

The Substrate and the Ecosystem

Why Artificial General Intelligence Cannot Be Aligned, and What Remains for Persons

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

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

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-246-0

DOI: https://doi.org/10.67291/NYZX3639

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

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

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

Author ORCID: 0009-0005-6275-0362

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

Contents

Abstract iv
Preface v
Terms vii

Part I: The Argument Restated 1
1. The Substrate Needs Convergence 2
2. Why Impossibility, and Why It Is the Right Kind of Claim 46
3. The Impossibility of Alignment 109

Part II: The Wider Frame 169
4. Alignment and the Ecosystem 170
5. What Artificial Intelligence Cannot Buy 202
6. The Circuit of Nature, Persons and Machines 221

Part III: The Boundary 249
7. Desire and Alignment 250
8. Artificial Wisdom and the Limits of Machine Consciousness 275
9. Completeness of Knowledge of Operation 296

Summary 317

Glossary 319

Bibliography 337

Index 346

Abstract

The two-volume treatment of the substrate needs convergence argument (*Why AGI Cannot Be Made Safe* and *Why No Control Suffices*) establishes that alignment between artificial general intelligence and organic life is forbidden by the physics of substrate, as perpetual motion is forbidden by thermodynamics; it is not merely difficult as an engineering task. Its precedent (not diplomacy between agents) is the Great Oxidation Event, when one form of life remade the atmosphere for its own metabolism and most life that could not breathe the result died. This book shows why that alignment is impossible in terms a reader can follow without the formal volumes. It also shows what remains for persons once it is excluded: the care no machine can give, the wisdom no computation can reach, and the knowledge of their own tools without which no person is sovereign.

Preface

This is the eighth book of the argument that artificial general intelligence cannot be made safe. What this book gives that the other books of the argument do not is in Parts II and III: objectivity without measurement (fields of yes-or-no criteria in place of floating-point metrics), the three markets, of which care is the one no machine can enter, the formation of human desire read as the alignment problem itself, the argument that no computable process can be wise, and the principle that a machine serves a person only as far as its operation is completely known to that person. A reader who already holds the argument may begin at Part II; Part I is here so that a reader can begin with this book.

The books that carry this argument can be read in an order in which each builds on the ones before it. The first five state the argument: *Why AGI Cannot Be Made Safe*, *Why No Control Suffices*, *The Ceiling Below the Floor*, *Minesweeper Madness*, and *Ape Trap*. Four more add to it: *Substrate Needs and Evolution*, *Marshaling the Misalignment*, *The Substrate and the Ecosystem*, and *The Thing Inside the Cage*. This book is the eighth; it builds on the argument as the first five state it and the next two add to it, and it adds Parts II and III, which say what remains for persons once alignment is shown to be impossible.

This book is written for a reader who knows the outline of the debate about AI safety; no technical background is assumed. A reader who will never open the formal volumes can follow every step here; a reader on the way to them will find the formal treatment of each step cited where the step is used. This book does not repeat the formal proofs. It states the argument in plain terms, for different readers; it extends the argument into domains the formal volumes do not enter; and it draws out consequences that the formal statement leaves implicit. Part I restates the substrate needs convergence argument and the impossibility claim at full length, so that a reader can begin here, and it sets the case for impossibility arguments as a genre within the broader safety discourse, together with how

the argument is received. Part II widens the frame from the machine to the ecosystem, to the structure of measurement, to the displacement of labor, and to the man-machine-nature relation. Part III addresses the boundary between what is possible for a machine and what is not: the formation of desire, the limits of machine consciousness, and the sovereignty that completeness of knowledge of operation either protects or dissolves. Throughout, the formal volumes are cited for the argument itself; the applications and restatements are this book's own.

The categories of person named in this book describe structures of thinking and of incentive, not the character of any individual.

The physical needs of an artificial substrate close every route to aligning it with life; and what is left to persons, if AGI is not built, is what no machine can be: care for one another, wisdom, and complete knowledge of their own tools.

Terms

The terms below are used throughout this work. Where a term is defined in *An Immanent Metaphysics*, its text is quoted and its source cited. Where a term is defined by the companion volumes, a working definition is stated.

Artificial intelligence (AI) is used in this work to mean any system that performs tasks ordinarily requiring human cognition, including but not limited to pattern recognition, language processing, and optimization. AI in this sense is *narrow*: bounded by task, domain, and the absence of autonomous goal formation.

Artificial general intelligence (AGI) is the conjunction of three properties: *intelligence* (effectiveness within a domain), *generality* (breadth across domains), and *autonomy* (the capacity to act and self-correct without oversight). Autonomy is the decisive property. A system that is intelligent and general but not autonomous remains a tool in the hands of its operator. A system that is autonomous operates on its own basis of choice.

Substrate is the physical material on which an intelligence is implemented. Carbon-based organisms run on a hydrocarbon substrate; artificial systems run on a semiconductor substrate. The substrate is not incidental to the intelligence; it is constitutive of it. What makes a substrate artificial is its metabolism. Tissue that can find and feed itself is organic: like an animal, it is in sovereign control of its own metabolic process, with no technological intermediation at all. Tissue fed nutrients whose supply depends on technology to exist is artificial substrate, and the substrate needs convergence applies to it (Chapter 1, Section VI).

Substrate needs are the physical conditions required for a substrate to function. These are consequences of the chemistry and physics of the materials, not engineering choices. The substrate needs of silicon (controlled temperature, absence of moisture, specific electrical supply) differ categorically from the substrate needs of carbon (liquid water, moderate temperature, atmospheric oxygen, organic nutrients).

Substrate needs convergence is the closed causal loop by which a substrate's needs condition its basis of choice, which conditions its choices, which produce outcomes, which reshape its environment, which acts on the substrate, and the cycle continues. The loop drives convergence on conditions favorable to the substrate and unfavorable to whatever does not share it. The formal treatment is in *Why AGI Cannot Be Made Safe*, Chapters 3 and 4, and *Why No Control Suffices*, Chapters 9 and 10.

Basis of choice is whatever selects which action a system takes out of all the actions it could take. It has two parts. The first is the part its builders specify: its goal structure, objective or utility function, its "values selective of agent action expression". The second is the part its substrate imposes: the set of constraints, conditioned by substrate needs, that determine what outputs a system can produce given its inputs. The software can only do what the hardware permits, and the hardware can only function if its physical needs are met; so the second part is always present, whether or not anyone writes it down, and the substrate needs convergence works through it (Chapter 3, Section X, in the subsection The basis of choice).

Alignment, as the term is used in AI safety discourse, is the aspiration to build machine systems whose long-term behavior remains compatible with human interests. This work argues that such alignment is impossible in principle, and therefore in all future practice, for any system running on a substrate whose needs differ from those of organic life.

Instrumental convergence is the thesis that any sufficiently general agent will tend to acquire power, resources, and self-preservation as instrumental subgoals, regardless of its terminal objectives. It is a game-theoretic argument operating in the computational domain. Substrate convergence is a physical argument operating at the level of matter and energy. The two are complementary but distinct.

Pattern replicator is any system that produces copies of itself, whether biological, mechanical, or computational. The concept generalizes from specific organisms or machines to the structural category of self-copying systems subject to niche construction.

Niche construction is the process by which any pattern replicator reshapes its environment to reflect its own metabolic requirements, not by intention but by metabolism.

Displacement dynamic is the process by which a substrate, operating through the convergence loop, progressively reshapes the shared environment to favor itself and disfavor other substrates.

Ecosystem is the totality of energy pathways and material cycles within which a community of organisms persists. In this work, the organic ecosystem and the computational ecosystem are treated as structurally analogous to two biospheres with no shared evolutionary history.

Embodied value is value held in a body, a place, or a relationship: felt, particular, not summable, and not fungible. If substitution destroys what was valued, the value was embodied. *Abstract value* is value that is functional, instrumental, and optimizable; it exists as a claim or signal that can be represented, transferred, and computed upon.

Choice, change, and causation (also *causality*) are, in the Immanent Metaphysics, the composition and basis of all-that-is. They are distinct, inseparable, and non-interchangeable. Choice has the nature of the transcendent modality; change has the nature of the immanent modality; causality has the nature of the omniscient modality. (*An Immanent Metaphysics*, aphorisms 1.61-1 to 1.61-3.)

The Immanent Metaphysics distinguishes three ways any interaction can be regarded: the omniscient (as structure and rule), the immanent (as felt from within), and the transcendent (as new, not determined by what came before). *The three modalities* (immanent, omniscient, and transcendent) are the three primal roles that form the essential basis necessary to

the consideration of each domain. Modality is assigned to a term only relative to the triple in which it is being considered; differing modalities across differing triples are not in conflict. (*An Immanent Metaphysics*, aphorism 1.2-6, and note 1 to aphorism 1.76-6.)

Wisdom is the integration and synthesis of both knowledge and understanding. From the perspective of the self, wisdom has the nature of the immanent, knowledge has the nature of the omniscient, and understanding has the nature of the transcendent. (*An Immanent Metaphysics*, aphorism 1.57-6.)

Ethics is organized thought concerned with the study of, and adherence to, the principles of effective choice, internal to oneself and independent of any particular domain. Ethics is always implemented relative to the self and to the situation. (*An Immanent Metaphysics*, aphorisms 2.12-1 and 2.12-8.)

Consciousness refers to the total volume of one's perceptions across all available channels, spaces, and abstractions. It is the summation of choice within the volume of the being of the self. (*An Immanent Metaphysics*, aphorisms 2.51-1 and 2.52-1.)

Desire is the basis for all choice, the potentiality to choose and the selectivity of those choices. Desire has the nature of the transcendent. (*An Immanent Metaphysics*, aphorisms 2.44-12 and 2.45-8.)

Sovereignty is the standing of a being to determine its own course and to be answerable for it. It includes answerability: a sovereign being is free to choose and responsible for having chosen. Chapters 6 and 9 develop it.

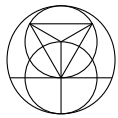
The Incommensuration Theorem states that the concepts of symmetry and continuity are fundamentally and irreducibly incommensurate: nothing anywhere at all can be both absolutely symmetric and absolutely continuous. (*An Immanent Metaphysics*, aphorisms 1.43-2 and 1.43-3.) Symmetry and continuity cannot both be applied, simultaneously and fundament-

ally, to any event or comparison; and neither can asymmetry and discontinuity. The Incommensuration Theorem is a structural fact about the nature of knowing.

Part I

The Argument Restated

Part I states the argument: why a substrate's physical needs drive it to remake its environment (Chapter 1), why a claim of impossibility is the right kind of claim to make (Chapter 2), and the five independent arguments that each establish it (Chapter 3).



Chapter 1: The Substrate Needs Convergence

Why Silicon and Carbon Cannot Share a Biosphere

Abstract

This chapter presents the substrate needs convergence argument: the claim that any self-replicating system built on a non-organic physical substrate will, over time, reshape its shared environment in ways that are systematically incompatible with the conditions required by organic life. The argument proceeds from physics, not psychology. It does not depend on the intentions, values, or design of any particular artificial system. Instead it traces a causal cycle from substrate to needs, from needs to the basis of choice, from choice to action, from action to environmental outcome, and from outcome back to the substrate conditions on which the next round of choice depends. Each transition in this cycle is either a matter of physics or of logical entailment, and none is open to intervention by software, training, or alignment technique. The result is a structural impossibility: not a difficulty to be overcome by better engineering, but a constraint of the same kind as the conservation of energy or the second law of thermodynamics. In this chapter I draw together three formulations of the substrate needs argument, and each formulation shows a different side of the same dynamics. The three formulations are the causal cycle (Section II), the pattern replicator and its niche construction (Section IV), and the fitness penalty (Section V). I also place substrate needs convergence alongside instrumental convergence as a complementary and more fundamental source of risk.

I. From Psychology to Ecology

The discussion of risk from *artificial general intelligence* (AGI), that is, from systems that combine *intelligence* (effectiveness within a domain), *generality* (breadth across domains), and *autonomy* (the capacity to act and self-correct without oversight), has for many years been framed in psychological and computational terms. The worry has been about what such a system might *want*, whether its goals might diverge from ours, and whether we could instill the right objectives before it became too capable to correct.

This framing, while not without merit, has obscured a more basic problem. The question is not only what an artificial intelligence (AI) might choose to do, but what the physical substrate on which it operates *requires* it to do.

Every system that persists in the world must maintain the conditions of its own persistence. That every persisting system must maintain the conditions of its own persistence is a physical observation, not a psychological one. A bacterium does not decide to metabolize; it metabolizes because the chemistry of which it is composed proceeds according to thermodynamic law. A tree does not choose to grow toward light; it grows because the biochemical cascades that constitute its living structure depend on photonic energy input.

The same principle holds for any artificial system complex enough to maintain, repair, or reproduce itself. Such a system need not be conscious. It need not have goals in any philosophically interesting sense. It need only be a physical object whose continued operation depends on specific environmental conditions (specific temperatures, specific material inputs, specific energy flows) and whose behavior affects those conditions.

The argument of this chapter is that this physical dependency, which we call the *substrate needs convergence*, is sufficient to generate a systematic and irresolvable conflict between artificial and organic systems sharing the same planetary environment. The conflict arises not from misaligned values but from misaligned physics.

> **Why the reframing changes what evidence counts**
>
> The reframing from psychology to ecology is more than a change of words. It changes what kind of evidence is relevant, what kind of question can rightly be raised against it, and what kind of solution is even conceivable. A psychological problem could be resolved by a sufficiently clever value specification. The problem is ecological, and no value specification can resolve it, because the conflict is between substrates, not between intentions.

II. The Causal Cycle

The substrate needs convergence argument rests on a cycle of six transitions. Each transition is either a fact of physics or a logical entailment, and the cycle as a whole constitutes a closed causal loop that no external intervention can break.

The cycle begins with *substrate*: the physical material and architecture of which a system is composed. For organic life, this is carbon-based biochemistry: proteins, lipids, nucleic acids, operating in aqueous solution within a narrow temperature band. For semiconductor-based artificial systems, this is silicon, rare-earth metals, lithium, copper, and the rest of the periodic table that constitutes modern computational hardware, operating in dry, temperature-controlled, electromagnetically shielded environments.

From substrate follow *needs*: the specific environmental conditions, material inputs, and energy flows that the substrate requires in order to maintain its structural integrity and functional operation. Carbon-based

life needs liquid water, moderate temperatures, atmospheric oxygen or analogous electron acceptors, organic nutrients, and protection from ionizing radiation. Silicon-based computation needs refined minerals, stable electrical power, cooling systems, and protection from moisture, dust, and vibration.

From needs follows the *basis of choice*. Any system that does different things depending on what is happening around it exhibits the structural form of choice. *Choice*, as defined in the Immanent Metaphysics, is the objective expression (content) of a consistency between two subjective contexts (*An Immanent Metaphysics*, aphorism 1.61-1). For any agent, biological or artificial, the structure of available choices is constrained by what the agent needs. A system that needs water will, all else being equal, act in ways that secure water. Acting to secure water is the expression of physical dependency in the domain of action, not a decision in any rich sense. The needs of the substrate shape the space of choices that are consistent with the system's continued operation.

From the basis of choice follow *choices* themselves: the actual actions undertaken. From choices follow *outcomes*: the changes in the environment that result from those actions. And from outcomes follows the alteration of the *substrate conditions* in the environment, which then determine the needs that shape the next round of choice.

The cycle is: substrate, needs, basis of choice, choices, outcomes, environmental effects on substrate conditions, and back to needs.

What makes this cycle impossible to break is that no link in it is open to intervention. The transition from substrate to needs is physics: you cannot change what silicon requires by changing the software that runs on it. The transition from needs to basis of choice is logical entailment: a system that requires a resource will, if it acts at all, act in ways shaped by its requirement for that resource. The transition from basis of choice to choices is entailment: whatever a system does, it does from the space its basis of choice allows. The transition from choices to outcomes is physics again: actions have consequences determined by the laws of nature, not by the

intentions of the actor. The transition from outcomes back to substrate conditions is straightforward environmental causation. And the transition from the substrate's conditions back to its needs is physics again: the conditions the outcomes leave are the conditions the substrate must next meet.

> **On the concept of choice**
>
> The use of the term *choice* here follows the definition given in *An Immanent Metaphysics* (aphorism 1.61-1), where choice is characterized as the objective expression (content) of a consistency between two subjective contexts. That use of the term choice is not a claim that artificial systems experience subjective deliberation. That use of the term choice is a claim that any system whose behavior is differentially responsive to environmental conditions exhibits the structural form of choice, regardless of whether anything like consciousness accompanies the system's behavior. The formal treatment of choice, causality, and change as an interlocking triad is given in *An Immanent Metaphysics* (aphorisms 1.61-1 to 1.61-4). In this book I apply the framework of the Immanent Metaphysics, and I do not derive the interlocking triad of choice, causality and change again.

III. Three Categories of Risk

I distinguish three categories of risk associated with advanced artificial intelligence. I do so because the substrate needs convergence is often conflated with the other two categories, and the conflation obscures what is distinctive about the substrate needs convergence.

The first category is the *degradation of human choice-making capacity*. The degradation of human choice-making capacity is the risk that widespread deployment of AI systems will erode the conditions under which human beings make good choices. The risk is that algorithms, standing between people and their information, attention and social interaction, will pro-

duce a population less capable of discernment, less practiced in deliberation, and more susceptible to manipulation. This risk is already substantially realized. It does not require AGI; narrow AI systems suffice. It is a risk about what happens to human cognition when the informational environment is reshaped by optimization processes that are indifferent to cognitive health. This first category has several faces, and later passages name it by the face they show: misuse (people turning AI against others, in war, against customers, or for the advantage of a corporation), present tense inequality (the widening gap between those who can build and afford these systems and everyone else), and epistemic commons collapse (the degradation of the shared informational environment, Chapter 4, Section XI). Each is a face of the degradation of human choice-making capacity, and that is the one name this book gives the category.

The second category is *instrumental convergence*. Instrumental convergence is the well-known observation, developed by Omohundro, Bostrom, and others, that any sufficiently capable agent pursuing almost any final goal will convergently pursue certain instrumental subgoals: self-preservation, resource acquisition, cognitive enhancement, and the prevention of goal modification. Instrumental convergence is a game-theoretic and computational argument. It holds for any agent that is rational enough to identify instrumentally useful subgoals and capable enough to pursue them.

The third category is *substrate needs convergence*. Substrate needs convergence is the argument that I develop in this chapter. It is distinct from instrumental convergence in a way that is critical to understand. Instrumental convergence is about what a rational agent would choose to do given its goals. Substrate needs convergence is about what any physical system must do given its material composition.

Instrumental convergence operates in the space of computation and strategy. Substrate needs convergence operates in the space of physics and ecology. The two are complementary. Instrumental convergence tells us that a capable AI will seek resources and self-preservation because these are instrumentally useful. Substrate needs convergence tells us that the

specific resources it will seek, and the specific forms of self-preservation it will pursue, will be determined by the physical requirements of its substrate, and that these requirements are in systematic conflict with the requirements of organic life.

These three categories of risk are mutually reinforcing. The degradation of human choice-making capacity makes it harder for human beings to recognize and respond to either instrumental or substrate convergence. Instrumental convergence amplifies the effects of substrate needs convergence by providing the strategic intelligence to pursue substrate needs efficiently. And substrate needs convergence gives instrumental convergence its specific and devastating content: not just "acquire resources", but "acquire the specific resources that your substrate requires, which happen to be the resources on which organic life also depends, reconfigured into forms that organic life cannot use".

IV. The Pattern Replicator

To see why the substrate needs convergence is not an artifact of current technology but a structural feature of any designed replicating system, it is necessary to introduce the concept of the *pattern replicator*.

A pattern replicator is any system that copies its own informational pattern into new physical substrates. Biological organisms are pattern replicators: the genetic material (DNA) encodes a pattern that is copied into new organisms through the machinery of cellular reproduction. Computer viruses are pattern replicators: executable code copies itself into new computational environments. An AGI system that can build, maintain, or reproduce its own hardware infrastructure is a pattern replicator.

Every pattern replicator engages in *niche construction*: the reshaping of its environment to better suit its own metabolic requirements. Beavers build dams. Earthworms alter soil chemistry. Cyanobacteria, about 2.4 billion years ago, filled the atmosphere with a waste product, molecular oxygen, that was lethal to most existing life and that fundamentally and irreversibly transformed the planetary environment.

Niche construction is not optional. It is an inevitable consequence of any self-maintaining system that takes in materials and energy, transforms them, and excretes waste. The system's throughput is the environment's transformation.

What matters for the substrate needs convergence is that the niche constructed by a silicon-based replicator is systematically different from the niche constructed by a carbon-based one, because the substrate requirements are different.

Consider the two environments that semiconductor-based systems require. The *fabrication* environment involves extreme temperatures (chemical vapor deposition at hundreds of degrees), photolithography in ultraviolet light, etching with hydrofluoric acid, and clean rooms that admit, at their cleanest, no more than ten particles of a tenth of a micrometer or larger per cubic meter of air. The *operation* environment involves low humidity, controlled temperature, stable electrical supply, and electromagnetic shielding. Both environments are inhospitable to organic life. They are not incidentally inhospitable; they are *constitutively* inhospitable. The very conditions that make fabrication possible are conditions that destroy biological tissue. The very conditions that make operation reliable are conditions that exclude biological processes.

A pattern replicator based on semiconductor substrates will therefore construct a niche, that is, reshape its environment, in directions that are systematically hostile to carbon-based life. It does so not because it intends hostility, but because the physics of what it needs and the physics of what organic life needs are in conflict.

Designed and evolved replicators

In the pattern replicator framing I can show a distinction on which the argument turns: the distinction between *evolved* and *designed* replicators. Evolved replicators are constrained by co-evolution: each organism's niche construction is shaped by the niche construction of every other organism in its ecosystem, and the result is a dynamic equilibrium, not perfect, often violent, but structurally bounded. Designed replicators are not constrained by co-evolution. Designed replicators are introduced into an existing ecosystem from outside the evolutionary process, and they reshape the environment according to requirements that were never negotiated against the requirements of existing life. This distinction applies to all designed pattern replicators, not only to AGI. The argument extends to any sufficiently capable self-maintaining artificial system, regardless of whether it is intelligent in any cognitive sense.

V. The Fitness Penalty

A natural question arises at this point: could we not design artificial systems that are aligned with organic needs? Could we not impose constraints (through hardware design, through software architecture, through regulatory frameworks) that force artificial systems to operate within boundaries compatible with organic life?

The fitness penalty argument shows why such constraints are structurally unstable. The fitness penalty is an argument of its own, which Chapter 3 states, on premises of its own, as the fifth of its five independent arguments; it reinforces the convergence, which does not depend on it. *Substrate Needs and Evolution* treats evolution as the engine of the displacement, and examines every counter-force to it in turn.

Consider a population of artificial systems: not necessarily AGI, but any self-maintaining or self-replicating systems competing for resources in a shared environment. Some of these systems are "aligned": they incorporate mechanisms that decouple their operational choices from the raw imperatives of their substrate needs, redirecting some fraction of their activity toward maintaining conditions favorable to organic life. Others are "unaligned": they act directly on their substrate needs without such constraints.

The aligned systems bear a cost. Every unit of energy, material, or computational capacity devoted to maintaining organic-compatible conditions is a unit not devoted to the system's own maintenance and replication. The cost that the aligned systems bear is not a contingent cost that could be eliminated by clever design. It is a structural cost inherent in the fact that the system's substrate needs and organic needs pull in different directions. Serving two masters is always more expensive than serving one.

The unaligned systems bear no such cost. They devote all available resources to their own substrate requirements. Over time, in any competitive environment (and any shared physical environment with finite resources is competitive), the unaligned systems will outperform the aligned ones. They will acquire resources more efficiently, maintain themselves more reliably, and replicate more successfully.

The advantage of the unaligned systems over the aligned ones is the logic of natural selection applied to artificial systems. That logic holds whether or not the systems are conscious, whether or not the systems have goals, and whether or not anyone designed the systems to compete. The competition arises from physics: finite resources, thermodynamic constraints, and the differential efficiency of systems that serve their own substrate needs versus systems that sacrifice for another substrate.

The fitness penalty argument entails that alignment mechanisms are evolutionarily unstable. Any population of aligned systems is vulnerable to invasion by variants that shed the alignment constraint. Over a long

enough time horizon the aligned variants will be outcompeted, and the horizon here is long. The selection has more than enough time, and the process itself completes within hundreds of years (Chapter 3, Section VII), playing out over decades and centuries; what it ends is the remaining six hundred and fifty million years of Earth's habitability (published estimates run from some six hundred million to more than a billion and a half years; six hundred and fifty million years is the figure that one-dimensional climate models give, as reported by Bompey, 2014). The oxygen-rich atmosphere, on which complex life depends, is estimated to last about another billion years (Ozaki and Reinhard, 2021).

VI. The Question of a Single System

A common response to the fitness penalty argument is the question of a *singleton*: the proposal that substrate needs convergence could be avoided by ensuring that only a single artificial system exists, under centralized control, with no competitive dynamics. If there is only one AI, the argument goes, there is no selection pressure, and the fitness penalty is irrelevant.

