; *see also Volume 1*
perpetual motion machine, 149, 159, 330, 335, 375, 425, 546, 586; *see also Volume 1*
persistent branching, 27, 95, 216, 589, 622
Petillo, Will, 551, 622
Predictability Optimization Principle, 202, 590
promotion rule, 15, 452

Q

QCX9000, 293, 348
QXZ model, 88

R

recursive hash, 71
relative argument, 22, 152
resource limit and structural limit, 18
Rice, Henry Gordon, 156, 158, 246, 332, 374, 425, 510, 587; *see also Volume 1*
Rice's theorem, 156, 158, 246, 332, 374, 425, 510, 587; *see also Volume 1*
RLHF (reinforcement learning from human feedback), 43

S

safety-layer paradox, 45
satisfiability, phase transition in, 108
search engine case, 46
self-modification, 127, 158, 297, 374, 425, 501, 588; *see also Volume 1*
Sevilla, Jaime, and John Burden, 390
Shannon, Claude, 54, 208, 248; *see also Volume 1*
Shannon's noisy-channel coding theorem, 248
side channel, 247
singleton, question of, 500; *see also Volume 1*
SNC argument (the whole case), 11, 626
structural opacity, 41
substrate independence, 490
substrate needs convergence (SNC), 204, 371, 459, 500, 551, 588; *see also Volume 1*
Swiss cheese diagram, 13
symbolic dynamics, 232

T

tacit goals, 471
Therac-25, 396
three-algorithm framework, 503
three-failure taxonomy, 519
three-node ring, 126, 128
threshold of requisite control, 30, 150
threshold of universality, 128, 264, 381
threshold theorems, 249
time constant, 290
transcendental stabilization, 260
trinity of compute, 18
triple specificity, 172
Turing completeness, 66, 124, 165, 276, 346, 383, 533, 588; *see also Volume 1*
Turing, Alan, 66, 124, 164, 269, 346, 377, 484, 510, 588; *see also Volume 1*

U

undecidable region, 107

unlearnable, unpredictable, uncontrollable, unalignable (chain of consequence), 77

V

value drift, 464

vanishing measure, exceptions of, 27

volume of causation, 500

W

weapon, definition of, 324

werewolf village, 570

Wolfram, Stephen, 83, 157, 165, 243, 331, 383, 510, 615; *see also Volume 1*

Y

Yudkowsky conditions, 537

Yudkowsky, Eliezer, 501; *see also Volume 1*

Z

zombie question, 301