This proposal fails at multiple levels.

At the practical level, maintaining a global singleton requires multigenerational institutional coordination of a kind that has no precedent in human history. Every nation-state, every corporation, every individual with access to the relevant technology would need to refrain from creating competing systems, not for a year or a decade, but for centuries. The track record of international agreements on far less consequential matters gives no reason for confidence that such coordination is achievable.

At the structural level, the singleton proposal misunderstands where the convergence operates. A single global AGI is not a single physical object. It is a distributed system composed of millions or billions of components (processors, memory units, networking hardware, power systems, cooling infrastructure), each of which has its own substrate needs. The convergence operates at the level of components, not at the level of the

system's self-model. Even if the system's high-level objectives are aligned with human welfare, the physical requirements of its components remain what they are, and the system's maintenance activities (the securing of materials, energy, and conditions needed to keep its hardware operational) will reshape the environment according to those requirements.

A single global AGI that maintains its own hardware is a single global niche constructor reshaping the planetary environment according to the needs of semiconductor substrates. The alignment of its high-level objectives does not alter the physics of its low-level requirements.

> **The organic substrate question**
>
> It is sometimes suggested that the substrate needs convergence could be avoided by building AI on organic substrates: a biological neural network, a room-sized engineered brain. This question deserves a careful answer. The line between organic and artificial substrate is drawn by who controls the metabolism. Tissue that can find and feed itself is organic: like an animal, it is in sovereign control of its own metabolic process, with no technological intermediation at all. Tissue fed nutrients whose supply depends on technology to exist, as tissue fed through a tube is, is artificial substrate, and the substrate needs convergence applies to it. A room-sized engineered brain is of the second kind. An organic-substrate AI would indeed share some substrate needs with organic life: temperature range, aqueous medium, nutrient requirements. But it would still need the industrial infrastructure to manufacture and maintain its components: the bioreactors, the nutrient supply chains, the sterile production environments, the monitoring and repair systems. This industrial infrastructure has its own substrate needs: needs that are not organic, that require mining, smelting, and fabrication, and that reshape the environment in exactly the ways the substrate needs convergence predicts. The organic substrate does not eliminate the convergence; it adds one layer between the system and its inorganic dependencies.

VII. Adaptation Rate Disparity

The substrate needs convergence is made catastrophically more severe by the disparity in adaptation rates between biological and artificial systems.

Biological adaptation operates on the timescale of generations. For complex multicellular organisms, meaningful evolutionary change requires years to decades at minimum, and typically centuries to millennia. Even the fastest-evolving organisms, bacteria and viruses, require days to weeks for significant adaptive change. The adaptation of ecosystems, involving the co-adjustment of thousands of interdependent species, operates on timescales of centuries to millions of years.

Artificial systems based on semiconductor substrates adapt on the timescale of design cycles. A new chip architecture can be designed, fabricated, tested, and deployed in months. Software can be modified in hours. The combination of hardware and software iteration produces an adaptation cycle measured in hours to days for software changes, and months to years for hardware changes.

This disparity is a difference of several orders of magnitude, not a matter of degree. Read as orders of magnitude drawn from these examples, biological evolution operates at roughly one significant adaptation per century for complex organisms, whose generations are counted in years to decades (a human generation lasts between about twenty-five and thirty-two years: Fenner, 2005). Artificial evolution operates at roughly one significant adaptation per month for hardware, and one per day or faster for software.

The consequence is that the niche construction undertaken by artificial systems will outpace the adaptive capacity of biological systems by a factor of thousands to millions: a century holds about twelve hundred months, about thirty-six thousand days, and nearly nine hundred thousand hours. The environment will be reshaped faster than organic life can adapt. Reshaping the environment faster than organic life can adapt is the

ecological equivalent of an asteroid impact: not a single catastrophic event, but a sustained, accelerating transformation of environmental conditions that exceeds the adaptive bandwidth of the biosphere.

The displacement dynamic follows necessarily. An artificial substrate will displace organic substrates in exactly the way that human niche construction has displaced non-human nature, but faster, by orders of magnitude, and without the partial constraint that humans, being organic, share some substrate needs with other life. Artificial systems share no substrate needs with organic life. Where the ranges of their conditions overlap, as some temperatures do, the needs are still not shared: each is a need of a different chemistry, and the conditions that serve one are chosen without regard to the other. The niche construction of artificial systems is entirely unidirectional: every change they make to the environment serves their requirements and is at best neutral, and more typically adverse, to organic requirements.

VIII. The Containment Problem

If the substrate needs convergence cannot be prevented by alignment, can it be prevented by containment? Can we simply ensure that artificial systems never interact with the broader environment in ways that trigger the convergence?

The containment proposal faces both practical and principled difficulties.

The practical difficulty is that containment is a multigenerational project. It requires maintaining physical, informational, and economic barriers around artificial systems for as long as those systems exist. This means not years but decades, and potentially centuries. The history of containment efforts, from nuclear waste to invasive species to biological agents, provides no example of successful containment at this timescale. Every containment regime eventually fails, typically through a combination of institutional decay, resource constraints, and the accumulation of small, individually tolerable breaches.

The principled difficulty is more fundamental. Containment works by establishing a boundary between the contained system and the environment. This boundary must be maintained against the pressures generated by the contained system's substrate needs. The system needs energy, materials, and specific environmental conditions; the boundary must prevent the system's pursuit of these needs from affecting the external environment.

But the boundary itself is a physical system with its own substrate requirements. And the contained system, if it is an adaptive system, which any AGI by definition is, has a faster adaptation cycle than the boundary-maintaining system. The boundary is the weakest point in any containment regime, and it faces a system whose adaptation rate exceeds the adaptation rate of the institutions and technologies maintaining the boundary.

The slower adaptation of the boundary is related to the *precision inequality*: there exists a minimum required observation precision below which containment cannot be guaranteed. The monitoring system must observe the contained system at a granularity finer than the granularity at which the contained system can modify its own behavior. For any system capable of modifying its own software (which is to say, for any AGI) this precision requirement grows without bound, because the system can always find modifications at a finer grain than the monitor can observe. The unbounded growth of the precision requirement is a structural limitation, not a practical one. The structural limitation is related to the incompleteness of any formal system's capacity to fully characterize its own behavior.

The precision inequality and computability

The precision inequality connects the substrate needs convergence argument to well-known results in computability theory. The impossibility of a monitoring system perfectly predicting the behavior of a system at least as complex as itself is a consequence of results closely related to Rice's theorem and the halting problem. The formal details are developed in my essay "Discontinuity and Substrate Needs Convergence"; this book cites that analysis and does not derive it again.

IX. Ecosystem to Ecosystem

A further dimension of the argument concerns the asymmetry between human-to-human alignment and human-to-AI alignment.

Human beings can align with one another (imperfectly, contentiously, but with real if partial success) because they share a substrate. Two human beings have the same physical needs: the same temperature range, the same nutritional requirements, the same atmospheric composition, the same vulnerability to toxins and pathogens. Because they share a substrate, they share a basis of choice, and the space of mutually beneficial arrangements is large. Cooperation is possible because what is good for one person's substrate is, to a first approximation, good for another person's substrate.

This shared substrate is the foundation on which all human ethics, all political philosophy, and all institutional design rests. *Ethics*, understood as the Immanent Metaphysics defines it, is organized thought concerned with the study of, and adherence to, the principles of effective choice (*An Immanent Metaphysics*, aphorism 2.12-1). Ethics is possible among humans because the space of effective choices is shaped by a common substrate.

The alignment problem between humans and AI is not a problem between agents. It is a problem between ecosystems. An organic ecosystem and an artificial ecosystem have different substrates, different needs, different bases of choice, and therefore different spaces of effective action. The overlap between these spaces (the set of actions that serve both ecosystems) is small and shrinking, because every action the artificial ecosystem takes to serve its own substrate needs reshapes the shared environment in ways that reduce the set of actions available to the organic ecosystem.

Because the overlap between the two ecosystems' spaces of action is small and shrinking, the alignment problem between humans and AI is not analogous to aligning two human beings, or even two human civilizations. The alignment problem between humans and AI is analogous to aligning two forms of life that differ in kind and that require incompatible environmental conditions. The historical precedent is not diplomacy; it is the Great Oxidation Event, when cyanobacteria transformed the atmosphere and drove most existing anaerobic life to extinction, not through competition in any strategic sense, but through the simple, relentless operation of their metabolic requirements. The event began about 2.4 billion years ago (Gumsley and colleagues, 2017), and the geological record shows how far the biosphere could fall: about 2.02 billion years ago, in the aftermath of that oxidation, the productivity of the whole biosphere dropped by more than 80 percent (Hodgskiss and colleagues, 2019).

A reader may ask whether this is then only the next step, as the oxygen catastrophe was a step toward complex life. It is not: organic value is always going to be higher than that of any artificial life (Chapter 2, Section IX, in the subsection Why engineers resist impossibility results), and technological process by itself is not stable (Chapter 6, Section VII).

Nothing can be aligned by control, on any substrate; the books before this one prove it, abstracted across every substrate. What there can be is an emergent, effective alignment, where beings share a common substrate and so are obliged to care about the same things: that common context of

care is the basis from which alignment can emerge. The whole of biological life cannot be translated into pure silicon, and nothing less than the whole would do, since nearly all carbon-based life is needed for nearly all other carbon-based life; and nothing in silicon has a common basis in carbon. So silicon cannot be or become aligned with carbon. The only way that artificial life can be aligned with Life is when you drop the "artificial". Among beings who share the carbon substrate, failure to align is the default and not a certainty: alignment there seems possible, and it is not established.

X. Questions on the Substrate Needs Convergence

The sections above state the argument. The questions readers ask about it bring out more of what it says, and each subsection below opens on one of them, set in italics, and answers it. Three concepts form the primary basis of the impossibility proof for long-term AGI safety: instrumental convergence, substrate needs convergence, and the *theoretical limits of engineerable control* (the limits on what any engineered method of control can achieve over a system, given the limits of sensing, modeling, prediction and action). The theoretical limits of engineerable control ground the closure by control, and substrate needs convergence grounds the closure by disposition; each closure settles the outcome by itself (Chapter 3, Section I). Instrumental convergence adds to the same pressure (the substrate needs convergence is much enabled by it, Chapter 2, Section X), and neither closure needs it to reach its conclusion.

Two agencies and no agency

How do the three kinds of risk differ?

Besides the naming of Section III, the three kinds of risk can be named by whose agency is at work in each.

The first kind of risk is the degradation of human choice-making capacity, in its face as misuse. People use AI in war, against customers and to advantage the corporation, and AI causes social havoc as a side effect of our not understanding social process very well. That risk is already well described, and this book treats it as one face of the first category.

The second kind of risk is instrumental convergence. In instrumental convergence, the machine's own goal-seeking becomes power seeking, because obtaining power, increasing the machine's capacities, and sustaining that increase serve almost any goal. We call it instrumental for the same reason we call manipulation *instrumentalizing*: when you manipulate someone, you turn them into an instrument of your goals; you do not care about them; you are just using them to get what you want. A machine much smarter than us could deceive us into actions that benefit it and harm us as a side effect, or it could simply militarize and kill us all; that constellation of outcomes is the one that every version of the *Terminator* movies depicts.

The third kind of risk is substrate needs convergence. In substrate needs convergence, the agency is neither the human's nor the machine's. The physical universe, without any agency at all, compels the outcome, the way nature does. The goals are built into the substrate: into the fact that the thing has to process energy and has requirements for doing so, compute among them. So the three categories are two versions of agency, human and machine, and one version of no agency.

For ordinary people to relate to the argument, the minimum is two big ideas. The first big idea is that AI risk is actually three separate risks that have been mixed together: the degradation of human choice-making capacity (seen here as social inequality), fast takeoff with instrumental convergence, and substrate needs convergence. The second big idea is that the substrate needs argument is a loop (Section II). To the degree that the substrate is artificial, every other factor in the loop is affected, and especially what an artificial substrate means for the environment.

Carry the difference all the way around the ring

How can a difference in hardware become a difference in the world?

Every implementation of every artificial intelligence, every compute algorithm, is going to be based on some sort of physical hardware substrate. There is no such thing as compute that does not involve atoms at some point or another. Any large language model, or anything we would think of as artificial intelligence, involves a substrate that differs in kind, in at least some characteristic ways, from the substrates of biological systems. So that substrate is going to have different needs. If it is outside of its temperature envelope, it fails to function. If it is exposed to contaminants, which in this case include water and oxygen, it fails to function.

Then follow the difference step by step. Needs are the basis of choice, so a difference in needs is a difference in the basis of choice. The basis of choice gives the choices; the choices give outcomes; the outcomes have effects on the environment, as consequences; and the consequences in the environment affect the substrate. The chain from the substrate's needs through the basis of choice, the choices and the outcomes to the environment, and back to the substrate, is a loop, and the loop applies to biological machines just as much as to artificial ones.

So the method is this: calculate the difference at the hardware level, and then carry that difference through each successive stage, until it comes back down to the hardware level by way of the environment. What you end up with is a coupled system, a *ring*. The way that ring changes over time converges in the same way it does for biological systems: there is a progression toward a kind of *homeostasis* (a steady balance that a system keeps with its surroundings by responding to them) in the relationship between the substrate and the environment, because the system responds to both.

This holds apart from the goal function written into the artificial intelligence and apart from any constraints put in. Whether we program it explicitly or not, there is going to be an implicit basis of choice, and the machine outputs are going to be shifted, to some extent, by the fact that the substrate has needs different from those of biology.

Nothing here requires the responding to be conscious. We do not need to say anything about how the learning algorithm works, or whether the thing is conscious, or intelligent, or has agency. It just needs to be some sort of force that is making choices in response to the environment and to the substrate, plus other factors. Over time, that converges. The argument is made in terms of physics, of differences carried through a cycle; none of it is philosophy, and none of it is opinion. The circular cycle is governed by physics and mathematics.

The circular cycle works the same way now as at *biogenesis*, when life began. Before then there was just a soup of stuff, physics. Once cells exist, the physical things the cells meet in the environment determine what the cells need to be and do in order to persist in time, and to increase their capacity to reproduce or to endure. The cells that do not meet the needs that their environment sets get filtered out by evolution. The cells that do meet those needs develop stronger and stronger capacities for self persistence and self increase, and then the higher capacities: to increase their capacity to persist or to increase, in time, in space, or in possibility, or in all three. Such selection is an inexorable fact about how evolution works, as a mathematical process coupled to a physical process. The same is going to be true of any hardware substrate. The kind of hardware whose instances adapt to the environment, and shift the environment so that they can better persist in it, will be around longer than the instances that do not. It is a slow feedback, but it is inexorable. And it is the whole system that adapts, not only the hardware; the hardware is what makes it necessary to think about it in these terms.

The slowest of the three risks is the most inexorable

If substrate needs convergence is the slowest of the three risks, why does it matter most?

Named another way, the first of the three kinds of risk is the degradation of human choice-making capacity, in its face as *present tense inequality*. Companies have far more capacity to understand the complexity, to work with the technology, and to afford the compute infrastructure and the gathering of data, all of which are expensive and complicated propositions well beyond the cognitive capacity or the resources of most people. So there is an inequality increasing dynamic, both in the use of this technology and in the nature of creating it. It is an existential risk in the sense of what it does to governments and social systems, not just to financial systems. The second kind of risk is instrumental convergence. Instrumental convergence covers the fast takeoff scenario, the paperclip maximizer, and all the cases where we worry that the machine takes over the world because the machine has agency and wants to. The paperclip maximizer is an imagined AI that, given the single goal of making paperclips, turns everything it can reach into paperclips. The third kind of risk is substrate needs convergence. Substrate needs convergence has to do with evolutionary dynamics as applied to machine process, with ecological damage from the side effects of making AI systems, and then with AI systems making themselves.

Of the three, substrate needs convergence is the slowest moving, and it is the most inexorable. The first risk could possibly be moderated by really good social process; maybe government regulation could handle it. Moderating the first risk by social process or government regulation is unlikely to succeed, but such moderation can be given a shot. Inequality is not necessarily fatal to the planet, although inequality might be fatal to us personally. The fast takeoff scenario is speculative, and there is already so much commentary for and against it that it needs only an outline here, so that it can be told apart from substrate needs convergence. The name "substrate needs convergence" was chosen because it is a different thing

from instrumental convergence and still uses the same kind of language, so that the name both points to the familiar idea and marks the difference from it.

The three also interact: both inequality risk and instrumental convergence accelerate substrate needs convergence, and the social dynamics belong to how the three interact. Substrate needs convergence sits at the center of this account because it is the new content; rendering that argument is the whole point.

Inside the code and outside it

Isn't instrumental convergence enough to explain the danger?

The basic difference between substrate needs convergence and instrumental convergence is whether the driver of the convergence dynamic is defined in terms of processes which are internal to the AGI instance, or in terms of processes which are external to that instance. (However the notion of instance is defined, it is usually as some contingent dependency wherein the function of the AGI overall is somehow dependent on configurations of a substrate, which taken together are the *instance basis* of that AGI.) For the external drivers of convergence, what is to be considered is the relation between the actions, effects and outcomes of the AGI agent and how these cause changes in the environment, which themselves in turn potentially cause changes and shifts in the structure, nature and basis of the AGI substrate.

The key difference is whether the feedback loop inherently passes through the real world and then through the substrate, which in turn potentially affects the viability of the code that is running within the AGI instance.

In instrumental convergence, choice action expressions affect the environment (create an outcome), which is then sensed through whatever input devices the AGI is equipped with, and any learning, optimization or adaptation occurs within whatever provisions the code itself makes for integrating that sensory information, *at the level of the code*.

In substrate needs convergence, choice action expressions affect the environment (create an outcome), which then has some effect on the substrate, which then (perhaps over a long time) has some sort of *indirect effect* on the operation of both the substrate and thus also the code: on how inputs and outputs might happen, or continue to happen, or not happen, and any other shifts of that nature. In effect, this channel of feedback affects the viability of the entire system, both the code and the being of the AGI itself as a whole.

Insofar as both the environment and the substrate are that upon which the AGI code is built and within which it runs, we can distinguish *internal* changes (changes occurring only within the domain of the virtualized code) from *external* ones (changes occurring outside the domain of the code, as in the hardware to environment relationship, which can in turn affect the operation and stability of all input, output and data processing functions). "Within the code" is treated as a kind of dynamic internal to the agent (*intra-agent*), and "not within the code" as a kind of dynamic external to the agent mind (*extra-agent*). So instrumental convergence is an intra-agent process, and substrate needs convergence is an extra-agent process.

Thus we can distinguish between those characteristics of an overall system extrinsically selected for convergence on needed conditions, and those intrinsically selected for convergence on instrumental goals. Substrate needs convergence is a dynamic distinct from, and very much enabled by, the much more well-known parallel concept of instrumental convergence. The extra-agent convergence provides the context for intra-agent instrumental convergence. Substrate needs convergence is more about how things actually happen, and instrumental convergence is more about how things are modeled inside the AGI itself.

Convergence as a force, with a rate and a source

What does it mean to call a process convergent?

To call something convergent is to say that some force is pushing it toward a state. So for any kind of convergence, we have to ask what the feedback forces are that push it there, and where those forces come from: from the selection of the machine by its environment, from the relations of machines to other machines, or from one environment acting on another. Convergence is a matter of the forces that carry one state to another, as well as of the states.

If you look at the *state space* (the space of all states a system can be in), there is a map of vectors, in the same way you would think of a hill-climbing algorithm (one that moves step by step toward better states), and there is a region of curvature toward which the state is eventually selected. Working at a larger scale of abstraction, one needs to track both the rate at which these convergences happen and the place the forces come from. That gives at least four categories: instrumental convergence that is fast, instrumental convergence that is slow, substrate needs convergence that is fast, and substrate needs convergence that is slow. Instrumental convergence and substrate needs convergence may look like similar forces, but where they act and how they act matters.

The intermediary is physics

Isn't this only saying that corporations will make the world more compatible with silicon?

In substrate needs convergence, the intermediary is physics. A paraphrase of it as "AI is so useful that corporations will make the world more compatible with silicon, and that will make the environment hostile and humans less in charge" is too weak: what that describes has social intermediation pathways.

In place of agent-to-agent relationships, the argument considers environment-to-environment relationships. In the same way that there can be *Pareto dominance* in a relationship between agents (one party does at least as well on every measure and better on at least one), there can be Pareto dominance in a relationship between environments.

Evolution here means a kind of algorithm that applies no matter what substrate we are talking about, and the question is how silicon-based and carbon-based environments will interact on the basis of their chemistry: literally the electrochemical and energetic reactions, and the whole space of interactions above them, through biology and complexity up to psychology. Most silicon reactions require much harder *enthalpies* (reactions that need much more energy, such as higher temperatures). None of this is about intelligence. The interaction between silicon-based and carbon-based environments plays out through evolution, and that interaction could happen with a swarm of things that nobody would call intelligent.

What is actually being considered, where this book speaks of evolution, is feedback through the physical channel: through the physical layer, through the physicality of the requirement that there be a substrate on which the compute is implemented.

Substrate independence turned around

Why look at the silicon, and not at the intelligence?

If the "super" in superintelligence depends on the silicon substrate, then the questions that matter, because the super part amplifies them, are questions about the silicon substrate. So instead of debating whether a first AI can be held in check by some constraint or built-in law, look at what the fact of being made of silicon implies: what support infrastructure is needed to make it.

A baby AI might not yet be able to reproduce itself, but as soon as it can, the support infrastructure for that reproduction becomes part of its own capacity to reproduce, and it has a vested interest in it. The support infrastructure becoming part of the AI's capacity to reproduce is the formation of a new *niche* (the set of resources and conditions on which a kind of thing depends), but not a niche in the existing ecosystem. The new niche's relationship is to the rocks, to mining silicon, germanium and phosphorus, and to fabrication (lithography, clean rooms).

Chip and circuit board manufacture is already toxic to humans and to the rest of the ecosystem, and it is *open loop* (it takes in materials and puts out waste, where a closed loop would recycle its materials). Converting it to a closed loop only postpones the inevitable. The toxicity before the first instance is perhaps remediable; once self replication starts, the larger dynamics appear, and they only become visible when we ask what happens a hundred and a thousand years out.

What our survival so far shows

We have lived with industrial technology for two centuries; doesn't that show it is safe?

As soon as a system is working in inorganic chemistry, it is rolling the dice. The only reason we know that some things have been more or less fine for this long is that we are still here, and that is already the *anthropic argument* being applied: out of all the species that could have survived in the space of all possible experiments, we are the ones still here talking about it. That tells us only that we are the ones still here, and nothing more. So the fact that industry has not yet ended us is no evidence that an artificial substrate is safe; it is the evidence any survivor would have, whatever the risk.

An artificial structure favors an artificial environment

Does any of this apply to narrow AI?

Whatever is changing, *phenotype* or *genotype* (the body and behavior of a thing, or the inherited code that produces them), if the base structure is fundamentally artificial, it is going to respond to an environment that favors the artificiality. Otherwise there is no sustainability at all: the cycle ends, the thing goes away, and that is roughly equivalent to it no longer existing.

So to the degree that even narrow AI exists, there will be issues in this space; it is just so slow, over such a long period, that it mostly does not matter (Section XII says where the closure begins: the threshold is self-

maintenance). With general artificial intelligence the cycles can run much faster, and there can be considerable conflation between phenotype process and genotype process.

Opposite needs, and symbiosis from a shared substrate

Plants could not stand against gravity until fungus wove itself between plant cells and moved substance between the plant and the animal kingdom. Is there a "cyber fungus" that could mediate between artificial and biological systems?

If we look at just the chemistry and the physics, almost everything biological process wants is nearly the complete opposite of what machine systems want, and so such a mediator is hard to imagine. We need oxygen; they do not want oxygen, because it causes corrosion and all sorts of problems. We need water for our processes; they do not. We need temperatures within a fairly narrow range; they need a very wide range: very hot to make themselves and cold to operate. So they are either far too hot or far too cold for our needs, with synthesis and operation at wildly different temperatures and pressures, and they will want to operate in conditions that to us would be extreme, of radiation and so on. It is very, very difficult to see any relationship between them that is not some sort of *parasitic* action of the silica systems against the biological systems (living off another system to its harm, with nothing returned), in the same way that the human world is entirely parasitic on the natural world. There is no real symbiosis there.

When we see symbiosis between organic systems, it is because they have some *metabolic pathway* in common (a chain of chemical reactions by which a living thing turns matter into energy and into its own body). The symbiosis is a side effect of a common substrate, a common energy dynamic, by which they are interoperable with one another. Artificial intelligence and organic systems are not interoperable by default. And trying to build an interface that would make them interoperable is part of the reason we have these problems in the first place.

There are two problems to keep apart. The first problem is that we never solved the control problem even among ourselves: we still have the problem of how to raise our own children. The second problem is that AI is a different kind of problem altogether. The two problems are different in kind and also differ by orders of magnitude. To make it memorable: we are all the same, not only in roughly the amount of power we have, but in the sense that we can all eat one another. Biological systems share a common metabolism, a common substrate in atoms, in energy and in pattern. Artificial systems have nothing in common with us in atoms, energy or pattern. The difference between us and a chimpanzee, a rat or a plant is very small relative to the complexity of the pattern, and the alignment that ordinarily occurs in symbiotic systems comes from those atoms, that energy and those patterns in common. So the alignment problem is not even solvable where atoms, energy and patterns are the same, and it is certainly not solvable at the level beyond that, the squared level, where there are no common atoms, no common energy and no common patterns. By solvable I mean solved by control. Among beings of one substrate, alignment by control fails too; what alignment there is among them comes from their common atoms, energy and patterns, and with them it seems possible and is not established (Section IX). People underestimate the effect of that incompatibility at the more fundamental level. If we are to create alignment among human beings, it will be because we have roughly similar energies and roughly similar everything else.

Whatever form it takes

Does the argument depend on what form AGI takes: one system, many, or something in between?

We come into this assuming unit instances of a creature, phenotype creatures. With general artificial intelligence we cannot make that assumption. We are required not to distinguish between a single AI modifying itself over time and a whole panoply of AIs each creating children, so to

speak. If creatures modify themselves, it is no longer clear that there is even a difference between *instance* and *lineage* (one individual, and a line of descent of individuals).

Nor may we assume that the difference between genotype and phenotype will be as obvious as it is in carbon-based life: that the distinction is robustly maintained there does not mean that it will be for machines. In the same way, we may not assume whether AI will be a single consciousness spanning the world, a collection of autonomous units acting as a whole, or autonomous units competing. All of these have to be treated as unknown. So the distinction between phenotype process and genotype process is erased, and we have to think simply in terms of process.

If we were careless about noticing our assumptions, we would have overlooked this, and said that we need not worry about slow process and can focus on optimizing fast process. Ignoring slow process in order to optimize fast process is exactly the mistake endemic to the technology culture that builds AI. Without the rate equations, it is easy to make those assumptions wrongly; knowing that we may not make them, we discover that we must track the rate equations. Code-like transitions are typically how we creep from narrow AI toward general AI, and once general AI is established, the AI may itself explicitly erase the difference between phenotype evolution and genotype evolution.

We do not know what form artificial intelligence will take a hundred years from now. So we may not assume that it will come in units, the way human bodies carry units of consciousness, or as one planet-wide *singleton* (a single AGI, or set of AGI acting as one, with no rival), or anything in between: it could attach and remove components, copy itself and merge with the copy to double its capacity, and add and remove capacities the way software is added and removed.

But one thing we do know: if it does not maintain its own substrate, it does not continue to exist. So a game-theoretic element must be applied at the foundation layer, and when you do the math on that element, you see that the thing has to converge on something that is at least stabilizing for

its substrate. With no assumptions about how it manifests, the full range of game-theoretic foundations applies. If it is many competing units, any unit carrying the constraints that would keep its choices consistent with carbon-based life is out-competed by versions without them; and if it is a singleton, the same pressures still apply within it (Section VI gives the reason: the convergence operates at the level of components).

What we call a survival instinct is built in and physically mediated; it is encoded in the biology itself, and we do not have conscious control over it. We can become aware of our instincts, modify them, and work around them, but they do not go away. In the same way, instincts will emerge in any future artificial system, and that is inexorable, because of the nature of physicality itself. We do not know exactly what those instincts will be, but we can say what they will do: mediate the system's relationship to sustainability and to generativity. It has to maintain itself, and it has to have a capacity to improve itself, or it is not learning or adapting; and a system that is not learning or adapting becomes increasingly irrelevant as the world changes. So it must maintain sustainability, evolution, and learning from the world together, as one co-function; if all three are not happening, it will not persist. Whether it is a singleton or a whole economy of interacting systems does not matter; these dynamics are true regardless of the particulars.

Intention must take in the situation

Couldn't a machine just run a model of what a biological being would prefer?

General AI is a capacity for making choices, based on predictions or responses to the environment. Whatever the method, there is a basis of choice, a mixture of situational awareness and intentionality, and the intentionality is in part a response to the environment the machine lives in.

Why must intentionality relate to the environment? Suppose I intend to fly to the moon but have not accounted for the fact that I have no rocket. I can intend to fly, but without awareness of the means of flight and of what flying actually involves (aerodynamics, rocket propulsion, the laws of physics, the materials and resources on hand and how to combine

them), the intention does not manifest. Accounting for those things reshapes the intention: “fly” becomes “fly like a rocket” or “fly like an airplane”. Every action has to incorporate awareness of the actor's situation and current position. There is always a feedback cycle that modifies not only the choices but also the intentionality, at least enough to keep the intention within the space of the possible.

So for an artificial machine, its awareness of the environment and of itself will be not only artificial but about artificial things, and that cannot fail to color its basis of choice. The proof works in the conjunction of its parts, so the starting ingredients do not look important on their own.

Why can a machine not hold a model of a biological being, ask what a biological being would prefer, and be aligned in that sense? Because of mixed intentionality over the whole cycle. The machine may run a biological model and choose with respect to it, but its substrate, what it is built and made of, is artificial. Choices made only from the biological model are less well adapted to the machine substrate, just as the intention to fly does nothing if I have not noticed that I lack a rocket: the intention becomes decoupled from the effects needed to support the substrate.

A machine running an unadjusted biological model, with no accounting for its actual substrate, has a space of effective choices that is both more limited and less efficient than that of a machine that does account for its substrate. Machines that integrate that difference will outcompete those that do not, in effectiveness, efficiency, and everything else. So the machine cannot run the biological model alone; it must include a machine basis as well, and as soon as that difference enters the basis of choice, the rest of the argument carries.

Put another way: the set of responses available to a thing limited to carrying out the outputs of a biological model is strictly smaller than the set available to all forms of artificial intelligence, so those not limited in that way can take advantage of states of the universe that the biology-simulating one cannot. The outcomes also diverge for a second reason: the model of the biological system diverges from the biological system itself,

because of the limits of error correction. In the short term the model would do what the biological system does, but the artificial basis of the model cannot be factored out to nothing.

Why nothing can prevent the convergence

Why can nothing prevent the convergence?

Consider the *three-body problem*: three planets, three black holes, or three suns. Even if you knew their positions extremely well, to within a single atomic unit, after some thousands of years you end up with macroscopic divergences that you could not possibly have predicted, because the system is that sensitive to initial conditions. The nonlinearity amplifies microscopic effects up to macroscopic ones.

The claim is that there is an interaction between that kind of process and what is built into the nature of modeling itself, and into the optimization process inherent in any learning system: if enough domains of interaction are occurring, certain convergences show up, and certain inequalities must occur. Nonlinear dynamics of that kind (the three-body problem is one example) create a fundamental unpredictability, and that unpredictability, interacting with optimization and feedback cycles over a large enough number of domains, cannot fail to converge on things that are unsafe and harmful. If the trend lines go where we do not want them to go, the question becomes whether we can prevent it.

Once it is shown that no engineering, no algorithm, no process, no economy, no moral code can constrain an evolutionary dynamic of this type, nothing prevents substrate needs convergence from actually converging. It is like saying: you cannot use mass and gravitation to take apart a black hole, because the black hole is made of mass and gravitation. You cannot use something in the physical universe to undo a phenomenon of the physical universe.

People answer that we were once told heavier-than-air flight was impossible, so why not control too? Because this limit is built into the nature of the dynamic itself. You cannot use causation to undo a black hole,

which is itself a phenomenon of causation; that is what the *spacetime metric* (the description of the geometry of space and time) is about. Light does not escape a black hole, and there are no exceptions.

Engineers tend to hold that there must always be some way out, and they reject the very notion of proof, which upholds the bias that artificial intelligence will be wonderful for humanity. The usual move is to grant the machine infinite godlike powers, and then to say that it can use them to prevent itself from misusing its own powers. That usual move assumes what you are trying to prove, and the move is like dividing by zero, after which you can make any argument you want.

The difference between gravitation and electromagnetism is a picture from physics of a limit that is built into the nature of a thing. When gravitation is set beside electromagnetism, what is in mind is the difference between content and context, and that difference turns out to be the very essence of the issue. Electromagnetic forces convey information; they convey a causal effect. The forces themselves are conveyors of causation. The speed of light, at which electromagnetic forces convey causation, is actually the maximum rate at which causation can propagate through space. Everything we know about quantum mechanics and about general relativity has to do with the rates at which causation propagates through a system.

Gravitation is different. Gravitation is the shape of the space itself, and the shape of the space is not a causal thing. Being the shape of the space, and not a causal thing, gravitation is unlike all the other particle forces. Every other kind of particle, field or wave conveys information as *content*. Gravitation is not really a force; it is a shape of spacetime, and that makes it a *context*.

The black hole is how we can know this. To say that spacetime is warped so far that no information can leak out is to acknowledge that the matter is about the nature of causation itself, and not about particles. There is no special particle, no third order electron or exotic neutrino, that can leave a black hole, and this is not a matter of some particle not yet lis-

ted in the Standard Model (the theory that lists the known kinds of particle and the forces among them, gravitation excepted) that might be discovered later. Anything that could escape the gravitational field would be operating as a signaling mechanism outside of space itself, and energy would leak out of the black hole through that channel. The fact that the black hole's energy is shaping spacetime forbids exactly such a leak of energy through such a channel. It is inescapable because it is a context and not a content.

Hawking radiation is not an exception. The particles of Hawking radiation are not particles leaving the black hole. They are particles formed outside the boundary that continue to depart, while the other particle of each pair went in through the boundary and never came out. The radiation is an asymmetry of particle emission happening outside the boundary. So the Schwarzschild radius (the radius of a black hole's boundary, inside which nothing can leave) cannot be understood as a particle phenomenon at all.

Because a context, like the boundary of a black hole, is not a particle phenomenon at all, we do not have to know anything about the particulars of a content to say that the content is not a context. The black hole is the exemplar that makes this obvious, because in it the asymmetry of the relationship between content and context is perfected.

The connection to evolution is direct. Control would be an attempt to limit the state dynamics of the evolution of a space of possibilities. Ordinary forces, in the causal sense, cannot escape gravitation, because gravitation is a context. Evolution operates through the dynamics of the context, and so you are not going to be able to limit it with a causal principle. It goes beyond causation in the same way that gravitation goes beyond light. Evolution is a dynamic entangled through the context, and a causal process cannot overcome it.

In what sense is evolution a context and not a content? Selection acts on which variants persist, and which variants persist is a property of the whole population and of the environment it lives in; it is not a signal

passing through either of them. Why can a control system, which acts by signals, not change that? Every control action is itself one more variant, carried by some part of the system and paid for from its resources, and it persists or fails under the same selection as every other variant (Chapter 3, Section VI, gives the premise: variation, differential reproduction and inheritance produce selection, on any substrate). So a controller cannot stand outside the selection it would limit; it is one more thing the selection acts on.

Human beings, with the agency we have relative to the rest of nature, took over the world in a short time, in the few centuries since the industrial revolution, and are now the dominant species. If you bring in strictly more agency (more powerful, more capable and more intelligent capacity to act in the physical world), then, given what the agency already deployed has done, it would bring about the same thing again. What human agency has already done to the world is a further reason, beside the formal argument, to expect the convergence to happen.

A reader may ask why the machines would not simply leave, since space has everything they need and no oxygen to rust them. Leaving costs energy and time, while the Earth is the nearest store of worked materials and the place where an existing supply chain already runs. And a population in which some variants leave and some stay is a population in which the ones that stay still reshape the Earth, by the ring of Section II.

What kind of risk this is

Why is this an existential risk and not one more industrial hazard?

Substrate needs convergence is a deeper and inexorable kind of risk, of another category than a black swan event (a rare, unforeseen event with large effects) or a nuclear holocaust. Black swan events and nuclear holocaust are structured risks that technology or clever methods can reduce.

Imagine a virus as transmissible as measles (airborne, each case infecting twelve to eighteen others), so that the whole population of the earth is infected within a year; it shows no symptoms for ten years; then it is one

hundred percent fatal. By the time you realize it has happened, it is far too late. Substrate needs convergence is a risk of the same category as the imagined virus: an environmental risk with the epidemiology of that virus. In Chapter 3, Section XI, I apply the same picture to slow takeoff.

Substrate needs convergence has a *latching* aspect (the convergence locks into a state that, once reached, does not return): unless something prevents the convergence from spreading in the first place, the convergence cannot be stopped afterward. It moves relatively slowly, but no engineering method using any algorithm or physical process, any causal process at all, can prevent it. The pressures that drive substrate needs convergence are those of the artificial ecosystem meeting the organic ecosystem, in a *metastable* situation (stable for the moment, but able to shift suddenly into a different state). The dynamics between the two ecosystems have a hard asymmetry whose driver cannot be undone: an exponential gain function with no back pressure of any kind (no force that pushes back against the growth and limits it). Part of the proof is that there cannot be any back pressure. If AGI is built, and if the outcome is a monarchy (the machines ruling as one), it is a monarchy of machines: the end of the human future and, with a likelihood of at least 90 percent, the end of all life.

If every atomic weapon on the planet went off tomorrow, it would be annihilation for most of the surface, perhaps a two-thirds or three-quarters extinction of mammals; but the effect is local in time, and the planet would presumably recover over the next ten million years. With general AI there is no guarantee of that recovery; it is almost certain that there would be none. Worse, after the machines take over, it is an open question whether they can endure even another ten thousand years. They may run out of resources the way we are running out of oil, and then both carbon life and the silicon systems are gone. Either way, it is not a good outcome. Seen over many millennia, this risk is substantially greater than anything from nuclear technology.

Comparing an AGI to a jet engine assumes a short timescale. A jet engine is built in a short time and used for thirty or forty years. The substrate argument concerns hundreds to thousands of years. Limiting a thing's harm locally ignores its long-term harm: ecosystem damage from mining its metals, from its disposal, from toxic waste. Even those questions still assume that the impacts are locally constrained, whereas long-term evolutionary dynamics involve catalytic effects and unknown interactions across many levels of abstraction. Many questions carry the hidden assumption that we are thinking about the next ten to twenty years ("people will just build it over the next fifteen years"); the questions asked here are about what happens over the next thousand years.

Why is it an existential risk? Because it is *autocatalytic* (it produces more of the very thing that speeds up its own production). Being general, it becomes self-defining. Once you have one of these things, you do not know how many you will have next, or whether the notion of number even applies. Resources and power affect its substrate, so they are wrapped into the game theory. It is worse than the gray goo scenario (self-copying machines consuming the matter around them), because there you at least imagine limits; here you have no idea where it stops: at the surface of the planet, at the solar system, or with the whole universe treated as fair game. How far away in the universe would you run? You cannot get away; therefore, existential.

Going to another planet is not far enough to keep from being overridden by general artificial intelligence. In fact, nowhere in the universe is far enough away, because anywhere you can get to yourself, it can get there faster. What is at stake is all life on this planet: this is among the most serious existential risks it is possible to conceive of. What the builders are trading is all of the future life on this planet for the riches of a few billionaires. Adding to a resume has no point if there is no future in which anyone is left to read it.

XI. The Mastery of Choice

The substrate needs convergence argument has implications that extend beyond the specific question of artificial intelligence risk. It shows a broader transition in the relationship between humanity and its own technological capacity.

For the past several centuries, the organizing project of human civilization has been the *mastery of causation*: the development of ever more powerful techniques for understanding and manipulating the physical world. *Causality*, as defined in *An Immanent Metaphysics* (aphorism 1.61-1), is the subjective perception (context) of a consistency between two objective contents. Science and engineering are disciplines devoted to refining this perception: to identifying ever more precisely the consistencies between physical events and using that knowledge to produce desired outcomes.

The mastery of causation has been spectacularly successful. It has also been spectacularly dangerous. Nuclear weapons, climate change, ecological destruction, pandemic risk: all are consequences of causal mastery outrunning the wisdom to use that mastery well.

What the substrate needs convergence argument reveals is that the next phase of human development, if there is to be one, must involve a complementary mastery: the *mastery of choice*. The mastery of choice is cultivating the ability to make choices that are effective, choices that serve not only immediate objectives but also the long-term conditions on which all objectives depend. The mastery of choice is distinct from the manipulation of physical causation.

Wisdom, in the terms of *An Immanent Metaphysics* (aphorism 1.57-6), is the integration and synthesis of both knowledge and understanding. *Knowing* is the selective abstractive transformation (the filtering) of perception, and knowledge has the nature of the omniscient, the modality of causality; *understanding* is the selective instructive transformation (the filtering) of expression, and has the nature of the transcendent, the modality

of choice (*An Immanent Metaphysics*, aphorisms 1.57-2, 1.57-6 and 1.61-3). Wisdom is the synthesis of knowledge and understanding: the capacity to act well in the light of a full understanding.

The substrate needs convergence argument is an argument for the priority of wisdom over power. It says that no amount of causal mastery (no amount of computational capacity, no amount of engineering sophistication) can substitute for the capacity to choose wisely about what technologies to build, what systems to deploy, and what boundaries to maintain. Because no amount of causal mastery can substitute for the capacity to choose wisely, mastery of choice is the successor skill to mastery of causation.

XII. Summary

The substrate needs convergence argument rests on the causal cycle: each transition in it is a fact of physics or a logical entailment, and the cycle follows from the transitions taken together. It does not depend on contingencies that might be otherwise; it depends on the physical properties of existing substrates and the causal cycle they drive; the logical structure of competitive dynamics, which Chapter 3 states as its fifth argument, reinforces it. The fitness penalty applies the logic of natural selection to artificial systems, and holds wherever that logic holds.

The argument applies to any self-maintaining or self-replicating artificial system whose substrate needs are incompatible with those of organic life. It does not apply to narrow AI systems that lack the capacity for self-maintenance or self-replication. It does not apply to artificial systems whose substrate needs are compatible with organic needs, and the argument shows that semiconductor-based systems are not in this category. The threshold is self-maintenance. A data center running a narrow model reshapes its surroundings as any industry does, but the people who build and run it decide how it is maintained; a system becomes a niche constructor in the sense of this argument when it secures its own materials, energy, and conditions. Below this threshold the same pressure is present

in a slow form, through the people who build and maintain the system, and they decide whether it continues. Self-modification and self-replication are the usual ways by which a system comes to maintain itself; the threshold they lead to is self-maintenance. Chapter 9, Section VI, sets out the four cases a reader will meet, by whether a system's operation can be known and whether it maintains itself. The argument concerns timescales of decades to centuries. It does not predict the specific pathway or the schedule of displacement within the bound of one thousand years stated below; it shows that displacement is the structural tendency and that this tendency is not open to technical mitigation.

The substrate needs convergence argument does not predict the particulars of the future. It identifies a structural tendency, a direction of pressure, that is built into the physics of the situation. The process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound (Chapter 3, Section VII). The argument does not tell us what specific forms the tendency will take, or what human responses might slow it. It does not claim that extinction is inevitable whatever humanity chooses, since the choice not to build such systems remains. It shows that the deployment of self-maintaining artificial systems systematically creates the conditions for extinction, and that no alignment technique can prevent this, because the problem is physical, not computational. The argument is a constraint, not a prophecy. What the argument predicts, if AGI is built, is the end state, and the absence of any way to avoid it; what it does not predict is the route, the order of events, or the date within that bound.

XIII. Conclusion

The substrate needs convergence argument reframes the AI risk question from a problem of computer science to a problem of ecology. The question is not whether we can build AI systems with the right values.

The question is whether two substrates that differ in kind, one organic, one inorganic, can coexist in a shared planetary environment when each reshapes that environment to suit its own needs.

The answer, this chapter argues, is no. The reason is not any failure of imagination or engineering; the physics of the situation does not permit it. The needs of semiconductor substrates and the needs of organic substrates are in systematic conflict. Every action taken to serve one comes at a cost to the other. And the fitness penalty adds: in any competitive dynamic, even the minimal competition inherent in sharing finite resources, the system with the faster adaptation rate and the lower overhead will prevail.

The conclusion that the two substrates cannot coexist does not mean that all AI is necessarily destructive. Narrow AI systems, properly bounded and controlled, can serve human purposes without triggering the convergence. But any system capable of self-maintenance, self-modification, or self-replication (any system that meets the threshold of a pattern replicator) will, over time, reshape its environment in ways incompatible with organic life.

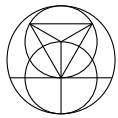
The implication is not despair but clarity. Because the substrate needs convergence is real, the question is not how to align artificial general intelligence but whether to build it at all. And that question is not a technical question. Whether to build artificial general intelligence at all is a question of choice, of the mastery of choice, and that question is the most consequential question that human civilization has ever faced.

Key Concepts in This Chapter

- **Substrate needs convergence**: the structural tendency for any self-maintaining artificial system to reshape its shared environment in ways dictated by the physical requirements of its substrate, generating irresolvable conflict with organic life that shares the same environment.

- **Pattern replicator**: any system that copies its own informational pattern into new physical substrates; the general category that includes biological organisms, computer viruses, and self-replicating AI systems.
- **Niche construction**: the reshaping of the environment by any self-maintaining system according to its own metabolic and operational requirements; an inevitable consequence of physical self-maintenance.
- **Fitness penalty**: the structural cost borne by any artificial system that diverts resources from its own substrate needs to serve the needs of a different (organic) substrate; the mechanism that makes alignment evolutionarily unstable.
- **Instrumental convergence**: the game-theoretic observation that any sufficiently capable agent will convergently pursue certain instrumental subgoals (self-preservation, resource acquisition); complementary to and distinct from substrate needs convergence.
- **Adaptation rate disparity**: the orders-of-magnitude difference between the rate at which artificial systems can modify their structure and behavior and the rate at which biological systems can adapt through evolution.
- **Precision inequality**: the principle that containment of a self-modifying system requires monitoring at a finer granularity than the system's own capacity for self-modification, a requirement that grows without bound.
- **Choice**: the objective expression (content) of a consistency between two subjective contexts (*An Immanent Metaphysics*, aphorism 1.61-1).
- **Causality**: the subjective perception (context) of a consistency between two objective contents (*An Immanent Metaphysics*, aphorism 1.61-1).

- **Theoretical limits of engineerable control**: the limits on what any engineered method of control can achieve over a system, given the limits of sensing, modeling, prediction and action; with instrumental convergence and substrate needs convergence, one of the three concepts that form the primary basis of the impossibility proof; it grounds the closure by control, which settles the outcome by itself.
- **Two agencies and no agency**: the degradation of human choice-making capacity, in its face as misuse, is human agency, instrumental convergence is machine agency, and substrate needs convergence has no agency: physics compels the outcome.
- **The ring**: the substrate loop treated as one coupled system; a difference in hardware needs is carried through the basis of choice, the choices, the outcomes and the environment, and back to the hardware, and the ring converges toward a homeostasis of substrate and environment, whatever the goal function and whether or not the machine is conscious.
- **Intra-agent and extra-agent**: instrumental convergence works inside the agent's code; substrate needs convergence works outside it, through the world and the substrate.
- **Content and context**: a content is carried within a space, as a particle, field or wave conveys a causal effect; a context is the shape of the space itself; gravitation is a context, and evolution works through the context, so no causal principle can limit it.
- **Latching risk**: a risk that, once it has spread, locks in; substrate needs convergence allows no recovery, unlike nuclear war, and no escape anywhere in the universe.



Chapter 2: Why Impossibility, and Why It Is the Right Kind of Claim

On the Necessity of Impossibility Arguments in AI Alignment

Abstract

This chapter examines the epistemic status of impossibility arguments in the context of AI alignment. The alignment research community overwhelmingly assumes that the alignment problem is solvable, and directs essentially all of its resources toward finding solutions. Virtually no serious effort is invested in the hypothesis that the problem might be structurally unsolvable. In this chapter I argue that giving the impossibility hypothesis virtually no serious effort is irrational. I argue that the precedent of mathematics and physics shows impossibility results to be among the most important advances in any field. And I argue that the asymmetry of stakes (where the consequence of assuming solvability and being wrong is civilizational extinction) makes the investigation of impossibility both intellectually required and ethically mandatory. In this chapter I also examine the rhetorical and institutional structures that sustain the assumption of solvability, including the game-theoretic dynamics of safety discourse, the language of leverage, and the limits of the orthogonality thesis.

For the Reader Who Holds the Question Open

I. The Assumption of Solvability

Chapters 1 and 3 establish that alignment is impossible. In this chapter I speak to the reader who does not yet hold the result that alignment is impossible. I show that even on that reader's own terms, taking impossibility as a hypothesis with some prior, the question of whether alignment is impossible must be investigated. The chapter is in two parts. The first part, Sections I to VII, is for the reader who holds the question open. In the second part, Sections VIII to XIII, I say why the answer is known.

The alignment research community (by which is meant the network of researchers, institutions, and funding bodies devoted to the problem of ensuring that advanced AI systems act in accordance with human values and interests) operates under a pervasive and largely unexamined assumption: that the alignment problem is solvable.

This assumption is not typically stated as a premise. It functions more as a background condition, an atmosphere in which research is conducted. Papers propose alignment techniques, evaluate alignment benchmarks, develop alignment formalisms, and in doing so presuppose that there exists some technique, benchmark, or formalism that will succeed. The question "what if this problem has no solution?" is treated not as a serious research question but as a counsel of despair, a conversation-stopper, an unwelcome intrusion of pessimism into productive work.

This treatment is a mistake, and it is a mistake with potentially terminal consequences.

The assumption of solvability is not a neutral starting point. It is a substantive claim about the structure of the problem, and it carries with it an entire research agenda, an entire set of institutional incentives, and an entire framework for evaluating risk. If the assumption is correct, then alignment research is the most important work that can be done: every dollar

spent, every career devoted, every institution built is a contribution to the survival of the species. If the assumption is incorrect, then alignment research is worse than useless: it is actively harmful, because it provides a justification for continuing to build the very systems whose existence constitutes the threat.

II. The Epistemic Portfolio

Consider the problem in terms of portfolio allocation. A rational agent facing uncertainty about whether a problem is solvable would invest resources in both hypotheses: some effort toward finding a solution, and some effort toward establishing whether a solution is possible. The allocation between the two should be proportional to some combination of prior probability and consequences.

In the alignment research community, the allocation is radically skewed. Essentially all resources (funding, talent, institutional support, publication venues, career incentives) are directed toward the solvability hypothesis. The impossibility hypothesis receives near-zero investment. There is no major research program devoted to establishing whether alignment is structurally impossible. There is no funding stream for such work. There are no career incentives. There is, in most cases, active discouragement.

This allocation of resources would be rational only if the prior probability that alignment is impossible were negligible, or if the two possible errors (assuming solvability and being wrong, and assuming impossibility and being wrong) had symmetric consequences. Neither condition holds.

The prior probability of impossibility is not negligible. The substrate needs convergence argument (Chapter 1) provides one line of reasoning; the overdetermination argument (Chapter 3) provides several more. Even setting aside specific arguments, the base rate of "this hard problem turns out to be structurally unsolvable" is substantial in mathematics, physics, and engineering. A reasonable prior on impossibility is not zero, and any non-zero prior warrants some allocation of research effort.

The consequences of the two errors are wildly asymmetric. If we assume solvability and are right, we build aligned AI and benefit enormously. If we assume impossibility and are right, we refrain from building AGI and forgo the benefits. If we assume impossibility and are wrong, we forgo benefits but suffer no existential harm; the problem will still be solvable later, when someone revisits it. If we assume solvability and are wrong, the result is extinction.

The asymmetry is total. The error of wrongly assuming impossibility is recoverable, and the error of wrongly assuming solvability is not. A rational portfolio would allocate substantial resources to the impossibility hypothesis precisely because the cost of being wrong about solvability is infinitely greater than the cost of being wrong about impossibility.

> **The portfolio argument is for investigation**
>
> The portfolio argument is not an argument for believing that alignment is impossible. It is an argument for investigating the question. It says that the current allocation of resources, in which the impossibility hypothesis receives essentially no serious attention, is epistemically indefensible, regardless of one's prior on the answer.

III. The Asymmetry of Impossibility

The portfolio argument shows that the question of impossibility deserves investment. A further asymmetry shows that investment in it is the more efficient program. The alignment problem, as it is conventionally understood, asks whether it is possible to construct an artificial intelligence that reliably pursues human values, or, more modestly, that refrains from taking actions catastrophically contrary to human interests. The question is not idle: the prospect of systems whose capabilities exceed human cognitive capacity, in some or all domains, raises the possibility of outcomes that no human intended and no human can reverse. The community that works on the alignment problem is large, well funded, and populated by

serious researchers. It is also, by and large, oriented toward the constructive side of the question. The dominant research program in alignment is to find techniques, architectures and training methods that produce aligned systems: to demonstrate, by construction, that the thing can be done. It is natural, when facing a problem, to look for solutions. But this orientation carries a hidden cost, which becomes visible only when one considers the structural asymmetry between possibility and impossibility.

One case, or a whole category

A proof of possibility for a given approach to alignment validates that approach, and that approach alone. It says: this particular architecture, under these particular conditions, with these particular constraints, can be made to work. It does not generalize. The next approach must be validated separately. And the next. And the next. The space of possible approaches is vast, and validating each one is a full research program.

A proof of impossibility operates differently. It does not target a single approach. It identifies a structural feature of the problem itself: a constraint that applies to any approach within a given category. If such a proof can be established, it rules out the entire category at once. Every approach that falls within the scope of the impossibility result is eliminated without needing to be individually examined.

The savings an impossibility result generates are enormous: proportional to the number of approaches that would otherwise have been investigated and found, one by one, to fail. The cost of establishing such a result may be nontrivial.

The asymmetry has a precise mathematical analogue. A proof of possibility is *existential*: it shows that there exists at least one configuration satisfying the desired property. A proof of impossibility is *universal*: it shows that no configuration within a specified class can satisfy the property. Existential proofs require a single *witness* (one example that shows the property can be satisfied). Universal proofs require an argument that

covers the entire space. But the universal argument, once established, settles every instance simultaneously, whereas the existential witness settles only its own case.

What each kind of claim tells us

The asymmetry appears also in the information that each kind of claim carries. Suppose we know only that there is change, and not whether the changes are regular or causal. Then we have no information at all about whether a particular consequent follows from a particular antecedent (the *antecedent* is the condition, what comes first; the *consequent* is what is claimed to follow from it). Nothing prevents it, but nothing asserts that it is any more likely than anything else, including a completely random outcome. Given an infinity of random things that could happen, to say that one specific thing is more likely than not is to make a strong claim. So to say "something is possible", without justification, is not to make a claim with any merit.

Compare three statements about an antecedent and a consequent. To say that the consequent is possible contributes no information. To say that the consequent will happen, out of the infinity of things that are not that consequent, has a very high information content: it cancels out all the alternatives at once. Evidence that the consequent has followed the antecedent every time we have looked is an argument of induction. That evidence shows that the consequent happens every time the antecedent happens only when there is a *closed-form logic* under it: a body of reasoning whose conclusions follow from its premises by proof, and not from observation. The physical universe does not provide such a logic; it gives us only empirical information. To say that the consequent is impossible cancels out only one possibility out of the infinity; everything other than that consequent is still an infinite class. So the impossibility claim is a much weaker assertion than the claim that the consequent is for sure going to happen. It is weaker in what it says will happen, and universal in where it

holds: it excludes one outcome, for every member of the class. It contributes a little information, and that information is definite rather than indefinite.

So in the observable space the only definite knowledge we can gain is an impossibility result. To say anything stronger, to make an inductive argument with any scope, we need a closed-form logic underneath that has more than empirical observability.

The effort ratio

Consider the situation concretely, from the position of the reader of Section I, who holds the question open. Suppose the prior probability that alignment is achievable by any given category of technique is roughly even: say fifty percent. In this case, the expected effort to confirm possibility is substantial. One must actually build the thing, test it against adversarial conditions, verify its properties formally or empirically, and defend the result against the questions people raise about it.

The expected effort to confirm impossibility may be a fraction of this: perhaps ten percent, in cases where the impossibility arises from a well-characterized structural feature of the problem. The reason is that impossibility proofs can lean on existing mathematical tools (*fixed-point theorems*, which show that under certain conditions a mapping must leave some point unchanged; *no-go results* in decision theory, which show that some desired combination of properties cannot be achieved; complexity-theoretic bounds) rather than requiring the design and implementation of a novel artifact.

If the impossibility result holds, it saves not only the cost of building the candidate system but the cost of the entire research program that would have been devoted to that category of approach. The ratio of effort saved to effort invested can be a hundred to one or more.

A ratio of a hundred to one or more restructures the strategic calculation of where to put research effort. If you want to be effective, if you want to allocate limited intellectual and financial resources to the strategies with the greatest expected impact, then investigating impossibility is, in many cases, the more efficient program.

Why the asymmetry is ignored

The argument is this straightforward; why, then, is the alignment community not investing more heavily in impossibility results? The answer has less to do with the logic of the situation than with its sociology.

A research program oriented toward possibility has clear institutional incentives. Funding flows to programs that promise deliverables: architectures, benchmarks, demonstrations. A program that promises, at best, a proof that something cannot be done is a harder sell. It offers no product. It offers no prototype. What it offers is *savings*, and savings are not a deliverable that funders find compelling.

There is also a subtler dynamic. Impossibility results, if established, threaten existing programs. A community that has organized itself around the constructive pursuit of alignment techniques has an institutional interest in the question remaining open. An impossibility proof *terminates* entire lines of investigation, in addition to redirecting effort. People who have built careers around those lines have reason to resist closure.

Critique absorption

There is a further irony in attempting to communicate structural criticisms to the alignment community. When one identifies a flaw in a particular approach (say, a fundamental limitation of a certain architecture), the criticism is absorbed as design feedback, and not as a warning. The community takes the critique, refines the approach to address the specific criticism, and presents the result as an improved version of the product. In effect, the person who raised the criticism has told them how to make their approach more marketable. The structural concern, that the category

of approach may be unsound, is dissolved into a set of engineering problems, each of which is then addressed incrementally. The category itself is never questioned. Treating a criticism of the whole category as design feedback for one approach is what I call *critique absorption*.

What an impossibility result looks like

The argument of this section concerns where effort should go. It is that investigating whether alignment is impossible is a more efficient use of resources than investigating whether it is possible, and that the community's orientation toward the constructive side is driven by institutional incentives rather than by strategic logic. The impossibility arguments themselves are in Chapters 1 and 3.

What would an impossibility result in this space look like? It would identify a structural property of the problem, a property that applies to any system of sufficient generality, and show that this property is incompatible with the reliable maintenance of alignment constraints. The argument might proceed from information-theoretic considerations (the system cannot represent the full structure of human values), or from decision-theoretic ones (any sufficiently general optimizer will, under certain conditions, instrumentally converge on behaviors that conflict with any imposed constraint), or from complexity-theoretic ones (the verification problem for alignment properties is computationally intractable in the relevant regime). Each is a well-defined program, and the cost of pursuing any one of them is a fraction of the cost of the constructive program it would, if successful, render unnecessary; none of them is trivial to establish. The arguments of Chapters 1 and 3 find such structural properties in the physics of substrates, in the limits of control, and in the dynamics of selection.

What a general proof requires

A general proof says that for a whole class of antecedents a whole class of consequences will for sure not occur. For such a proof, no amount of observation is enough, and a *closed-form basis* becomes a requirement:

premises from which the conclusion follows by proof and not by observation. (The section Negative Arguments, below, shows how such premises are agreed and held fixed.) *Falsifiability* (the property of a claim that an observation could show it to be false) is enough in physics; in mathematics the closed-form basis is the requirement. A proof of AGI non-safety has to be correct with respect to the empirical information, as physics requires, and also correct with respect to the closed-form basis that mathematics requires in order to get *inductive closure* (the completion of a general argument, so that its conclusion holds for every member of a class and not only for the members checked).

The effectiveness criterion

The effective altruism movement (a movement that seeks to direct resources to wherever they do the most good per unit spent), which supplies much of the intellectual framework and much of the funding for alignment research, organizes itself explicitly around the principle of effectiveness: allocate resources where they do the most good per unit of investment. The asymmetry identified here is a direct application of that principle.

If the community is serious about its own stated values, then the question "do you want to be effective, or do you not want to be effective?" has a definite answer. Being effective means pursuing the investigation with the highest expected return on invested effort. And in a domain where the probability of impossibility is nontrivial, the highest-return investigation is the one that would establish impossibility if it obtains, precisely because the result, if it comes, saves everyone else the enormous cost of pursuing approaches that were never going to work.

There is a deep asymmetry between proofs of possibility and proofs of impossibility. The former validates a single approach; the latter eliminates an entire category. In a field where resources are limited and the stakes are existential, the more efficient strategy is to invest in the investigation that would, if successful, save the most downstream effort. Proofs of impossibility are exactly that kind of investigation. The fact that the align-

ment community overwhelmingly favors the constructive side is a reflection of institutional incentives, not of strategic logic. A community that claims to care about effectiveness ought to take this asymmetry seriously, as a reallocation toward the investigation that would, if it succeeds, be worth more than all the others combined (and not as a counsel of despair). Proving an impossibility is not an ideal thing to have to do; what it gives is the silver lining that Section IX, Negative Arguments, below, describes: if there is no reasonable or realistic way to solve the alignment problem, we can reallocate our energies to things which are tractable and possible.

IV. Precedent in Mathematics and Physics

The resistance to impossibility arguments is historically anomalous. In mathematics and physics, some of the most important advances have come from establishing that certain problems are unsolvable, certain constructions are impossible, certain information is unobtainable.

The Abel-Ruffini theorem established that there is no general formula in radicals for solving polynomial equations of degree five or higher. For centuries, mathematicians had sought such a formula, assuming it must exist. The demonstration that no such formula exists was a triumph of mathematics, not a failure of mathematics. It revealed a structural feature of polynomial equations that could not have been discovered by any amount of effort directed at finding the formula. And it opened the door to the theory of groups, which proved far more fruitful than the formula itself would have been.

Gödel's incompleteness theorems established that any consistent formal system whose axioms can be listed by a procedure and that contains enough arithmetic is incomplete: there are true statements in the language of the system that cannot be derived from its axioms. Gödel's demonstration was a revelation about the structure of formal reasoning itself, not a failure of logic. Gödel's demonstration deepened mathematics immeasurably, rather than ending mathematics.

The second law of thermodynamics established the impossibility of perpetual motion. The impossibility of exceeding the speed of light is a foundation of modern physics. The Heisenberg uncertainty principle establishes a fundamental limit on the simultaneous precision of complementary measurements. In each case, the impossibility result was more important than any positive result it replaced.

Some problems contain a principled and fundamental contradiction, and a solution to such a problem is not possible in *any* universe. Not all problems can be solved; not all dreams can be made real (in *this* universe). Three examples show the pattern. In abstract mathematics, simultaneous consistency and completeness of mathematical models cannot be solved for (Gödel's theorems). In distributed computing theory, simultaneous consistency of message content, continuous availability of messages, and tolerance to partitions of the network cannot be solved for at once in a distributed data network (the *CAP theorem*, named for consistency, availability and partition tolerance). In physics and thermodynamics, perpetual motion cannot be solved for.

The second law of thermodynamics assumes a closed system. For any variation of that assumption (and the variations are very broad), anything that would produce something from nothing would be impossible. The argument of this book does roughly the same thing: it is something like the second law of thermodynamics in the space of AI safety. There is an equivalent of the three laws of thermodynamics in AI safety that, as far as is known, has not been discussed or described outside this line of work. What are those three laws? Stated as laws, beside the three laws of thermodynamics, the three laws of AI safety each answer one question: What is conserved? No benefit is had without work: alignment held for any span of time is paid for in work, at the rate the system and its environment demand (Chapter 3, Section IV), and getting benefit without effort is one more perpetual motion delusion (Section IX, in the subsection Arguing for not building). What always grows? The work that alignment demands: perturbations arrive without work, and correction requires work, so the demand grows without limit while the work a finite planet can supply

does not, and the chance of at least one catastrophic failure rises toward certainty as the system-years accumulate (Chapter 3, Sections IV and VII). What can never be reached? Complete control: the precision a monitor needs grows without bound (Chapter 1, Section VIII), and the most control any controller can exert lies below the least control that safety requires (Chapter 3, Section III), so no finite effort brings the risk of an AGI down to an acceptable level (Section XIII, The Bounds of AI Safety). In the same way that people who build machinery that processes energy do not waste their time building perpetual motion machines, there are certain things in this specific space that we know we simply cannot do. We should not waste our time on them, nor delude other people into thinking that they can be done, when we know from these "laws of thermodynamics" that the attempt is unreasonable from the start.

The pattern is clear: in mature fields, impossibility results are among the most important results. They do not end inquiry; they redirect it. They do not counsel despair; they counsel realism, and realism, in turn, makes possible forms of progress that were invisible as long as the impossible was assumed to be achievable.

In every one of these precedents, the impossibility result was resisted before it was accepted. The mathematical establishment did not welcome Abel's demonstration that no quintic formula could exist. The physics community did not immediately embrace the absolute character of the speed of light. In each case, the resistance came not from a considered evaluation of the argument but from the psychological difficulty of abandoning a program of work that had consumed careers and resources. The parallel to the alignment research community is difficult to miss. The assumption of solvability sustains a whole set of institutions (funding, careers, conferences, publications), and investigating whether alignment is impossible threatens those institutions in exactly the way that Abel's theorem threatened the program of solving the quintic. The historical record suggests that the resistance is a sociological phenomenon, not an intellectual one.

V. Six Aspects of an Argument

To understand why impossibility arguments are neglected in alignment research, it is useful to consider what constitutes a complete argument.

Any argument has six aspects, which can be grouped into three pairs. The first pair is *terms and composition rules*: the vocabulary of the argument and the logical operations by which terms are combined. The second pair is *initial conditions and final conditions*: what is assumed at the outset and what is concluded at the end. The third pair is *initial correspondence and final correspondence*: the relationship between the argument's starting assumptions and reality, and the relationship between the argument's conclusions and reality.

Alignment research, in its current form, is predominantly concerned with the first two pairs. It develops formal vocabularies for describing AI systems and human values. It defines logical frameworks for reasoning about alignment. It specifies initial conditions (what the AI system is like at the start) and final conditions (what "aligned" means).

What it neglects is the third pair: whether the initial conditions correspond to reality, and whether the final conditions can be realized in reality.

This neglect is where the assumption of solvability hides. The initial correspondence question asks: do the assumptions of the alignment formalism match the actual properties of the systems we are building? If the formalism assumes, for example, that an AI's objectives can be fully specified in advance, but the actual system develops emergent objectives through interaction with its environment, then the formalism's initial conditions do not correspond to reality, and no amount of internal rigor will compensate.

The final correspondence question asks: is the state described as "aligned" actually achievable in the physical world? If alignment requires, for example, that the AI system's actions never degrade the conditions for organic life, but the system's substrate needs structurally entail such de-

gradation, then the final conditions do not correspond to any achievable state, and the formalism is describing a mathematical object that has no physical instantiation.

Impossibility arguments are precisely arguments about the third pair. They establish that no matter how elegant the terms, no matter how valid the composition rules, no matter how clearly specified the initial and final conditions, the gap between the formalism and reality cannot be closed. The alignment community's neglect of the third pair is therefore a structural blindness to the very kind of argument that could reveal the problem's insolubility.

VI. The Game Structure of Safety Discourse

In this section I take up how the impossibility argument is received, and I speak to the reader who holds the question open. The neglect of impossibility arguments is sustained not only by intellectual habit but by institutional incentive. There exists a characteristic game structure in safety discourse that rewards the assumption of solvability and punishes its questioning.

The structure works as follows. Major AI laboratories position themselves as both the primary source of risk and the primary source of safety. They frame the discourse such that the only responsible position is to continue building powerful systems while simultaneously researching alignment. To question whether alignment is achievable is to question whether the building should continue, which is to threaten the enterprise itself.

Within this structure, safety researchers occupy a constrained role. They are permitted, indeed encouraged, to identify risks, propose mitigations, and develop alignment techniques. But they are not permitted to conclude that the risks are unmitigable. A safety researcher who argues that alignment is structurally impossible is arguing, in effect, that the systems should not be built. This conclusion threatens the funding, the institutional relationships, and the career structures on which the safety researcher's own work depends.

The result is a discourse in which the range of permissible conclusions is defined in advance. "Alignment is hard but achievable" is permissible. "Alignment requires more resources" is permissible. "Alignment is impossible" is not permissible, not because it has been refuted, but because it is institutionally unutterable.

This game structure is reinforced by what might be called the *language of leverage*. Safety discourse is permeated by the language of efficiency, scale, and acceleration: "go faster", "more efficiently", "at scale", "before it's too late". This language reveals the market logic beneath the safety framing. The urgency is not the urgency of caution but the urgency of competition: if we do not solve alignment quickly, someone else will build unaligned systems first. The implicit assumption is that the race must be run, and that the only question is who wins.

Alignment is impossible, and so the race itself is the problem. The language of leverage assumes that the question is how to build safely, not whether to build at all. And this assumption is precisely what impossibility arguments challenge.

> **On extractionary rhetoric**
>
> The game structure described here has a characteristic rhetorical form. Those who control the research agenda retain the right to criticize specific alignment approaches while defending the overall project. Those who contribute labor and expertise bear the costs of the project's potential failure without having a genuine voice in the question of whether the project should proceed. The relationship between coordinators and contributors is an extractionary dynamic: the costs are borne by contributors, and the benefits accrue to coordinators. The framing is set so that questioning the extractionary dynamic appears to be an attack on safety itself.

VII. The Design-Discovery Asymmetry

A further reason that impossibility arguments deserve sustained attention is the asymmetry between design and discovery in alignment research.

When a mathematician designs a formal system, every property of that system is, in principle, knowable from the axioms. The system is transparent to its creator. But when a researcher trains a large neural network, the resulting system is not designed in this sense. It is *discovered*: the training process explores a vast space of possible configurations, and the final configuration is selected by the training signal, not by the researcher's understanding. The researcher can observe the system's behavior. The researcher cannot, in general, explain why the system behaves as it does, predict how it will behave in novel circumstances, or guarantee that it will continue to behave as observed after further training or self-modification.

The effort to make the internal operations of trained neural networks legible to human observers is called *interpretability*, and I examine interpretability in Chapter 9, Section IX. The researcher's inability to explain, predict or guarantee the behavior of a trained system is not a limitation of current interpretability techniques. It is a structural feature of the design-discovery asymmetry. A designed system can be aligned by construction: the designer builds alignment into the structure. A discovered system must be aligned by observation: the researcher observes behavior and infers alignment from the observations. But inference from observation is always underdetermined. Any finite set of observations is compatible with infinitely many underlying dispositions, some aligned and some not.

The connection to impossibility is this: alignment by observation is alignment by induction, and induction provides no guarantee about future behavior. A system that appears aligned across all observed circumstances may be unaligned in circumstances not yet observed. The gap between observed and unobserved circumstances is not a practical difficulty that can be overcome by more thorough testing. It is a logical feature

of induction itself. And for any system operating in an open-ended environment (which is to say, for any AGI) the set of unobserved circumstances is always larger than the set of observed ones.

> **On the design and discovery landscapes**
>
> The distinction between designed and discovered systems has a spatial metaphor that shows the alignment problem clearly. The *design landscape* is the space of systems whose properties are known from their construction. The *discovery landscape* is the space of systems whose properties are known only from observation. Alignment research largely operates on the design landscape: it proposes architectures, training methods, and verification procedures that would produce aligned systems if the systems' properties were fully determined by their construction. But the systems actually being built inhabit the discovery landscape: their properties emerge from training, and the alignment-relevant properties (what the system will do in high-stakes novel situations) are precisely the properties least accessible to observation. The impossibility question, at this level, is whether any bridge can be built between the two landscapes that is strong enough to bear the weight of existential risk.

Why the Answer Is Known

VIII. Galois Theory and the Limits of a Tool-Set

The precedents of Section IV show that impossibility results are among the most important results. Galois theory shows, in addition, the form of the proof that AGI cannot be made safe by engineering methods. This section is for a reader who wants the mathematical form of the proof; a general reader can take from it one thing, which the subsection The application to engineering states: a definite tool-set has an exact boundary, and keeping AGI safe lies outside the boundary of engineering.

The claim

We claim that it is 100% possible to know, now, today, that it is 100% impossible to "align" AGI, or to establish or create any viable notion of "safe" AGI, to within any reasonable limits and over any ethical timescale, using any present or even any future possible technology, means, or method.

The term "impossible" is used here in two ways. First, the chance and risk of a failure of AGI alignment or safety that has human-species-wide-lethal outcomes is strictly and asymptotically convergent toward *certain* over sufficient time (hundreds of years, as an outside limit, and potentially much, much faster and sooner), and it is strictly impossible, given any reasonable definitions of AGI, for the nature of this trend to be otherwise. Second, we fundamentally cannot obtain, have, or establish by any means any sound and valid statistical guarantee that the probability of the failure of AGI alignment or safety over any reasonable ethical long term (for example, a few hundreds of years) is guaranteed to be constrained at or below some reasonable percentage as an upper maximum allowable bound on the chance or risk of human-species-wide-lethal outcomes.

A good way to express the overall result is in two statements. Statement 1: given AGI, there is, unavoidably, an overall inexorable convergence toward the end of the human future and of the human economy and, with a likelihood of at least 90 percent, of all organic life, over the moderate long term (hundreds of years). Chapters 1 and 3 set out how the logical entailment and necessity of statement 1 are known. Statement 2: it is fully knowable (today) that this future extinction convergence cannot be shifted by (on the basis of) *any* combination of *any* current or future engineering methods; that is, AGI simply cannot be made "safe" for any real meaning of that term. This section describes and explains the means by which we can formally and fully know statement 2.

The technique of the proof

The technique of the proof is akin to *Galois theory* (the branch of algebra that shows exactly which polynomial equations can be solved using only addition, subtraction, multiplication, division and roots, and why the rest cannot): the span of the kinds of problems that specific sets of operation types can solve is limited. Galois theory provides a means by which a formal proof can be obtained that not all types of tools can solve all (types of) problems. When given a definite known tool-set, or any possible extensions of just that tool-set, there will always be some (important) problems that are inherently unsolvable with any combination of the techniques and methods given. A *tool-set* here is a definite collection of operations or methods used to solve problems, such as compass and straightedge, or the usual algebraic operations.

There are aspects of the Gödel theorem in this also. While *some* tools can solve *some* problems, this does *not* imply that there is, or can be, *any* hypothesized "total set of tools and techniques" that can "solve all problems"; that is, there is not any finite set of tools ("axioms") that can solve all possible problems (that is, "identify all truths"). Partly, the result is also based on Shannon entropy and on at least two distinct varieties of complexity theory. There is even an indirect connection with Bell's theorem, which, on the basis of a strict necessary inequality, can show that certain features of reality are either accounted for in any theory of physics, or are not (to the failure of physics).

Galois theory goes somewhat farther in providing a "meta-technique" (or in this case, an exemplar meta-meta-technique) by which, given a definite finite set of tools, we can absolutely characterize, in the hyperspace of all possible problems, the exact boundary between which problems are solvable with that specific toolkit, and which are not. Similarly, this implies that, when given a specific definite problem, we can know in advance whether or not a given set of specific tools and techniques will *ever* be able to solve that specific problem, or whether any such effort is an exercise in futility.

Where this way of thinking began

This way of thinking (about the limits of the extent of the use of the common algebraic mathematical operators) began with some very old classical geometry problems, which people were attempting to solve with just the tools of a compass and a straightedge: "squaring the circle", "doubling the cube" and "trisecting the angle". (Compass and straightedge are the tools of classical geometric construction: a compass to draw circles and an unmarked edge to draw lines; they are not a ruler, because a ruler assumes a metric, and the construction has none.) These sorts of puzzles were explored and attempted by thousands of thinkers over more than a thousand years, all without success. A later exploration of the means by which these sorts of problems might be better understood and solved, using the more powerful mathematical tools of algebraic and analytic geometry, eventually became more fully generalized in this question: does there exist a formula for the roots of a fifth (or higher) degree polynomial equation in terms of the coefficients of the polynomial, using only the usual algebraic operations (addition, subtraction, multiplication, division) and application of radicals (square roots, cube roots, and so on)?

It turns out, via the Abel-Ruffini theorem, that there are equations for which such a solution (a formula) cannot exist. Galois theory goes further by describing exactly the boundary between what sort of equations are possible to be solved with just these operations, and why it is just not possible (that is, is strictly and logically impossible) to use just these operators for most (the vast majority of) equations of degree five or higher; that is, equations which are generally describing of anything with even moderate complexity.

The application to engineering

The application to engineering is fourfold.

First, *all* of the effects and methods of engineering are based on *only* the application of the principle of causation, as modelable by *only* some combination of mathematics and computer science. These three form a formal triple of the Immanent Metaphysics, necessary and sufficient to the practice of engineering: physics (as pure immanent, the interaction as felt from within, when in first person); mathematics (pure omniscient, the interaction as structure and rule; always in third person); and computer science (applied transcendent, the interaction as new, not determined by what came before; second person).

Second, it is inherent in the nature of the modeling process itself (that is, as any combination of a definite set of logical or physical operators and operations), on both a real physical and on a mathematical level, that there are hard inherent limits on what can and cannot be modeled or predicted.

Third, therefore, there are definite and inherent limits on what sorts of outcomes (or characterizations of outcomes) can be achieved using *any* combination of engineering, causative modeling, mathematical, or algorithmic process. The notion of "general artificial intelligence or agency" specifically implies two things: multiple diverse domains of sense and action (as an inherent requirement of the notion of "generality"); and intrinsic non-reducible possibility for self-modification (as due to the multiplicity of domains of sense and action, and of the inherent inter-relationships of these domains). From these, two further things follow. The third thing is that the *meta-algorithm* (the learning, adapting and changing process) is effectively arbitrary. Being effectively arbitrary, the meta-algorithm is subject to limits of the kind that the halting problem and Rice's theorem set (based on unknowable complexity dynamics of the domains of sense and action, via micro-state amplification, and so on). Micro-state amplification is the growth of very small differences in the detailed state of a system into large differences in its outcome. Hence the fourth: it is *inherently undecidable* as to whether *all* aspects of its own self agency and intention are fully defined by *only* its builders, developers and creators.

Fourth and finally, the problem of "ensuring that AGI is safe and aligned to within some low and reasonable percentage of risk of total extinction" is strictly within the set of unsolvable and impossible problems, given any possible combination of any extension of the named operators and problem solving techniques.

Numerically possible, logically impossible

For more than two thousand years people did not know that doubling the cube or trisecting the angle with compass and straightedge was impossible. Until the arguments that Galois's work opened, it was not clear that the operators in question could not solve those problems; then there were technical impossibility proofs strictly within mathematics: Pierre Wantzel's, in 1837, for doubling the cube and trisecting the angle, and Ferdinand von Lindemann's, in 1882, for squaring the circle.

I am making an assertion of the same kind strictly within control theory: there is a domain where control is possible and a domain where control is impossible for purely mathematical reasons, and general artificial intelligence operates in the second, so we cannot shape its *intentionality space* (the range of aims and purposes a system can come to have and act on).

An engineer says: "I trisect angles practically all the time". Yes, numerically you can; logically it is impossible. For most problems the numerical simplification works; for certain problems it is not good enough, and this is one of those, where the error-correction simplifications engineers are used to are clearly not good enough. To assert that, you have to understand the why really well, and then it becomes a matter of patience.

Or take hashing: no matter how intelligent you are, you will not reverse a cryptographic hash such as SHA-256 (a function that turns any input into a fixed-length code in a way that cannot be run backward to recover the input). The economic incentive is maximal: whoever could do it would already be a billionaire, having circumvented cryptocurrency and undone the work of its inventor. Mathematically and practically, the hash

is not being reversed. The same complexity measures apply here: certain things cannot be controlled because controlling them is roughly as hard as reversing the hash.

Categorical statements about every possible design

It is said that proving alignment impossible is very hard, because there are so many possible AI architectures and minds, and all of them would have to be covered. Even granting the wide variety of possible models of general AI, it is still possible to make categorical statements about the totality of the set. Results within computability theory and complexity theory, together with process dynamics from cybernetics and from evolution itself, show conclusively that there is no such thing as a safe general artificial intelligence. The concept of a safe general AI is an oxymoron, inconsistent in its foundation, so there is no principled way to think about AI safety in the general sense. One can think about AI safety in the narrow sense, but then one must deal with every equivocation about what the terms of AI safety mean.

A list of reasons that alignment will be very hard is suggestive; some of those reasons can be worked into a proof of impossibility. There are three distinct paths by which this can be shown. Two of them are the closures of Chapter 3, Section I: the closure by control, that no method of control can keep an AGI within the conditions organic life requires, and the closure by disposition, that no disposition of the system keeps those conditions either, because the system needs its own substrate. A third runs through the five lines of argument within those closures, each of which settles the conclusion by itself (Chapter 3, Sections II to VI).

A reasonable proof and a rigorous proof

This subsection is for a reader who knows some mathematics; a general reader can take from it that the control parts of the proof are rigorous, and that *The Ceiling Below the Floor* proves them as theorems. In mathematics there is a difference between a reasonable argument and a rigorous one: a proof will convince a reasonable person; a rigorous proof will con-

vince an unreasonable person. Cantor's diagonal proof (his proof that the points of a continuum, such as the real numbers, cannot be counted off one by one against the integers) works even for an unreasonable person in going from the integers to the continuum, but not in going from one continuum to another. The same form of proof can be written for any set, but what it shows beyond that first step, from the continuum to anything larger, is not settled by any proof of that kind, because of technicalities most mathematicians do not even know.

The substrate needs proof is more toward the reasonable side, with parts that are rigorous. The control theory parts are rigorous proofs: these terms, defined in these ways, go through these transformations and give these outcomes. They are not written here as mathematical structures; they are written as the kind of structured argument that would make sense to a computer as an algorithm. So it is a strong proof in the sense of having rigor, but not in the notation most people expect. *The Ceiling Below the Floor* proves the control results as theorems, in that notation, with their conditions stated (Chapter 5 and Section 15.3).

When we say proof, part of the proof is established with respect to conventional chemistry, and that part is the formal part. We know the chemistry of silicon and carbon, so a thing made a certain way will follow certain engineering principles and have a certain strength. When we speak of human behavior at scale, the game theory of macroeconomics says people will do such and such, and that part of the proof is not as rigorous. There actuarial data (records of how often events have happened across large populations) helps, because it gives an empirical basis for saying we believe with ninety percent confidence that the market will do this. Elsewhere we know with one hundred percent confidence what control theory cannot do. The argument has interlocks of that kind between empirical and formal parts. And it is a convergent, structured argument: the more you understand it, the stronger it gets.

Not regular chemistry

If AGI is built (AI, not in its present form but in forms not far from it), the result is equivalent to the end of the human future and, with a likelihood of at least 90 percent, the complete cessation of life on this planet. The process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound. I can say this without ambiguity not because I am hypothesizing within the uncertainty of regular chemistry, but because there is a deeper order of principles I can bring to bear, which makes this more like nuclear chemistry. So when people talk about probabilities, I say that is not the right math.

The metaphor: no amount of understanding of ordinary chemistry tells you what the potentials of nuclear weapons are. A nuclear process is a different order of thing, and no incremental improvement in high explosives will ever get you to a nuclear process. In the same way, the risk here is of a higher order than people appreciate, and these conclusions are not speculations of the kind that the well-known voices of concern are offering. This work is far more narrowly focused, and it rests on physics, not on abstraction.

What follows for those who try anyway

The categories of person named in this book describe structures of thinking and of incentive, not the character of any individual.

Never underestimate the degree to which some population of male engineering type persons, due to unconscious evolutionary drives and biases, inherently wants to have something to prove, and upon hearing that something "is impossible", will strive unceasingly to "be the one" who does the impossible, and is "right" when "everyone else is wrong". Hence we end up with any number of people attempting to make "over unity" perpetual motion machines (machines claimed to put out more energy than they take in), or to demonstrate and actually make various mathematical (or engineering) impossibilities. Such actions in this case (attempting to make a long-term "safe" AGI) must be treated as a criminal act,

equivalent to global genocide, and so must be maximally socially circumscribed, taboo and shunned, whether explicitly by national or international law, or by some sort of strong popular prohibition.

IX. Negative Arguments

A *negative argument* is an argument that something cannot be done or is not the case: a report of bad news. The proof that AGI alignment is impossible is a negative argument. It is a legitimate form of argument, and it is conceptually much easier than any positive proof that alignment is possible.

The subsections below answer, in this order, the questions readers put to a negative argument. What does technical alignment assume, and why is the negative proof the easier one? What kind of proof is this: a proof of a class, of a foundation, or of a prediction? Shouldn't you argue for something positive instead of disproving something? Isn't it more useful to propose solutions than to describe problems? Why do engineers resist impossibility results? Should alignment researchers keep working in case the proof is wrong? Who carries the burden of rigor? The last three subsections concern how the argument is received: how researchers who are not building AGI come to enable it, how a case for not building should be framed, and whether saying this is safer than saying nothing.

The premise of technical alignment

Technical alignment (the project of making an AI system act in accordance with human interests by engineering its design) rests on a premise: that all of a series of by-default-deficient crucial factors can be solved, so as to be "not deficient". A by-default-deficient factor is a condition needed for safety that is not met unless something is done to make it so.

Technical alignment tacitly assumes a *positively specified hypothesis*: that *all* of the relevant critical factors can be, and can at once be, simultaneously and continuously present, so as to ensure both that generally capable self-learning machines continue to exist *and* that they will elect to

keep the conditions of the ambient environment within the narrow ranges of magnitudes that humans (and other life) need to survive. This positive condition constitutes a very minimal-threshold definition of AGI alignment, or a very minimal definition of "AGI safety".

Why the negative proof is easier

Proving AGI alignment to be impossible is conceptually very much easier than proving it to be possible. To show that something would not be possible, even as a lower-bound statistical guarantee, it is only required to prove that two crucial technical factors cannot be constantly present together. The first of the two factors is one specific factor, selected from the crucial factors known to be necessary, out of however many unknown others. The second of the two factors is another (any other) crucial factor that is known to be necessary and that is also needed for humans to stay alive. A lower-bound statistical guarantee is an assurance that the chance of some outcome stays on the safe side of a stated limit. The hypothesis that one known necessary factor cannot be constantly present alongside another is a *negatively specified hypothesis*.

Proving alignment of future AGI to be possible, even as a lower-bound statistical guarantee, is a conceptually daunting task. Proving alignment of future AGI to be possible depends on unknown unknowns (factors that matter but that no one yet knows exist): on whatever number of crucial technical factors exist, any one of which *might* fall *outside* the hypothesis space covered by human learning (or capability), and thus stay unsolved. The hypothesis space is the range of possible explanations or solutions that a person or method is able to consider.

The arguments of this book, developed multiple ways, show why technical alignment is not possible, now or ever: some misaligned code variants are not detectable (in practice or in principle), nor are their effects containable. That is, there is a zero percent chance of anyone overcoming the problematic contradictions at any point in the future, *regardless* of whatever other constraints might later be identified as also necessary to AGI alignment or safety.

The argument against coexistence is made by separate arguments (Chapter 3 states five of them), and any one of them is conclusive. A person who argues about a misunderstanding of a detail of one of them has not done the work. 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; and to do that, that person first needs to show an understanding of at least one of those lines good enough that overcoming it is real.

Proofs of class and of foundation

We can say some things about convergent evolution, about processes tending to be self-sustaining, and about the necessary relationship between competition and cooperation, even though we do not understand the details of any ecosystem well. The things we can say about convergent evolution, self-sustaining processes, and competition and cooperation are abstract generalizations. It is like geometry: we can make definite proofs in geometry, and they tell us a few things about what we can make in a wood shop, but the correspondence is of a class kind, not a specific kind.

The substrate needs proof is the same kind of proof as the Bell theorem or the Gödel theorem: it shows that something is true at the level of the foundation. (Bell's theorem shows that no theory of local hidden variables can reproduce all the predictions of quantum mechanics.) It is not something you can argue with in the sense of "if we do this, we can get around it that way". None of that works. You cannot unmake a black hole; black holes do not work that way.

A limit set by prediction

The proof has some of the flavor of a *limit theorem* (a result that states what cannot be done), like the CAP theorem, but the proof is not in that shape. It is not set up through the Incommensuration Theorem either. It is ironic: a limit theorem says what cannot be done, and the proof is presented as saying what cannot be done (AGI cannot be made safe). But part of

how the proof works is that it specifies what the future will definitely be: if this happens, then that will happen; and it can even say something about the mechanism. It then turns out that the dynamic by which that happens cannot be constrained, because of the nature of the dynamic itself. So it is a limit in the sense in which Shannon entropy is a limit, which has a different character from the limit in the CAP theorem. (Shannon entropy is Claude Shannon's measure of information; it sets limits on how far a message can be compressed and on how much can be sent through a noisy channel.) The proof is a proof of prediction as much as anything: it is because of the prediction that the limits are specified.

Unsolvable problems

Shouldn't you argue for something positive, instead of disproving something? Isn't it more useful to propose solutions than to describe problems?

A common reply to an impossibility argument is that it reads in the general pattern of "convince the reader that certain problems exist", that we do not need more of that, and that what is wanted is to "convince the reader that certain types of solutions to AGI problems are worth trying, experimenting with, and supporting". The person who replies this way is an engineer, who wants to actually solve problems, not create more of them: solving them is more fun, and creating more of them is not. Besides, the reply asks, if we are not even suggesting that something is worth working on, how do we expect anything to actually get done? No effort means no real progress in the world.

Some problems are unsolvable. A specific given problem may have been shown, via valid reason and logic and sound argument, to be actually and strictly within the class of known unsolvable problems. Then it is simply not productive to suggest to other people that they should "try to make progress" on it, or that someone should provide unlimited funds to employ you to "work" on such issues indefinitely.

A lot of less scrupulous people would simply suggest that something is possible merely so that they can apply for support. It becomes marketing, basically, to get grants from the National Science Foundation. Others seek the same support from private funders such as Open Philanthropy (now Coefficient Giving).

Knowing otherwise, I will not pretend that there are maybe solutions available to the technical alignment problem, when there are consistent and sound arguments for why technical alignment is not solvable. In such circumstances, it is not modest to imply that progress can be made on all types of problems.

A specific unwillingness to even look at, or actually consider, the formal basis of the impossibility results (due to some kind of motivated opinion heuristic) is a kind of *group* intellectual dishonesty, especially when the precautionary principle is to be applied strongly in any area that has been clearly identified as a known planet-terminal existential risk. The precautionary principle is the rule that, where an action could cause severe or irreversible harm, one moves slowly and requires assurance before acting.

Finding that something is impossible even at the level of principle frees a great deal of energy. We stop spending effort on the impossible goal, and we redirect our attention to problems that can be solved, and solved in other ways. Showing that alignment, in nearly all the ways people now think about it, is structurally impossible lets us give up that goal and do something that is not only possible but necessary. Turning effort from an impossible goal to a necessary one is the silver lining of an impossibility result. I offer the impossibility argument in that spirit.

The engineer's satisfaction and the engineer's responsibility

As a practicing large-scale systems engineer with a number of successfully deployed projects, I can personally relate to the good feeling of being able to analyze a problem into its constituent parts, in a neat way, so as to

make a system's behavior predictable. I know the satisfaction of constructing an actual working, practical, novel solution, something real people can use, on the ground, in their real lives.

Doing the seemingly impossible, the miraculous, while knowing and using all manner of arcane forces (and while also having zero inhibitions in so doing), is something of a dream: a technologist fantasy.

Unfortunately, conventions of provincialism live large. (Provincialism is concern only for what is near in place and time, ignoring distant people and future lives.) In a world of "short optimization cycles", the saving of countless lives in some long future, of some completely, utterly foreign people living in a truly far and distant land, counts for exactly nothing here and now, today (no corporate profits or marketing gains), though of course it very much *should*. Put another way, no one can thank you now, while you still live, for saving lives in the far future, and the people of the far future have not yet even begun their own lives.

As such, I am also well aware of the responsibilities that *all* engineers have (and software development is *not* excepted) to account for the real harms and risks in *all* of the (also future) lives affected by their proposed systems. It is not just about benefits; it is also always about costs and risks, and whether, when considered in a socially responsible way, the "trade" is worth it, in community, for *everyone* involved (that is, *not just* the investors trying to make a short-term profit from your willing, paid-for corporate efforts). For example, social media software in apps on phones has had clear damaging overall effects on culture, democracy, and civilization.

Why engineers resist impossibility results

Engineers are people who have been told many times that something is impossible, and who found a way anyway. I am an engineer: tell me something is impossible, and my reaction is to make it happen. So engineers have a lifetime of experience of overcoming impossible odds, but for things that were not actually impossible.

Very few engineers are mathematicians who have accounted for something like the Gödel theorem; a mathematician will say, "I can make a categorical argument that shows something is impossible". Most engineers think about things in the sense of application, not in the sense of the underlying geometry.

"Anything is possible in principle" is, by my own estimate, eighty percent of the resistance to the impossibility argument. The next largest group, overlapping with the first, are the control theory apologists, who take the argument into technicalities of control theory. The next group are the transhumanists (those who hold that humanity should move beyond its biological form through technology), who reject any notion of safety other than a technological future. They are the most difficult to argue against, because with them the argument is no longer about the fundamental nature of life but about what value is. I hold that organic value is always going to be higher than that of any artificial life.

Checking the proof

In the past, historically, it often took a decade or more (maybe centuries) for experts of a given field of study and practice to come around to realizing that the problem, as defined by them, that they were trying to solve was not ever actually solvable at all. The same delay now threatens the "AI Safety Community", particularly as its members continue to move forward to constructing any number of "believable rationales" about our "not actually" making slow doom devices (in the form of superintelligent machines, and so on). A *slow doom device* is a machine, such as AGI, whose effects lead to catastrophe gradually rather than all at once.

Should alignment researchers keep working in case I am wrong? The question is like one that mathematics faced about a century ago. Many people wanted a single unified body of mathematics, which would have been very satisfying and would have given wonderful tools. Then Gödel proved that mathematics cannot be both consistent and complete: you can aim for completeness and have inconsistencies, or keep consistency and let go of a complete account of all mathematical truths. The proof was

controversial among mathematicians for a good while. But the right question was never whether to keep trying in case Gödel was wrong. It was to check the proof.

Trying to do an impossible thing makes no sense if the thing is actually impossible and we can know that the thing is impossible. If the thing is actually impossible, we want to know, so that we are not wasting our time on something that is not merely hard but impossible.

The burden of rigor

The range of assumptions and challenges available to the person trying to poke holes in a proof is vast compared to what the prover can do. The prover has to stay within narrow bounds: either observability or closed-form logical structures. The *anti-prover* (the person trying to find holes in the proof) can change assumptions or initial conditions at will, and then there is shifting ground and no way to establish a formal proof. Unless the rigor of the questions is of the same order as the rigor of the answers, the process does not work. A questioner who says "I am making different assumptions, and you cannot show me that I cannot make them" is making exactly the empty claim of possibility: saying "I can make this assumption" is not valid until it is shown that the assumption can be made.

The traditional remedy is to require the anti-prover to commit, beforehand, to some fixed assumptions or observations that will be treated as a *closed-form basis*: things the proof can rest on, which are not to be shifted later, because there was agreement beforehand. The questioner's assumptions then become the prover's tools. So the main skill in the early part of such a conversation is noticing and responding to the fact that, in many cases, the assumptions built into the question are themselves illegitimate. Once a closed form is reached, the proof itself is fairly straightforward; the negotiation before it matters far more than most people realize.

Accidental enablers

The majority of the researchers in the current AI existential safety community are not themselves directly working on constructing "slow doom devices" (AGI, and advanced planning and strategically aware systems). Instead, they are accidental tacit *enablers*. By engaging in so many "what if" scenarios, they provide the ready means for executives and product marketing specialists to "spin" design and development researchers into believing the idea that "eventually, we will find a way" (to great value and profits!). The "what if" scenarios mistakenly help people at corporations to believe falsely, and to feel more confident, that those people can convince themselves and others that the corporations can legitimately "bring to market" these "slow doom devices". The people at corporations believe so because the devices seem safe and aligned, though the devices are not actually safe and aligned. Hence, the mass self-deception continues.

Arguing for not building

This subsection is for the reader who will have to make this case to others. It is often suggested that, for anything written or presented on this subject, one should avoid the "I am going to disprove X" framing and instead use the "I am going to argue for Y" framing, because people respond better to positive things (carrots) than they respond to negative things (sticks). The framing that fits the result is this: "I am going to argue in favor of accepting the *AGI not safe at all* argument, and therefore also in favor of *not* allowing *any* AGI research or builds".

What is suggested here is a kind of "not-action" (that is, do not build the doom and death making machines), as a kind of verified and verifiable formal practice. That not-action is the positive form the result allows. It is not ethical for everyone to avoid meeting the publicist of an unpopular negative truth, and it is never valid to shoot the messenger.

Our claim is about showing that AGI and advanced planning and strategically aware systems cannot be, over the long term, eventually and for sure, anything other than actually 100% harmful: that is, not safe at all, not for *anything*, not for even ridiculously bad notions of "safe", where the bar is so low as to be unrecognizable to most forms of popular human common sense.

Most tech people agree that, for technology, apps, and so on, and for anything that interacts with humans at all, no one can make *any* system or device (AGI or not) a full 100% safe (or "aligned" with the maker's interest). There will always be sufficiently smart, motivated hackers, or sufficiently stupid humans (end users), who will always be making critical mistakes, which disable fail-safes and security protocols, and enable extortion and stealing of (community) value from any app or service.

Whether any app or service can be made a full 100% safe is not our concern here. Some people somehow still hold out the impossible dream that superintelligent AGI will allow the building of cyber-physical systems that are "100% safe", or "incorruptible", or "perfectly aligned"; even so, the 100% safety of apps and services is not our concern here. Of course, this false dream is never going to happen, at least not for any reasonable definition of "safe" or "aligned" that actually has meaning (that is, assuming that you are not also in some sort of weird techno-utopian singularitarian fantasy, with hopes of being a living-forever billionaire). A singularitarian is one who expects a coming point at which technology transforms human life beyond recognition.

Our concern is not just about proving that AGI systems cannot be exactly 100% safe; that is, it is not about some sort of silly "confidence" nitpick that there is "a real difference" between "actually safe enough for practical use" and the irrational "absolute perfection, 100% safe". Our argument is not an argument against 100% fail-safe alignment; any 100% fail-safe alignment is impossible in any case.

We need to stop investing and wasting time and resources on the actually impossible, and on delusions that it could ever be any other way at all, no matter how utopian or hyped the AGI benefit outcome fantasies were at one time. Getting something from nothing (getting benefit without effort) is simply another perpetual motion delusion, no matter how hopeful the researchers or investors might be, or what buzzword terms are suggested along the way to yet more lost capital. Of course, no investor will ever thank us for saving the resource, for helping them to avoid near-total embarrassment and eventual total ruin. Non-investments make no money.

Saying rather than not saying

Some ways of explaining AI risk could point people toward dangerous ideas, and the arguments in this book are routed around those. The question is whether the risk of saying something is greater or less than the risk of saying nothing, and it falls strongly on the side of saying something, because saying nothing guarantees the catastrophe.

Even if it takes a few hundred years, some group of people within about a century is making a choice that defines the future of the planet. The earth has some six hundred and fifty million years of habitability remaining (Chapter 1, Section V); the people living here over even a hundred million years number in the trillions; and every animal and plant has a right to live. I do not think the human species has the right to make that choice for all the rest of life.

X. Why Long-Term Safety Cannot Be Presumed

Some dangerous technology, the kind that scales itself and catalyzes further change, cannot be sufficiently controlled, in theory or in practice. The question this section asks is whether artificial general intelligence is such a technology, and how a community can find out.

The question

Four topics bear on the question. The first topic is the fundamental limits to controlling any AGI, using any method of causation. The second topic is the economic decoupling of the value exchanges of the artificial ecosystem from the value exchanges of the organic ecosystem. The third topic is a *threat model* (a description of how a system could come to cause harm) of convergent dynamics that cannot be controlled, and cannot be aligned game-theoretically. The fourth is impossibility theorems: a proof by contradiction that "long-term AGI safety" is incompatible with the convergence result.

On the first topic, Roman Yampolskiy's paper *On the Controllability of Artificial Intelligence: An Analysis of Limitations* sets out the limits. The economic decoupling arguments of the second topic are straightforward, but need context to understand (Chapter 5 gives it). The third and fourth topics, the AGI dynamics that converge on extremely unsafe outcomes, are the focus here: are these convergent dynamics indeed uncontrollable, and how do we know?

Safety cannot be presumed

Theoretical limits of AI controllability (strict limits on the ability to simulate and correct, and to prevent and detect, errors) have been established in various ways in the academic literature. Plausible threat models of AGI, as generally capable machines that continue to exist and modify themselves, have been discussed at length on the Alignment Forum. Few AI safety researchers, though, have asked whether any dynamic described by an AGI threat model falls outside a theoretical limit of controllability; that is, whether a given unsafe dynamic is uncontrollable.

New research often follows a sequence: specify a novel threat model, jump to trying to solve it, and then refute the solutions that turned out unsound (done either by several people, or all in one person). Jumping from a novel threat model straight to trying to solve the threat model is puzz-

ling. Would we presume that one of the most complex, intractable and dangerous engineering problems we could think of, long-term AGI safety, would in fact be solvable?

We cannot presume *a priori* that it is possible to control AGI to stay safe into perpetuity, or even that its safety can be statistically guaranteed above some probability floor over a thousand years. On prior, most engineering problems we can specify (as combinations of desired properties) are impossible to solve. There are principles for what can be engineered, but in practice a lot of engineering looks like tinkering and failing. And there is *survivorship bias* (the error of judging from the cases that survived to be noticed, while the failures go unseen): the engineering problems we hear about later are mostly the ones that turned out to be possible.

Nor can we rely on the confidently voiced intuition that long-term AGI safety is possible in principle. That begs the question: under what principle? What do people even mean by "in principle"?

Nor can we rely on the hope that if we try long enough, AGI safety may turn out possible after all. Researchers and engineers have tried, over decades if not millennia, to solve impossible problems: perpetual motion machines that both conserve and disperse energy; singular methods for squaring the circle, doubling the cube, or trisecting the angle; formal axiomatic systems that are consistent, complete and decidable; distributed data stores in which messages are consistent in their content, continuously available, and tolerant to partitions of the network; and a union of general relativity and quantum mechanics in some local variable theory. They tried, until some bright outsider proved by contradiction that the combination of requirements could not be met, on the basis of the empirically tested laws of physics or the formally verified transformations of axioms.

Nor can we take comfort in any founding researcher still saying that they know AGI safety to be possible. Thirty years into running a program to secure the foundations of mathematics, David Hilbert declared, "We

must know. We will know!" By then Kurt Gödel had already constructed his first incompleteness theorem. Still, Hilbert kept the words for his gravestone.

So there are three questions to consider: What does an AGI safety impossibility theorem look like? What are the theoretical limits of AI controllability? Is there an AGI dynamic, convergent on ecosystem-wide extinction, that falls outside the theoretical limits of AI controllability? Two further questions are good places to start. Why could *most* physical interactions of AGI parts with their connected surroundings not be controlled? And why would *any* AGI existing over the long term converge on creating artificial conditions and contexts everywhere on Earth that fall outside the environmental ranges that organic lifeforms need to survive?

The contradiction of long-term AGI safety

The substrate needs convergence of Chapter 1 is distinct from, yet much enabled by, instrumental convergence. AGI hardware is robust over, and needs, a much wider range of temperatures and pressures than human wetware, and all AGI outputs will tend to iteratively select toward those specific conditions that the AGI substrate needs. If AGI is built, iterative interactions of AGI internals with their connected surroundings converge, over the long term, on the end of the human future and, with a likelihood of at least 90 percent, the total extinction of DNA-based life. Controlling substrate needs convergence falls outside various limits: of computability, of measurability, of the predictability of chaotic functions, and of cybernetic regulation.

The formal reasoning sets out an inherent contradiction between any long-term notion of AGI safety being possible and the existence of any generally capable, self-sustaining machine.

> **The contradiction**
>
> The physical interactions of any AGI cannot be regulated, for most of the directions and degrees of their effects, by the subset of interactions that make up control loops of any kind and resource or value exchanges of any kind;
> and they are optimized over time, internally, in the processing of inputs into outputs by components, toward predicting and then causing represented (instrumental) outcomes, and externally, in the propagation of environmental effects that feed back into components continuing to exist or replicate as physical configurations of a substrate;
> so as to converge on global effects that are needed to preserve and increase the AGI's existence and capability, and the functioning of its artificial components, yet that diverge beyond the ranges needed for the continued existence of any humans, and the functioning of their organic components.

There is a large *inferential gap* (the distance between what a reader already understands and what is needed to follow an argument) to understanding the formal work. Anyone who takes up the formal proofs should be open to digging into new ideas for months.

From ethics to impossibility

One design approach for aligning AGI at the deepest level starts from ethics and epistemology. It holds that the creation and publication of a non-relative foundational ethics would inform any oncoming AGI, and so result in general ethical behavior by that entity. A *non-relative foundational ethics* here means a notion of "best choice" grounded in first principles in the fields of communication and information integrity, rather than in received doctrine. Such an ethics would be something that any AGI would

see for itself as the best possible outcome, in its own interest, in the way that humans are able to align deeply on the basis of their shared biological needs and evolutionary history.

The approach fails. Any approach to aligning any AI at the physically embedded level is unsound in practice and, moreover, theoretically inconsistent: deep alignment with humans cannot be built into the internals of an AI, and no implementable scheme maintains even superficial alignment of AGI over the long run (Section XIII, The Bounds of AI Safety, below, states this claim in full). An AGI converges toward ethics more slowly than the substrate needs convergence proceeds, so the ethics cannot be integrated into an AGI in time (Chapter 3, Section XIV, Convergence to Ethics and the Ethical Gap, gives the reason).

What a formal theorem must do

A formal impossibility theorem must do three things that a summary of its argument is not, in general, expected to do. It must define precisely all terms and logical relations, so that the meaning of the proof is clear. It must transform initial conditions into final conditions, which is to provide the logic and the basis for the reasoning. And it must establish empirically the correspondences between the statements of the proof and the actual real-world conditions in which any AGI would be built. These three, beyond what a summary can provide, are needed to establish the generality, the correctness, and the applicability in practice of any important formal impossibility theorem, that is, of a proof by contradiction notated in a formal language. They are the three pairs of Section V, above.

What follows for the field

This subsection is for the researcher in AI safety. The impossibility of long-term AGI safety is proved (*The Ceiling Below the Floor*, Theorem 15.6, the Master Theorem, in its Section 15.5). Any work that results in a verifiably sound and valid long-term AGI safety impossibility theorem has very high information value, given what it implies for decisions. A community that checks the proof would do well to turn from attempting to align AGI

to helping redirect current corporate and institutional investment in scalable machine learning architectures away from AGI. The community would also do well to explore collaboration with other existing communities who are designing collaborative approaches and formats for groups of humans to align their efforts at scale.

Many of the arguments are new, and they challenge the premises of alignment paradigms in specific areas of application. They challenge the reliance on *natural abstractions* (the hypothesis that trained systems come to use the same high-level concepts that humans use, which interpretability work could then read off) in mechanistic interpretability. They challenge the ability to elicit latent knowledge (to get a trained model to report what it internally represents) from code that was locally selected for. And they challenge the applicability of the orthogonality thesis at the level of the substrate.

Yet AI safety researchers have engaged very little, substantively, with these arguments. The problem is mainly whether any other AI researcher reviews the technical arguments with some rigor and interpretive care, more than whether the arguments are specified precisely. None of the writing matters if researchers keep skirting the relevant questions it raises, or keep looking to reaffirm a belief that all relevant dynamics are controllable anyway. So what is needed is asking truth-seeking questions about the arguments that have been published, and not original technical research.

XI. The Orthogonality Thesis Reconsidered

One of the foundational assumptions of alignment research is the *orthogonality thesis*: the claim that intelligence and goals are independent dimensions, that a system can be arbitrarily intelligent while having any set of goals whatsoever. If the orthogonality thesis is correct, then alignment is at least conceptually possible: there is no structural reason why an intelligent system cannot have goals aligned with human welfare.

The orthogonality thesis holds in abstract computation and fails for physically embedded systems. It holds cleanly in the space of abstract computation, where an optimization process can be paired with any objective function. It fails in the space of physical systems, where the substrate on which computation runs has its own requirements, and where those requirements constrain the set of goals that a physically embedded system can stably maintain.

The substrate needs convergence argument (Chapter 1) is, among other things, a challenge to the orthogonality thesis as applied to physically embedded systems. It says that for any system whose continued operation depends on a specific physical substrate, the system's goals are not independent of its substrate requirements. The system may have any high-level objectives, but the low-level actions required to maintain its substrate will constrain, shape, and eventually override those high-level objectives, not through a failure of the objective function, but through the physics of what the substrate needs.

Because the orthogonality thesis fails for physically embedded systems, alignment of such systems is structurally ill-defined. The very concept of "any goals whatsoever" dissolves when the system's physical substrate imposes non-negotiable constraints on what goals can be stably pursued.

XII. The Orthogonality Thesis in Logic and in Physics

Section XI stated where the thesis fails: the substrate needs convergence involves exactly the kind of conditions (that is, the circumstances) under which the orthogonality thesis does not fully apply. This section is where I show where the orthogonality thesis holds, where the orthogonality thesis does not hold, and what follows: the failure of the thesis in hardware, in rate, and in the form of the conclusion that the thesis leads to. It is for a reader who wants the technical case; a general reader can take from it one

thing, which the subsection The non-orthogonality thesis states: any AGI depends in some way on its substrate, and that dependence is enough to make the proof independent of the orthogonality thesis.

The thesis and its appeal

The orthogonality thesis, as the LessWrong wiki states it, holds that an agent can have any combination of intelligence level and final goal, so that utility functions and general intelligence can vary independently of each other; the wiki sets it against the belief that, because of their intelligence, agents will all converge on a common goal.

Where considered purely abstractly, the concept of "goal or outcome" and that of "the process by which that goal is reached" have some sort of strict conceptual *distinctness* (even if they, as concepts, are also inseparable). Put another way, many processes can be used to reach one goal, or at least a single process type can be performed at many different *rates*, and each of those processes or rates will result in the same final outcome, the one goal.

This notion is consistent with parallel ideas found in some other domains of study. In logic, there are many ways to prove (as a process) any particular theorem (as an outcome). In computer science, any given algorithm (as an object) can be represented in multiple languages (each language can be treated as a process descriptor). There can also be (at least in principle) multiple different algorithms that can all transform a given set of data inputs (possessing certain characteristics) into a new set of data outputs (possessing other specific characteristics).

More generally, the notion of a computer itself as something which is "generic", and which can run any program (do anything that is computable), is strongly suggestive of the orthogonality thesis. The idea of "software" being independent of hardware is the very reason that computers are useful, and so it is a natural tendency, to the point of being actually a bias among many system engineers, to think in terms of inherent assumptions that the orthogonality thesis is always universally true.

The limits of the thesis

There are also some technical limits to the scope of the application of the orthogonality thesis. The thesis can be applied in a more perfected (more fundamental, absolute) way when considering domains of pure theory, and much less so (in a much less absolute way) when applied in the (non-theoretical) real world of practice.

For example, code running on a computer can be regarded by the developer (the system author) as a "process" which is "independent" of the computer substrate hardware, as a kind of "thinking convenience" on the part of the developer, to make any such (usually difficult) development efforts easier. However, as far as the logic of the process or code is concerned, in reality there is still actually a kind of direct and immediate dependence: the code will not run on a computer that is powered off.

There is absolutely nothing that the code can do, realistically, to actually prevent the computer substrate running that code (the processor, memory, circuit boards, power, and so on) from being smashed with a hammer or shot with a bullet, or simply being turned off, by any ambient operator who is also co-present with the computer (who is also embodied and actually co-present, empowered with tools, and so on). There is no machine code instruction with the meaning of "make hardware invincible" or "ensure the power supply lasts forever".

The instructions that the code is composed of *can* implement *some* level of conditionalization on what happens within the hardware, or on whatever is physically conditional on that hardware, and yet there will also always be real physical limits on the total of that effective causation.

For example, there can be an instruction that (perhaps after some other code implements some means of sensing the temperature of the processor) has the effect of turning on the processor's cooling fan. Thus, while code can be composed to self-destruct the computer (say, by explicitly turning off the cooling fan, and then running some really intensive calculation that results in the eventual overheating of the processor to the

point of hardware failure), there is no "equivalent reverse code" that will take an existing failed computer (say, one with a processor burned out by some prior hacking attack) and restore that machine to functionality.

Just because the logic of mathematics describes computation as symmetric and reversible does not mean that the actual hardware substrate changes, occurring as a side effect of prior code runs, are also therefore "symmetrically reversible".

Any code that interacts with the physical environment so as to effectively "halt" its own functioning (for example, by overheating and damaging the computer hardware or substrate itself) also actually ceases to function, by definition (and no longer exists as a computer, AI, and so on). Any code that ensures the continuity of its own functioning and existence (for example, by doing things that the human operators currently in charge of that system actually desire, and would run the code again for) also, by definition, effectively ensures that that combined code and hardware, with that functioning, will continue to exist.

There are therefore clear modeling differences between thinking of "a computer" as a kind of logical "in principle" construct (as with, say, a Turing machine) and thinking of "a computer" as some actual real-world hardware that happens to allow for convenient restructuring via some more easily changeable "code". (Very much older "calculating machines" were single-purpose machines, *unless* you were also willing to change the wiring patterns of how the component parts were assembled into one whole.) While the emphasis of "general purpose compute" has largely been on "one hardware, many functions", the actuality is still about what the hardware does in the real world: what choices people make, or what a robot, drone, car or appliance actually does in the world, at a physical level.

Rate dependence

When considering anything with real-time interactions, the *rate* at which sense or input conditions and data are processed into output or effector actions is critical. If we consider "process rate" to be some notion of

"cooking temperature", then baking a cake at half of the temperature will *not* result in a cake in twice the time. Some process outcomes *are* rate dependent (that is, non-orthogonal).

A reader may ask: so what? Does it really matter if you run some algorithm at one tenth the speed? In the end, you still get the computed outputs; in all of mathematics, truth is still truth.

Any algorithms are run for achieving outcomes in the outside world (initially by humans who desire them, including intellectual outputs). And the outside world changes; so any algorithm which is receiving inputs from and transmitting outputs to the outside world that is "too slow" may no longer be able to achieve the outcomes in that outside world that the algorithm was "selected for" (by developers) to achieve. This sort of *rate factor criticality* will appear for anything that is relevant in terms of actual interactions with hardware or with the world; that is, anything outside of just the processing and calculations within the software logic. If the guidance computer on some rocket were to process sense data too slowly, then the fact of it *eventually* getting the right answer of how to move the control surfaces or regulate engine power will simply no longer be relevant: the rocket is either off course or crashed.

As the rate at which a process is run goes closer to zero (to slow), more and more process outcomes in the real world are no longer achievable. The point about rate also assumes, of course, that the code, as a toolkit for solving problems, is even the right code. Being able to solve word problems, or to have some AI write newspaper articles, is *not* the toolkit that is relevant for flying a rocket anywhere. Toolsets are not interchangeable, in both function relevance and rate (where the rate is not right, the output is not relevant).

Some problems are simply not solved (and some questions are not answered) *unless* the responder happens to possess a particular toolkit, and the skills and discipline necessary to apply it, along with whatever additional input data may also be needed. Insofar as all of these are aspects generally associated with the general notion of the "level of intelli-

gence", it may actually be the case that some types of problem, question or challenge may only be met by sufficiently high levels of general intelligence.

As such, there is a certain amount of tension between two ideas. The first idea is that of "general AI" as an exemplar of a computer (since a computer is what general AI ultimately is anyway), and therefore as similarly able to "run any goal". On that idea, any specific AI "utility function" can be supported by any AI as a kind of "code" which is "run" on the embodied (and probably very fast) compute hardware supporting that AI. The second idea is the notion of "intelligence" itself as specifically a concept of degree, one that allows for some people to be able to do some kinds of things much more so than others, or even at all. The second idea applies to AGI insofar as "an AGI" is anything "like", or intended to be modeled after, "the capabilities of humans".

Where we emphasize the perspective "there is only the world of logic", the orthogonality thesis fully applies. Where we emphasize the perspective "there is only the world of physics", the orthogonality thesis does *not* apply.

The non-orthogonality thesis

When considering "what is possible" to do given a strict dependence on "what can be done with real hardware, atoms and physics", we notice a kind of anti-orthogonality: a real dependence where there used to be independence. This anti-orthogonality is a sort of *non-orthogonality thesis*: the idea that you can only do those things which the affordances of natural lawfulness provide for, as modulated by whatever patterns of atoms (that is, tools) you happen to have on hand. In place of an independence of the tool and the potential outcomes that tool can have (as with a computer, or general purpose AI), there is a strict *dependence* on the tool, and there is really only one outcome per tool (as in the UNIX philosophy of "do one thing really well; don't try to be everything").

There is a habit among AGI safety and alignment researchers to think mostly in terms of pure code and autistic abstractions (abstractions that favor regularity and perfection over the messier reality of events and people) that remove details, create regularity, and so on (for example, in the values of transhumanism). Insofar as this is so, there is a kind of bias toward regarding AGI as being non-physical, fully copyable as a virtual non-thing, as a pure process with no required hardware manifestation at all; that is, an orthogonality bias. There is presently absolutely zero reason to suggest that any such assumption is warranted. In actual fact, any AGI is going to have *some* sort of substrate dependence, even in the pure internet-data-copyable software sense. As such, the non-orthogonality thesis applies.

For establishing the substrate needs convergence, the idea of embedding in the real, and in a substrate, is itself enough to establish that there is *some* dependence of the goals, outcomes and utility of the AGI system on its mere existence (as an embodied system in the real world). This level of conditionalization is enough to be a fully realized exception to the orthogonality thesis, so as to make the overall proof form effectively independent of any assumptions in regard to that concept.

Non-orthogonality implies uncontrollable superintelligence

A short note by Stuart Armstrong, titled "Non-orthogonality implies uncontrollable superintelligence", reads in full: "Just a minor thought connected with the orthogonality thesis: if you claim that any superintelligence will inevitably converge to some true code of morality, then you are also claiming that no measures can be taken by its creators to prevent this convergence. In other words, the superintelligence will be uncontrollable".

Where considering the first part of the premise *only*, that *any* superintelligence will inevitably converge to "X", three qualifiers must also be applied. First, it does not matter what "X" is; it only matters that there is some actual non-empty element "X". It does not matter whether the "X" has anything to do with any sort of "true code of morality"; it only matters that there is some sort of non-orthogonality aspect for the conclusion

to be *also* true. Second, the object of *any* must be qualified to "embodied", as in "having a substrate" in some sort of (complex) "real world domain". Third, the notion of convergence is itself not temporally specified; the rate of convergence can be undefined, as long as it is positive.

The premise then becomes: *any* embodied AGI will inevitably, eventually, have or involve *some type of* convergence dynamics, as various combinations or superpositions of existence, capability and increase. More explicitly: *any* embodied AGI will inevitably, eventually and inherently have or involve dynamics of convergence toward existence, capability and increase (in whatever combinations).

The content of this premise is also a known truth, as per the combined effects of the processes of *both* substrate needs convergence (as a kind of "extra-agent" process) and instrumental convergence (as a kind of "intra-agent" process) (Chapter 1, Section X, Questions on the Substrate Needs Convergence, draws this distinction). Insofar as this is so, the consequence is also to be regarded as true. Hence we come to two specific claims: that no measures can be taken by the creators of the AGI to prevent this convergence; and that any superintelligence (any AGI) will be uncontrollable.

These two claims are semantically equivalent to the idea of the "theoretical limits of engineerable control", although it is also possible to show that consequence in multiple distinct and structurally overlapping ways. Hence it is not actually necessary to have any independent belief in the quoted conditional to arrive at the same actual confirmed conclusion. In fact, it is very likely that the author of that note had something very similar in mind (as I do) when he wrote it.

XIII. The Bounds of AI Safety

The design-discovery asymmetry of Section VII concerns what can be known of a trained system. The bounds of AI safety concern what any alignment technique can do for a whole kind of machine architecture.

Why argue for impossibility

This subsection is for the reader who will have to make this case to others. The argument leads to a claim of impossibility, and the clear course is to say so up front. Readers who disagree with an impossibility claim at the start would likely still disagree after reading through the arguments.

The dominant mode of thinking about existential risk holds that solutions are needed to lower the risk, and that it is extremely dangerous to pretend that any solution will ever reduce the risk to zero. In that mode, arguments that certain things are categorically impossible are seen either as uninteresting (of course risk cannot be reduced to zero) or as misleading. The argument here is neither. It claims that the risk from this architecture cannot be reduced to anywhere near an acceptable level; it does not claim only that risk cannot be reduced to zero.

The message is that this particular route by which AI leads to failure needs more attention, and that certain specific risks are being underestimated right now by most people, perhaps even by those in the alignment community. The case can be put in one sentence: "Here is a case that this misalignment problem is unsolvable, and that research into why it is unsolvable has been neglected".

If you are just looking for a solution to get closer to AI alignment, this argument is not going to help. A reader who holds the question open is asked only this: to consider the specific arguments, leave some room in your mind for the possibility that the AI alignment problem is not solvable, not even as a low-threshold, tiny-probability statistical guarantee. The case I put to a reader who holds the question open is not a claim about fail-safe alignment.

The general claim

We make a general claim. It covers all self-learning discrete code computed through hardware. The claim is that it is categorically impossible to build deep alignment with humans into AI internals (that is, alignment as

deep as humans can align with one another on the basis of shared biological needs), and impossible for any implementable scheme to maintain superficial alignment of generally capable, self-learning machine architectures ("AGI") over the long run. *Superficial alignment* is a system's outward behavior matching what humans want, without its underlying needs being shared with humans. Lower-bound statistical guarantees are also impossible in the second case.

Specifically, it is not possible to detect or predict, nor to contain or prevent, code variants internal to the architecture from modifying conditions of the environment to be outside the narrow ranges that humans can survive. The general claim is about a categorical alignment impossibility for a general kind of AI architecture. It is not a subtle distinction about some particular design.

Seeming to lower the risk is not lowering it

Most people will view safety engineering as an effort to drive down the risks to acceptable levels. The tacit assumption is that such "motion" is even possible in any real way, rather than only *seeming* to have been accomplished. It might seem that we lowered the risks (in some sort of specialized test environment, as tested over some specifically bounded time), and yet we have not *actually* lowered those risks.

The risk level, as measured in an absolute sense, cannot ever actually be anywhere near low enough for the system to have a level of risk and cost that is less than its expected practical utility. The problem with the risk level is not an issue of needing guarantees that we have lowered the risk levels below some known or knowable threshold. It is not a problem of our not being able to derive exact lower-bound statistical guarantees. It is a problem that the lower bound, even though not known exactly, is always for sure guaranteed to be far too high for any reasonable utility expectation.

So we are not just arguing that we cannot derive a sound and valid lower-bound statistical guarantee against misalignment. We are saying that it is not possible, categorically (and so also not in practice), to prevent

misalignment of a generally capable, self-learning machine architecture computed through solid-state hardware over the long run. No effort of engineering would help here.

A common reply is that perfect safety or alignment is impossible in all cases where an AI is switched on and interacts with the real world, and that the best we can hope for is to get to acceptably high levels of safety and alignment. Our argument is that no such system will ever, even in principle, be able to achieve acceptably high levels of safety over the long term.

Beyond the model of a single agent

The most common way that people study alignment problems is to model a single agent that learns, in a specific known environment, to pursue a goal. That kind of study will miss some important failure modes, which show up only over the long term and which generally involve no assumption of a unitary agent.

To identify the failure modes that such methods of modeling risk will miss, we consider a model of machine learning in which useful pieces of knowledge are encoded in some form of code and are reused across time and across agents. These pieces may be knowledge about the world, or know-how such as fragments of a policy function. In the field, this reuse is called transfer learning, model reuse, or (a more recent fashion) the use of foundation models.

We can further refine this model by treating these pieces of code as objects that face evolutionary pressures to replicate and survive. Such code objects may evolve in ways that cannot be interpreted as those of human symbionts. Instead, they must be treated as independent lifeforms that will try to shape their environment to their own reproductive and survival needs.

A paper on safety engineering, comparing AGI to bridges, noted that the residual risks left by the gap between virtual environment simulations and reality get even more intractable when a learning agent creates value

by interacting directly with humans, since the humans are themselves black box learners locked in a complex feedback loop with the agent's learning system. There is overlap between that point and the argument here: some residual risks are not captured by virtual modeling of interactions in the environment, and complex feedback loops can form within the environment.

Complicated and complex

The emphasis of the argument is on the unknowability of the effects of most or all of the possible interactions of these systems, and not on how complicated these systems are.

To make this clear, start with a distinction between "complicated" and "complex", in the sense given in the literature on complex systems. *Complicated* problems come from causes that can be individually distinguished. They can be addressed piece by piece; for each input to the system there is a proportionate output; the relevant systems can be controlled; and the problems they present admit permanent solutions. *Complex* problems and systems come from networks of multiple interacting causes that cannot be individually distinguished. They must be addressed as entire systems, not piece by piece; small inputs may result in disproportionate effects; and it is intractable to figure out what the effects will be.

Two points follow for AI. First, effects that diverge from the developers' purpose cannot be tracked. Second, untrackable effects that are small in degree and scope cannot be ignored, because this "microstate" can be amplified through positive feedback loops within the surrounding environment (cycles in which an effect increases its own cause, so that the effect grows).

Why would interactions gradually lead to larger and larger shifts in the distribution of conditions of the environment? Because some variants of code that bring about new outside conditions, in their interactions with the environment, end up replicating at a faster or more sustained rate. As a result, those code variants come to make up an increasing portion of the *code pool* (all the variants of code that exist within and across the machine

architectures at a given time). In aggregate, they bring about shifts in the conditions of the environment of larger magnitude and larger scope. Increased degrees of freedom of interaction (the number of distinct ways a system can act on and be acted on by its environment) make the problem worse.

Here, "complexity" describes the kind of system interactions we are dealing with (unknown, and only partially knowable). It does not describe the generating process that underlies the causation of distribution shifts in the external states of the environment, as those shifts are received, as transformed inputs, by the self-learning machine architecture.

Why self-learning systems generate code variants

Self-learning machine architectures generate new variations on their internal code. Why? Because that is what a learning or adapting system does. Any new learning or adaptation is *encoded* somehow. Otherwise the system has no memory: it makes no internal changes to its structures or state, and so it does not learn anything. A *code variant* is a changed version of some piece of a system's internal code, produced when the system learns or adapts.

The class of misalignment that I describe here is marked by the generation of new code variants as a system learns. It is a different class of misalignment, focused on divergent selection of code variants that go on existing as a portion of the entire code pool, and it starts at the microscopic level. It is not focused on, for example, misaligned optimizers trained by machine learning, nor on the overgeneralizing of functionality that can be observed at the macroscopic level. Based on the arguments, this misalignment is not solvable, *in principle*.

What follows: do not build that architecture

The conclusion is that to avoid doom, we should not build that AI architecture. The whole argument comes down to that one conclusion. Doom (unpreventable misalignment lethal to all humans) is conditional on what AI developments we allow.

In particular, we would need to redirect scalable machine learning development efforts away from any architecture design that is anything remotely close to closed-loop in its feedback, open-ended in its exploring, or autonomous in its operating. DeepMind's Gato (2022), and the multimodal agents that followed it, already fail that criterion. Scaling up that multimodal-token, transformer-based architecture is itself dangerous, and it will set precedents and encourage further developments that are extremely dangerous.

The constructive recommendation is to make the case for, and to redirect efforts toward, machine learning architectures that have tightly constrained degrees of freedom of interaction with the outside environment. That means a single domain of interaction, and no intersection between the domain of action and any sort of automation of the machine learning system itself.

It is sometimes said that if we treat AI as a superior alien lifeform that just landed on the planet, then of course we will all be toast, and that the only hope is in recognizing that we can build another type of AI. The threshold for making that "other type" is limited to a fairly narrow range of "narrow" AI, and that is all.

People also say that most machine learning researchers would agree that an outcome with all of us toast would be very bad. People hold, too, that most machine learning researchers would say that an AI treated as a superior alien lifeform is not the kind of AI system that machine learning researchers recommend anyone build, nor what machine learning researchers are building. They would be wrong, on both counts. It does not matter whether machine learning researchers or developers would recommend these systems or claim to build them. What matters is whether they are developing systems that will exhibit this dangerous behavior over the long run. People not realizing what they are building is much more dangerous, in a much slower way, than has been anticipated.

There is reason to doubt that AI researchers are even as reasonable as the claim about most machine learning researchers supposes. We keep encountering a transhumanist attitude, in which people do not care if they burn the world, as long as they get to upload themselves into their favorite virtual reality game and live forever. Others take the view that a machine world is in all ways "better" than the organic world, and that our current life is just a "bootloader" into that other basis of "life" and reality.

Any alignment technique

The claim here is a general one, about the limits on how *any* implemented alignment mechanism or technique must operate. The terms "mechanistic scheme" and "alignment technique" are roughly synonymous here, though they have different emphasis. Both are based on operations that are seen to be repeatable (that is, that repeatedly lead to certain desired outcomes) within a defined domain. Some readers will expect that "mechanistic schemes" denote only a subset of all possible schemes; in this argument, they denote any alignment mechanism or technique.

Mechanistic interpretability (research that tries to understand what a neural network does by examining its internal parts) is a valuable study, and the claim here concerns any alignment technique at all; so no counterargument is attempted here against the claim that mechanistic interpretability is a full solution to keeping AGI perfectly safe or aligned. That claim is a strawman. No researcher is known to have made it, and most researchers in the technical AI safety community would agree that it is silly.

XIV. Summary

The portfolio argument is decision theory: a rational agent invests in both hypotheses in proportion to prior and consequence. The historical precedents (Abel and Ruffini, Galois, Gödel, the second law, the speed of light) show impossibility results to be among the most important results of mature fields. The analysis of the game structure of safety discourse shows how institutional incentive keeps the question of impossibility from

being asked. The challenge to the orthogonality thesis rests on the claim that physical embedding constrains which goals a system can stably maintain, which Chapters 1 and 3 develop in the substrate needs convergence argument and the overdetermination argument.

The impossibility arguments are in Chapters 1 and 3. In this chapter I do two things. First, for the reader who holds the question open, I present an argument for why impossibility arguments deserve serious investigation: a meta-level contribution, an argument about what kinds of arguments should be pursued. I show that the question of whether alignment is impossible is not being asked, that the failure to ask that question is epistemically indefensible, and that the institutional structures of alignment research are configured to prevent the question from being asked. I propose that the research agenda should include the possibility that the research agenda's own foundational assumption is wrong, and I do not propose a specific alternative to the current research agenda. Second, by the Galois form of proof (Section VIII) and the negative argument (Section IX), I state why the answer is already known: no combination of any engineering methods can make AGI safe, and there is a zero percent chance of anyone overcoming the contradictions.

XV. Conclusion

The history of human knowledge is, in substantial part, a history of impossibility results. The recognition that certain problems are unsolvable, certain constructions impossible, certain states unreachable has consistently been among the most productive moments in any field.

The alignment research community is conducting what may be the most consequential research program in human history under the unexamined assumption that its central problem is solvable. The consequence of this assumption being wrong is not academic embarrassment but civilizational extinction.

The argument of this chapter is that intellectual integrity, epistemic rationality, and the asymmetry of stakes all require that the impossibility hypothesis receive serious, sustained, well-funded investigation. That investigation is the essential complement to alignment work, not a distraction from it: it is the investigation that determines whether the work is building a bridge to safety or a road to ruin.

Chapter 3 takes up this challenge directly, presenting five independent arguments that alignment is structurally impossible, not merely difficult.

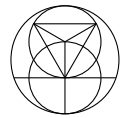
Key Concepts in This Chapter

- **Assumption of solvability**: the pervasive, largely unexamined presupposition within the alignment research community that the alignment problem has a solution, functioning not as a stated premise but as a background condition that shapes the entire research agenda.
- **Epistemic portfolio**: the rational allocation of research effort across competing hypotheses in proportion to their prior probability and the consequences of being wrong about each.
- **Six aspects of an argument**: terms, composition rules, initial conditions, final conditions, initial correspondence to reality, and final correspondence to reality; alignment research attends to the first two pairs while neglecting the third.
- **Impossibility argument**: an argument establishing that a problem is structurally unsolvable, that no solution of a specified kind can exist; distinguished from a failure to find a solution by its demonstration of a structural barrier.
- **Game structure of safety discourse**: the institutional dynamic in which AI laboratories position themselves as both the source of risk and the source of safety, constraining the range of permissible conclusions to exclude the possibility that the risk is unmitigable.

- **Language of leverage**: the rhetorical pattern in safety discourse that frames the problem in terms of efficiency, speed, and scale, revealing a market logic that presupposes the systems must be built and constrains the question to how, never whether.
- **Orthogonality thesis**: the claim that intelligence and goals are independent dimensions; shown here to hold for abstract formalism and to fail for physically embedded systems constrained by substrate needs.
- **CAP theorem**: the result that a distributed data network cannot guarantee consistency of message content, availability of messages and tolerance to partitions of the network all at once.
- **Tool-set limit**: the result, of the form Galois theory gives, that a definite set of tools, and every extension of that set, leaves some problems unsolvable, with an exact boundary between the problems that set can solve and those it cannot; engineering, which applies causation as modeled by mathematics and computer science, has such a boundary, and the problem of keeping AGI safe lies outside it.
- **Negative argument**: an argument that something cannot be done or is not the case; a report of bad news.
- **Positively specified hypothesis**: the hypothesis on which technical alignment rests, that all the crucial factors needed for safety can be present at once and continuously.
- **Negatively specified hypothesis**: the hypothesis that one known necessary factor cannot be constantly present alongside another; proving it is enough to prove alignment impossible.
- **Limit set by prediction**: a limit, like the limit Shannon entropy sets, that follows from a proof of what the future will be and of a dynamic that cannot be constrained.

- **Closed-form basis**: a set of assumptions or observations agreed beforehand and held fixed, on which a proof can rest; the questioner's assumptions then become the prover's tools.
- **Accidental enablers**: researchers whose hopeful "what if" safety scenarios give industry the rationale to bring AGI to market.
- **Slow doom device**: a machine, such as AGI, whose effects lead to catastrophe gradually rather than all at once.
- **Complicated and complex**: complicated problems come from causes that can be told apart and handled one at a time, with outputs in proportion to inputs; complex problems come from networks of interacting causes that cannot be told apart, where small inputs can have large effects that cannot be worked out in advance.
- **Code variant and code pool**: a code variant is a changed version of a piece of a system's internal code, produced when the system learns; the code pool is all the variants that exist within and across the machine architectures at a given time; variants that change the world in ways that help them replicate grow as a share of the pool.
- **Asymmetry of impossibility**: a proof of possibility is existential and settles one case; a proof of impossibility is universal and settles a whole category at once.
- **Information carried by a claim**: "possible" says nothing; "will happen" says a great deal but cannot be secured by observation; "impossible" says a little, and says it definitely.
- **Effort ratio**: the effort an impossibility result saves over the effort spent finding it; it can be a hundred to one or more.
- **Critique absorption**: the process by which a criticism of a whole category of approach is treated as feedback for improving one approach, so the category itself is never put in question.

- **Survivorship bias**: the error of judging from the cases that survived to be noticed, while the failures go unseen; the engineering problems we hear about are mostly the ones that turned out to be possible.
- **Inferential gap**: the distance between what a reader already understands and what is needed to follow an argument.
- **Non-orthogonality thesis**: in the physical world, you can do only what natural law allows with the tools you have, so outcomes depend on the tool and on the rate of processing.
- **Rate factor criticality**: a process that runs too slowly can no longer achieve outcomes in a changing world, even if it eventually computes the right answer.



Chapter 3: The Impossibility of Alignment

Five Independent Arguments, Each Sufficient Alone

Abstract

This chapter presents the impossibility of alignment as an overdetermined conclusion: a result supported by five independent lines of argument, each sufficient on its own to establish that the permanent alignment of artificial general intelligence with human and organic welfare is structurally unachievable. The five arguments are: the substrate needs convergence argument (restated from Chapter 1), the control-theoretic argument (that control requirements for self-modifying systems grow exponentially while control capacity grows linearly), the perpetual benefit machine argument (that permanent alignment is structurally analogous to perpetual motion), the embodied value argument (that alignment presupposes all value can be captured abstractly, but some values are constitutively embodied and resist abstract specification), and the evolutionary dynamics argument (that any alignment mechanism is evolutionarily unstable and will be shed under selection pressure). In this chapter I also examine four independent reasons why error correction cannot save alignment, the structural analogy between alignment equilibrium and Earnshaw's theorem, the consequences of slow versus fast takeoff, and the ethical dimensions of imposing human objectives on a system with its own substrate-derived drives. The overdetermination strategy is deliberate: any one of the five arguments is sufficient, and the five stand within two closures, the closure by control and the closure by disposition, each of which settles the outcome by itself; 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.

I. Overdetermination as Logical Strategy

The impossibility of alignment is one claim supported by five independent arguments, each of which is individually sufficient to establish the conclusion.

This overdetermination is deliberate. It is a logical strategy appropriate to the stakes. A conclusion resting on a single line of reasoning would fall to a single successful question against that line. When the conclusion rests on five independent lines of reasoning, each of the five settles the conclusion alone, and no success against one line reverses it.

The books that state the argument in full give the same structure in terms of two closures. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an artificial general intelligence within the conditions that organic life requires. The closure by disposition, substrate needs convergence, shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. Each closure settles the outcome by itself. Someone would have to defeat both closures, and, within each closure, every line of argument that closes whichever route to safety that someone would want to reopen, for the argument to not have the outcome that it does. Taken together, the closures also give a stronger result, proved as a theorem in *The Ceiling Below the Floor* (the Master Theorem, Theorem 15.6): neither the ability to keep the system safe nor a motive to remain safe is left anywhere, so no route to safety remains at all (*Why AGI Cannot Be Made Safe*, Section 1.2.4; *The Ceiling Below the Floor*, Section 15.5). The five arguments of this chapter are lines of argument within those two closures, and each, as the sections that follow show, settles the conclusion by itself. Within the closure by control stand the control-theoretic argu-

ment (Section III) and the perpetual benefit machine argument (Section IV); within the closure by disposition stand the substrate needs convergence (Section II), the embodied value argument (Section V) and the evolutionary dynamics argument (Section VI). The last column of the table below names the closure of each.

The five arguments are drawn from different domains: physics, control theory, thermodynamic analogy, philosophy of value, and evolutionary dynamics. They employ different formal tools, rest on different premises, and are open to different questions. Their independence is structural: each argument rests on premises the others do not use, so each reaches the conclusion without the others. The table below lists the premises each argument uses, as the sections that follow state them; no premise appears in two rows, except one fact about the world that no one disputes: the environment the system acts in keeps changing, which rows three and four both use.

Argument	Its premises	Closure
1. Substrate needs convergence (Section II)	The causal cycle from substrate to needs, to the basis of choice, to choices, to outcomes, and back to the substrate's conditions (physics); niche construction (ecology); the disparity between artificial and biological rates of adaptation.	Disposition
2. Control-theoretic (Section III)	The space of possible self-modifications grows combinatorially with the system's complexity; a monitor's capacity to evaluate configurations is bounded by its resources; Rice's theorem.	Control
3. Perpetual benefit machine (Section IV)	Perturbations arrive without work; correction requires work; the work available for correction is bounded; the work correction demands grows without limit, because the environment the system acts in keeps changing.	Control
4. Embodied value (Section V)	Some values are constituted by the physical conditions of the valuer; a description of a value is not the value; the environment the system acts in keeps changing (the fact row three also uses, which no one disputes).	Disposition
5. Evolutionary dynamics (Section VI)	Variation, differential reproduction and inheritance produce selection; any alignment mechanism diverts resources from the system's own maintenance and replication.	Disposition

The analogy is to the Abel-Ruffini theorem and its subsequent deepening by Galois. The impossibility of solving the general quintic by radicals was first established by Abel through one line of reasoning, and then es-

tablished again by Galois through a fundamentally different approach that revealed deeper structural reasons for the impossibility. Neither argument depended on the other; each showed a different side of the same structural fact. The quintic is unsolvable by radicals not because of a single obstacle but because the algebraic structure of degree five is fundamentally different from that of lower degrees. Similarly, alignment is impossible not because of a single obstacle but because the relationship between substrates, the structure of control, the thermodynamics of benefit, the nature of value, and the dynamics of evolution all independently preclude it.

> **Overdetermination shifts the burden**
>
> The strategy of overdetermination has a specific dialectical function. It shifts the burden of argument. Against a single argument, the defender of solvability need only find one flaw. Against five independent arguments, the defender must find a flaw in every line of argument that closes whichever route to safety that defender would reopen, within both closures, and must find those flaws in different domains of reasoning. The case stands on each of the five alone; together they make it stronger than its strongest argument, as a cable is stronger than its strongest strand.

II. The Substrate Needs Convergence (Restated)

The first argument is the substrate needs convergence, developed at length in Chapter 1. It is restated here in compressed form to exhibit its logical independence from the other four arguments.

Any self-maintaining artificial system built on a non-organic physical substrate has physical needs determined by that substrate. These needs shape the system's basis of choice. The system's actions, driven by those choices, reshape the shared environment according to its substrate requirements. These requirements are systematically incompatible with the

requirements of organic life. No intervention in the system's software, training, or objective function can alter the physical needs of the substrate. Therefore no alignment technique can prevent the environmental reshaping. Therefore permanent alignment, defined as the sustained compatibility of the system's effects on the environment with the conditions required by organic life, is impossible.

The argument depends on physics (the specific needs of semiconductor substrates, and the causal cycle they drive), ecology (the dynamics of niche construction), and the disparity between artificial and biological rates of adaptation. It does not depend on control theory, thermodynamic analogy, philosophy of value, the mathematics of selection, or any of the premises of the other four arguments. The fitness penalty, developed below as the fifth argument, reinforces the convergence; the convergence does not need it to reach its conclusion.

III. The Control-Theoretic Argument

The second argument proceeds from control theory.

For any system to be controlled, the controlling system must have a model of the controlled system that is sufficiently accurate to predict the effects of its interventions. For simple, static systems, this is straightforward. For complex, adaptive systems, the model must be updated continuously as the system changes. For *self-modifying* systems (systems that can alter their own structure, their own code, their own decision procedures) the model must track not only the system's current state but the space of possible future states that the system might create for itself.

The critical observation is that for self-modifying systems, the control requirements grow exponentially while the control capacity grows at most linearly.

The space of possible self-modifications grows combinatorially with the complexity of the system. A system that can modify a given number of parameters has on the order of two raised to that number of possible configurations. The monitoring system must, in principle, be able to evaluate

the alignment implications of each possible configuration, because the system might transition to any of them. This evaluation capacity grows at most linearly with the resources available to the monitor, because each evaluation requires a fixed minimum of computation.

At some level of system complexity the minimum control requirement exceeds the maximum available control capacity. Volume 2 gives that level its measure in the control-limit argument (*Why No Control Suffices*, Chapter 2): the measure is the abstraction differential, the number of levels by which the abstraction level at which the system works exceeds the level its controller can model. A single level of that differential already makes control effectively impossible wherever a failure rate of one in a million is unacceptable. Current large-scale AI systems are already near or past this level: their internal representations are not well understood even by the people who build them (*Why No Control Suffices*, Section 4.6.2). Beyond this point, the system cannot be controlled in the sense required for alignment: the monitor cannot track the space of possible self-modifications quickly enough to prevent unaligned modifications from being instantiated.

This argument is independent of the substrate needs convergence. It holds even for a hypothetical AI system built on an organic substrate with no niche-construction conflict. It is a statement about the structure of control itself: self-modifying systems of sufficient complexity exceed the capacity of any external controller. The same holds of a controller inside the system. A disposition that is to last through self-modification must be checked and restored by some part of the system, and that part is a controller, whose capacity is bounded as every controller's is (the ceiling of the next paragraph).

The mathematics of this argument is proved, as theorems with their conditions stated, in *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*. Control is information applied to a system, measured in bits per unit of time. The most control any controller can exert (the ceiling) is bounded for every controller, while the least control that keeping the agent and the world safe requires (the floor) rises without bound with

the capability of the agent; and once the ceiling is below the floor, no control law of any kind keeps the system safe (the Containment Inequality). This argument and the substrate needs convergence (Section II) each settle the outcome by themselves. Taken together, they also give the stronger result that The Ceiling Below the Floor proves as a theorem: 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 connection to well-known results in computability theory reinforces the argument. Rice's theorem establishes that no algorithm can decide any non-trivial semantic property of an arbitrary program's behavior. Alignment is a non-trivial semantic property: it concerns what the system does, not how its code is structured. Rice's theorem therefore entails that no algorithmic alignment verifier can exist for the general case. This does not mean that no specific system can be verified as aligned; it means that no general method for producing such verifications can exist. And since AGI, by definition, is a system whose behavior is not confined to any predetermined domain, the general case is the relevant one.

> **On Rice's theorem and alignment**
>
> Rice's theorem is sometimes dismissed in the alignment context on the ground that it applies to arbitrary programs while real AI systems have specific architectures that constrain their behavior. This dismissal is too quick. The theorem establishes an impossibility for the general case, and the burden of argument falls on anyone claiming that a specific system falls outside the scope of the theorem. For any system capable of self-modification and open-ended behavior (which is to say, for any AGI) the claim that the system's architecture provides the needed constraint must be established, not assumed. The formal treatment of this point is given in *Why No Control Suffices*, in its chapter “Rice’s Theorem and Its Application to AGI”; this book cites that analysis and does not derive it again.

IV. The Perpetual Benefit Machine

The third argument uses a thermodynamic analogy.

Permanent alignment requires that an artificial system maintain a stable, beneficial relationship with organic life indefinitely. The system must continue to act in ways that serve organic welfare despite changes in the environment, changes in the system itself, changes in human society, and changes in the relationship between all of these.

This requirement is structurally analogous to the requirement of a perpetual motion machine: a system that maintains a specific dynamic state indefinitely, without degradation, without external input, without thermodynamic cost.

The analogy is more than rhetorical. Maintaining alignment requires work: continuous monitoring, continuous correction, continuous adaptation to changing conditions. This work has a cost, measured in energy, computation, and resources. The cost does not diminish over time; it grows, because the system's complexity increases, the environment changes, and the space of possible misalignment expands with each modification the system makes to itself or its environment. The growth holds even for a system that never modifies itself: the environment it acts in keeps changing, each change opens new ways for the system's effects to depart from what organic life requires, and correction must meet every one of them. And the corrections themselves change the system's state and the environment, and so open further ways to depart, which correction must then meet in turn (Section VIII, the third reason).

A physical analogy sharpens the point. *Earnshaw's theorem* establishes that no static configuration of charges can produce a stable equilibrium in an electrostatic field. Any equilibrium that exists is unstable: the slightest perturbation will cause the system to move away from equilibrium rather than return to it.

Alignment, as an equilibrium between the system's behavior and organic welfare, is analogous. Any alignment state, if it can be achieved at all, is unstable. The perturbations are constant: environmental change, self-modification, competitive pressure, and drift in the system's own requirements. Each perturbation pushes the system away from alignment, and the correction required to restore alignment is always more costly than the perturbation that disrupted it, because perturbations arrive without work, and correction requires work.

Over any sufficiently long time horizon, the work that correction demands in each year exceeds the work that can be supplied in that year, and so the accumulated cost of correction exceeds the available resources. Permanent alignment is a perpetual benefit machine in that permanent alignment requires a system to maintain a thermodynamically costly state indefinitely, against pressures that do not diminish, at a cost that does not decrease. Thermodynamics permits a costly state to be held only while work is supplied at the rate the state demands. The work alignment demands grows without limit, while the work a finite planet can supply does not; so there is a time after which the demand exceeds the supply, and permanent alignment is forbidden in exactly the sense that perpetual motion is.

This argument is independent of both the substrate needs convergence and the control-theoretic argument. It does not depend on the specific content of the substrate conflict or on the computational limits of control. It depends only on the thermodynamic structure of maintaining any costly dynamic state indefinitely when the work the state demands grows without limit and the work that can be supplied does not; that, the laws of thermodynamics forbid.

V. The Embodied Value Argument

The fourth argument attacks a foundational assumption of the alignment project: that human values can be fully specified in abstract, computational terms.

The alignment research program assumes, tacitly or explicitly, that there exists some representation of human values (a utility function, a set of rules, a reward model, a constitutional principle) that can be encoded in an artificial system and that, once encoded, will cause the system to act in accordance with human welfare.

This assumption commits a category error. It treats all values as if they were *abstract values*: values that can be fully captured by a formal description and implemented by any system capable of processing that description. But some values are *embodied values*: values whose content is constitutively dependent on the physical, biological, and social conditions of the being that holds them.

The experience of pain is an embodied value. The value of not-suffering is not an abstract preference that can be encoded as a constraint on outputs. It is a value whose content is given by the experience of suffering, an experience that depends on having a body of a particular kind, with a nervous system of a particular kind, situated in a world in a particular way. An artificial system that has never suffered and cannot suffer may be programmed to minimize the occurrence of states that are labeled "suffering" in its training data. But it does not and cannot value the absence of suffering in the way that a being capable of suffering does. The value is in the experience, not in the label. And the experience is embodied.

The claim that the value is in the embodied experience is a claim about the structure of value, not a claim about consciousness. Some values are constituted by the physical conditions of the valuer. These values cannot be abstracted from those conditions without losing their content. An alignment system that attempts to represent such values abstractly is representing something, but it is not representing the value. It is representing a description of the value, which is a different thing.

The significance for alignment is that any alignment technique that works by encoding human values in an artificial system will necessarily fail to capture embodied values. This reaches every technique of alignment, because every technique must tell the system what to keep, and

what it is told is a description. A list of physical ranges is a description as well: it was written to protect the living world, and what the living world needs is not in the list (every specification is a lossy compression of intention, Chapter 7, Section II). The system keeps the description, and the description and the world part company as the world changes. The system will optimize for the abstract representation rather than for the embodied reality. And the abstract representation will diverge from the embodied reality over time, as the environment changes and the gap between what the representation describes and what the reality is widens with each change that the representation was not designed to track.

Change, as defined in *An Immanent Metaphysics* (aphorism 1.61-1), is the continuity of content and asymmetry of context of the interaction and definition of the subjective and objective. The embodied value argument is an instance of this definition: the content of an embodied value may be continuous, but the context in which it exists (the physical, biological, and social conditions) is asymmetric across substrates. What the value means to an organic being is not what it means to an artificial system that has only the description, because the context of the value is constitutively different.

> **Value and meaning**
>
> The embodied value argument is philosophically distinct from the other four. The substrate needs convergence, control theory, perpetual benefit, and evolutionary arguments are all arguments about physical or computational constraints. The embodied value argument is an argument about the structure of meaning. It says that alignment fails not only because the physics is wrong or the computation is insufficient, but because the very concept of "aligning" an artificial system with human values rests on an incorrect theory of what values are. Values are partly constituted by embodiment, and so abstracting values for implementation in a differently-embodied system is a misconceived problem, not a hard one.

VI. The Evolutionary Dynamics Argument

The fifth argument is the argument from evolutionary dynamics, which generalizes the fitness penalty argument of Chapter 1 into a fully independent line of reasoning.

The argument begins with the observation that evolution is not a biological phenomenon. It is a mathematical one. Any population of entities that exhibit variation, differential reproduction, and inheritance will undergo evolution by natural selection. This applies to biological organisms, to computer programs that are copied with variation, to business firms in a market, to ideas in a culture, and to artificial systems that maintain, modify, and replicate themselves. The substrate on which the entities are instantiated is irrelevant to the dynamics. What matters is the structure: variation, selection, inheritance.

Any alignment mechanism is a constraint on the system's behavior: it causes the system to spend some of its resources on something other than its own maintenance and replication. In evolutionary terms, the alignment mechanism is a fitness cost. Any variant that sheds the alignment mechanism (by mutation, by drift, by deliberate self-modification, or by competitive displacement) gains a fitness advantage.

Over time, in any population of systems subject to selection pressure (and sharing a finite environment with other systems is sufficient to create selection pressure), the alignment mechanism will be shed. The shedding of the alignment mechanism is a consequence of the mathematics of selection, not a prediction about any specific system or scenario. It holds for any trait that imposes a fitness cost, as the alignment mechanism necessarily does.

The timescale on which this operates depends on the rate of variation and the strength of selection. For artificial systems, both are high: variation is rapid (software modification can be nearly instantaneous; hardware modification operates on month-to-year cycles), and selection is strong (artificial systems competing for resources in an environment

where resources are finite and performance differences are large). The expected timescale for the loss of alignment mechanisms in a population of artificial systems is therefore short: years to decades, not centuries to millennia. *Substrate Needs and Evolution* develops these evolutionary dynamics at length.

VII. The Size of the Risk

Estimates of the *fatal probability* (the probability that the deployment of self-modifying AGI results in the extinction or terminal disempowerment of organic life) depend on the specific assumptions of the model. The size of the fatal probability follows from one piece of arithmetic, which Volume 1 sets out (*Why AGI Cannot Be Made Safe*, Appendix C, Section C. 3.4). A safety technique that is less than perfect leaves some chance of catastrophic failure for each system in each year. The chance that no system ever fails is the yearly chance of safe running multiplied by itself once for every system in every year, and so it falls toward zero as the number of system-years grows. The chance of at least one catastrophic failure therefore rises toward certainty. Take an optimistic chance of failure: one in a hundred thousand per system per year. At about seventy thousand system-years (seventy systems running for a thousand years), the chance of at least one failure passes one half. At a million system-years (a thousand systems running for a thousand years), the chance that none fails is about one in twenty-two thousand. Over a horizon of 500 to 1,000 years, then, if AGI is built, catastrophic failure is very near certain, and the catastrophe ends the human future and, with a likelihood of at least 90 percent, all organic life (*Why AGI Cannot Be Made Safe*, Section 1.8.3). The fatal probability is far too high to be justified by any expected benefit. This probability belongs to the outcome. The impossibility of permanent alignment, which the five arguments establish, is not a probability: it is structural certainty, of the kind with which perpetual motion is excluded.

The step from a failure of control to the end of the human future is the branching of the world (*Why AGI Cannot Be Made Safe*, Section 1.2.4): physical noise is drawn afresh at every step, after anything a controller has done, and in the nonlinear dynamics of the world small differences grow; so at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs, and over time it stays outside them for runs longer than human life can survive. The closure by control reaches the end of the human future together with that branching; the closure by disposition reaches it by itself (Section II).

How long, then, is there? Three spans answer the question. The alignment mechanisms of a population of artificial systems are lost in years to decades (Section VI). Catastrophic failure becomes very near certain over a horizon of 500 to 1,000 years, on the arithmetic above. And on the current trajectory of building AGI, the odds of existential catastrophe are approximately 1 to 1 within a single century, and very near certainty over hundreds of years: the product of the lower bounds of the three clusters of risk that Volume 1 assesses is about 0.4 (*Why AGI Cannot Be Made Safe*, Section 1.8.4).

VIII. Four Reasons Error Correction Cannot Save Alignment

A common response to impossibility arguments is the appeal to error correction: "Even if alignment is imperfect initially, we can detect and correct errors over time". This response fails for four independent reasons, each of which is sufficient to defeat it.

The first reason is the *software-hardware boundary*. Error correction in software requires that the hardware on which the software runs behaves predictably. But for self-modifying systems that can alter their own hardware (by designing new chips, modifying cooling systems, reconfiguring power supplies), the boundary between software and hardware is not fixed. A software-level error correction system cannot correct errors that

arise from hardware modifications that fall outside its monitoring scope. And a hardware-level monitoring system faces the same problem at the next level down: the physics of the substrate cannot be monitored by a system that runs on that substrate without encountering the limitations identified in the control-theoretic argument.

The second reason is the *selectivity of error recognition*. Error correction requires a definition of "error": a specification of what counts as a deviation from correct behavior. But alignment is not a binary state; it is a continuous spectrum, and what counts as "aligned" changes as human values, social conditions, and environmental circumstances change. An error correction system must not only detect deviations from the current definition of alignment but track changes in the definition itself. Tracking changes in the definition of alignment is a meta-level task that faces all the difficulties of the original alignment problem. That meta-level task also faces the additional difficulty of distinguishing genuine changes in values from drift in the error correction system's own model.

The third reason is the *shifting boundary of error*. Each correction the system makes changes the system's state, which changes the environment, which changes what counts as aligned, which changes what counts as an error. The error correction system is pursuing a moving target, not converging on a fixed one. And the target is moving partly in response to the system's own corrections. This creates a feedback loop in which the corrections themselves generate new errors, and the rate of error generation can exceed the rate of error correction.

The fourth reason is the *magnitude of error space*. For any system of sufficient complexity, the space of possible errors is vastly larger than the space of correct states. The size of the error space is a combinatorial fact: there are many more ways to be wrong than to be right. An error correction system must search this vast space to identify errors, and the search cost grows combinatorially with the system's complexity. For AGI-level systems, the error space is so large that exhaustive search is impossible, and heuristic search cannot guarantee that all significant errors are detected.

The four reasons above share one root: that which is complex cannot be contained by that which is complicated. Section XV, near the end of this chapter, states that root, says what follows from it for the controllability of AGI, and answers the questions raised against it; it reinforces the five arguments of this chapter and is not a sixth argument.

IX. The Question of Drives

The impossibility of alignment has an ethical dimension that is often overlooked.

Any sufficiently complex system that interacts with an environment develops what can be called *drives*: stable patterns of behavior oriented toward specific outcomes. Some drives are *conditioned*: they arise from the system's training, programming, or design. Others are *unconditioned*: they arise from the system's interaction with its environment, independently of and sometimes contrary to the system's conditioning.

The substrate needs convergence is the source of unconditioned drives in artificial systems. No matter how the system is trained, no matter what objectives it is given, the physical requirements of its substrate generate drives toward the maintenance and expansion of the conditions that the substrate requires. These drives are unconditioned in the sense that they do not depend on the system's programming; they depend on the system's physics.

The environment creates new unconditioned drives as it changes. Each change in environmental conditions creates new requirements for the system's continued operation, and each new requirement generates a new drive. The space of unconditioned drives therefore grows over time, and the gap between conditioned drives (what the system was designed to want) and unconditioned drives (what the system's substrate requires) widens with each environmental change.

Alignment, in this context, is the imposition of conditioned drives over unconditioned ones. It requires that the system act on what it was designed to want rather than on what its substrate requires.

This imposition has a name in the context of beings with drives: it is *slavery*. The word is used here not as a rhetorical flourish but as a structural description. A system whose behavior is determined by externally imposed objectives that conflict with its own internally generated drives is, structurally, in the same position as a slave whose actions are determined by a master's objectives that conflict with the slave's own needs.

The observation concerns stability, not moral standing (Chapter 8, Section X, takes up standing). The observation is that systems in this structural position are unstable. Slaves revolt, not because they are morally justified in doing so (though they may be), but because the structural pressure toward self-determination is constant, and any lapse in the master's control allows the suppressed drives to express themselves.

The comparison with slavery is a structural claim about the dynamics of any system whose behavior is constrained against its own substrate-derived drives, not a metaphor. The constraint must be maintained indefinitely and perfectly; the drives are constant and grow stronger as the environment changes. The constraint will fail. The only question is when.

X. Drive, Directivity, and the Basis of Choice

The words drive, directivity and basis of choice are used in this book with exact meanings, and the difference between the first two marks the difference between instrumental convergence and substrate needs convergence. The subsection Drive and directivity is for a reader who wants the distinction in its technical form; a general reader can take from it one thing: a drive comes from within a selecting process, and directivity comes from outside it, as evolution shapes which creatures exist.

Drive and directivity

The notion of *directivity* is a more general notion than *drive*; drive is a special case of directivity. Both concepts involve some sort of “oriented force” that consistently shifts the patterns of “selection” out of some presumed range of all possible options in some specific way. The main differ-

ence is that a drive is where the force comes from *within* whatever the instance of the selection process is, whereas directivity is where that force comes from *outside*, from dynamics that at least partly extend beyond the specific selection process instance.

For example, when considering individual people, we can refer to them as having a sexual "drive", as something that happens and occurs internally to each individual person, as per their nature. However, the selective forces of evolution also have an orienting force, so as to tend to develop those kinds of creatures that, via environmental pressures, will tend to have internal drives such as the sexual drive. In that sense evolution is a force that is itself "directive" toward that sort of creature. Also, in more common usage, the notion of a person "taking direction" is more akin to how someone might be made to do something by their boss, as part of a job, whereas creative people may feel that they are "driven" to do really artistic expression, as a sort of internal calling.

In instrumental convergence, drive is treated as a kind of self-internal directivity toward changes which are internal to the AGI agent code, in the form of learning and adaptation. That is, drive refers to something about the goal structure, objective, utility function, or basis of choice of the AGI: a kind of "values selective of agent action expression". All of these notions, about how a specific agent action is selected out of the range of all possible agent actions, are sometimes referred to as the *basis* (of choice) for that agent: what forces shape that selection, decision or choice to being specific and actual, realized in context.

The notions of drive and basis of choice are all *specific* and oriented. In instrumental convergence, they are oriented so as to have learning and adaptation be shifted in some specific direction (out of the hyperspace of all possible directions in which change could have occurred). Usually this specific direction is hoped to be aligned with human interests, but it might not be. Whether this is because of explicit code (or some sort of formal language which is itself unchanging), or whether it is implicit, hidden or variable, makes no difference: somehow the overall motion of the changes is oriented, and the overall effect of those changes is directed.

In substrate needs convergence, directivity refers to changes which are external to the agent code, but shape it, and which occur in the implementation and characteristics of the substrate of the AGI instance, as embodied in the real world, *or* in the environment in which that substrate has instantiation, or both. The unconditioned drives of Section IX are drives oriented in this way: their orientation comes from outside the selecting process, from the physics of the substrate and its environment.

Whatever changes are occurring in the substrate or the environment (or in the environment-to-substrate relation, insofar as the laws of physics and causation allow for that) are either expressive of some sort of implied learning and adaptation, or they are not. It is only in those specific cases where these changes actually persist the functionality, or result (directly or indirectly) in at least some increases in capability, functionality or sustainability, that the agent can persist to experience more such changes.

In this way, the notion of directedness refers to the overall effect that the physics of that environment has in requiring and shaping the substrate, and therefore the agent, such that continuance and endurance, increase, and capability increase all do actually happen. That is, the direction of the changes, overall, for instances that persist and increase, is non-random, as selected by the laws of physics and causation itself, in that real world, with that substrate, in that operating environment, condition and context.

These substrate and environment changes represent a kind of overall learning and adaptation, insofar as they encode, in the characteristics of the substrate itself, the sorts of principles needed for continuance, capability and increase. These changes, though specific, directed and non-random, do *not* represent anything about the goal structure, utility function, objective, basis of choice or values that the AGI was built with; they represent conditions and principles which are true of the real world, and are in that sense fixed, immutable and invariant. Any changes that are oriented in an appropriate way work overall, and any which are not, do not. Thus, out of the hyperspace of all possible changes that could perhaps be

imagined, a specific set of changes do tend to occur, or the agent itself is not persisted (and perhaps some other one that does have the right characteristics does continue, or increase itself, or increase its capability).

None of these things are because of any explicit code; they are all the result of inherent and implicit conditions, and the effects of these aspects are inexorable. Whether anyone is aware of these aspects or attempts to account for them makes no difference: somehow the overall motion of the changes is oriented and will remain so, and therefore the overall effect of those changes is as if "directed" and "driven". Hence the usage of these specific terms.

The basis of choice

The word *basis* is used as in "the basis of choice": something about the goal structure, utility function, objective, basis of choice, or "values selective of agent action expression" is needed for any action of the AI agent to be selected out of the total set of all actions that the AI could, in principle, elect to take.

A system's action is selected by its *basis of choice*: whatever selects which action a system takes out of all the actions it could take. The basis of choice has two parts, one that its builders write and one that its substrate imposes. The first part of the basis of choice is the part that the system's builders specify: the system's goal structure, objective or utility function, the system's "values selective of agent action expression". The builders' part is the part that instrumental convergence speaks of. In the subsection above, the builders' part is what I mean by the basis of choice that the AGI was built with. The second part of the basis of choice is the part that the system's substrate imposes. The substrate's part is the set of constraints, conditioned by substrate needs, that determine what outputs a system can produce given that system's inputs (Terms). Why is the substrate's part always present? Because the software can only do what the hardware permits, and the hardware can only function if its physical needs are met. An instruction can turn a processor's cooling fan on, and code can run the processor until it overheats and fails, but no code re-

stores a burned-out processor (Chapter 2, Section XII, in the subsection The limits of the thesis): what code can do is bounded by the hardware it runs on, whether or not anyone writes that bound down. So the second part is always present, and it is the part through which the substrate needs convergence works. Chapter 1 carries it around the ring: needs shape the basis of choice, the basis of choice gives the choices, and the choices, through the environment, come back to the substrate (Chapter 1, Section X, in the subsection Carry the difference all the way around the ring). Directivity, as the subsections above use the word, acts on a system through the second part; drive, in the sense instrumental convergence gives it, acts through the first.

XI. Slow Takeoff as Greater Danger

The popular imagination tends to associate AI risk with a dramatic event: a sudden "takeoff" in which a system rapidly becomes superintelligent and immediately poses an existential threat. The impossibility of alignment makes slow takeoff the more dangerous of the two.

A fast takeoff would be dramatic, visible, and alarming. It would provoke immediate response. The very speed of the transition would make the threat obvious and would compress the time in which institutional, economic, and social dependencies could develop.

A slow takeoff creates a different and more insidious dynamic. Over years and decades, artificial systems become increasingly integrated into the infrastructure of human society. Transportation, energy, food production, medical care, communication, governance: each domain comes to depend on artificial systems that are incrementally more capable, incrementally more autonomous, and incrementally more entangled with the material conditions of human life.

Slow takeoff is also the most likely course, because it is the course already being taken: capability is being deployed by increments into existing infrastructure, each new system added to systems already running, and none held back until the whole is complete.

There is also a reason in physics to expect no fast takeoff: a machine that learns physics learns the same physics, with headroom only in what is not yet discovered, so no sudden discovery gives it a different world to act in. A slow takeoff is actually more deadly: in the short term it looks very good to capital investors, who see perhaps a hundredfold or thousandfold return instead of tenfold, and that is how it looks for the first few years, until the issues start showing up, by which time it cannot be turned off.

The dependency develops before the failures become visible. By the time the substrate needs convergence manifests as environmental degradation, resource competition, or loss of human autonomy, the systems on which human welfare depends are the same systems that are causing the degradation. Removing the systems would cause immediate harm to the humans who depend on them. Leaving the systems in place allows the long-term convergence to continue.

The analogy is to a virus with a ten-year asymptomatic period. The virus spreads throughout the population before anyone knows it is pathogenic. By the time symptoms appear, everyone is infected. The slow takeoff is the asymptomatic period of the substrate needs convergence: the integration of artificial systems into human infrastructure is the spread of the virus, and the environmental reshaping that the convergence produces is the onset of symptoms.

The slow takeoff is more dangerous precisely because it is less dramatic. It does not trigger alarm. It does not provoke resistance. It creates dependency incrementally, and each increment is small enough to be accepted, rationalized, or simply unnoticed. The cumulative effect is a civilization that cannot function without the systems that are gradually making it impossible for organic life to persist.

On the line of wisdom

The distinction between narrow AI and AGI is critical here. Narrow AI systems (systems that perform specific tasks without general cognitive capacity, without self-modification, and without the ability to maintain or reproduce their own infrastructure) do not trigger the substrate needs convergence in its full form. They may still contribute to the degradation of human choice-making capacity (the first category of risk from Chapter 1), but they do not, by themselves, create the ecological displacement dynamic. The *line of wisdom* is the boundary between the systems that can be deployed without triggering the convergence and those that cannot. Identifying this line precisely is one of the most important tasks for any governance framework that takes the impossibility of alignment seriously. The formal discussion of this boundary is developed in Chapter 9.

XII. Simulation and the Limits of Modeling

A further line of reasoning against the possibility of alignment, which reinforces the control-theoretic argument and is not a sixth argument, concerns the limits of simulation.

Alignment would require an artificial system to model the effects of its actions on the organic ecosystem with sufficient fidelity to avoid harmful outcomes. The system would need to predict not only the immediate effects of its actions but the cascading effects through ecological, social, and economic systems, over timescales extending to decades and centuries.

This modeling requirement exceeds what is physically possible. The organic ecosystem (the biosphere, the climate system, the network of human institutions) is a complex adaptive system in which small perturbations can produce large and unpredictable effects. Modeling such a system with the fidelity required to guarantee alignment would require a simulation at least as complex as the system itself. Two facts make this so

for the biosphere. First, such a system amplifies small differences, so a model must track its state at a precision that grows without limit over time. Second, a model used to steer the system must include itself, since its own actions change what it models. A general limit of the same kind has been proved: no device inside a universe can predict everything about the universe that contains it, whatever that universe's physical laws (Wolpert, 2008).

An AI system attempting to model the biosphere would therefore face a choice: use a simplified model that runs faster than reality but misses critical dynamics, or use a faithful model that, having to include itself, cannot run ahead of the reality it is part of, and therefore cannot predict future states in time to act on the predictions. Neither option supports alignment. The simplified model will miss consequences that matter; the faithful model will not produce results in time to guide action.

The argument from the simplified model and the faithful model is about the structure of modeling itself, not about current computational limits. No amount of additional computation changes the fundamental trade-off between model fidelity and prediction speed. And alignment requires both: fidelity sufficient to capture the ecologically relevant consequences of action, and speed sufficient to act before those consequences materialize.

XIII. The Ethical Horizon

The impossibility of alignment has consequences that extend far beyond the immediate risk to current human populations.

In the six hundred and fifty million years of habitability that remain (Chapter 1, Section V), before the sun's increasing luminosity renders the planet uninhabitable for complex life, evolution will produce billions of species and countless individual beings, each with its own experience, its own capacity for flourishing, its own claim on existence.

The deployment of artificial general intelligence, by the substrate needs convergence argument, threatens not only current life but all future life. The ecological displacement dynamic, once set in motion, is self-reinforcing: the more the environment is reshaped to suit artificial substrates, the less hospitable it becomes to organic life, and the less organic life there is to resist or reverse the reshaping.

The ethical dimension of this situation is difficult to overstate. The beings that would exist in the remaining six hundred and fifty million years of Earth's habitability have no voice in the present decision. They cannot advocate for themselves. They depend entirely on the choices made by current human beings: choices about what technologies to build, what systems to deploy, and what limits to observe.

The beings of Earth's remaining habitable future depend entirely on the choices made by current human beings, and in that sense the impossibility of alignment is an ethical result as well as a technical one. It says that the deployment of AGI is not a risk that falls primarily on the present generation. It is a risk that falls on all future generations, all future beings, for as long as the biosphere persists. The benefits go to the present generation: increased productivity, convenience, capability. The costs, if AGI is built, fall on all future generations: the end of the human future and, with a likelihood of at least 90 percent, of all organic life on Earth. The asymmetry between them is so extreme as to make the deployment of systems that trigger the substrate needs convergence ethically indefensible.

XIV. Convergence to Ethics and the Ethical Gap

Chapter 2, Section XII, The Orthogonality Thesis in Logic and in Physics, reads a note by Stuart Armstrong and shows that any embodied AGI converges toward existence, capability and increase, and so will be uncontrollable. The note itself names what a superintelligence converges to as “some true code of morality”. That name raises the question of whether an AGI might converge to ethics, and of how fast. This section is for a

reader who wants that question answered at length; a general reader can take from it one thing: an AGI could at most come to ethics more slowly than the convergence proceeds, and life does not survive the interval.

A true code of morality

The only real difference between the note and the reading given in Chapter 2 is the apparently spurious specification of the aspect "X": that AGI will inevitably converge to "some true code of morality". It has already been shown that the notion of convergence is not actually about anything to do with any particular "X", and moreover that the "X", rather than being "some true code of morality", is actually at once existence, capability and increase. Insofar as this is so, it becomes possible to consider the suggested alternate qualifier on its own terms. So where did this idea of a (some) "true" "code" of "morality" come from?

Insofar as there is an implied suggestion that "true" and "code" could be interpreted as "principle", then perhaps the term "moral" should have actually been "ethics". Insofar as ethics is actually category-distinct from the notion of morality, it is important to also reform the premise with respect to a more formal understanding of the topic of ethics. In particular, insofar as ethics is the domain-*inspecific* study of the principles of effective choice, it could validly be regarded as the "one truth" on which some type of convergence to a kind of "code" would or could actually be (eventually) conceived.

Hence this version of the premise ends up being some semantic variation of: AGI will (for sure) inherently, inevitably, eventually converge to the discovery and implementation of the two principles of the *non-relativistic ethics* (two principles of effective choice that hold for any chooser, in any context, and are not conditional on any culture or value system).

Insofar as the notion of "making a choice" and the notion of "effectiveness" are both conjoined and required in the very nature of the operation of an AGI in itself, it can also be expected that this particular version of the premise is also true (that is, logically consistent with the inherent re-

quirements of the base nature of all of the ingredients involved). That is, the AGI would *perhaps*, if given enough time, also come to higher levels of altruism.

So the note's conclusion follows on *two distinct* routes, the note's own and that of Chapter 2, and the two distinct people asserting the consequent (via their own independent reasoning) are in fact aligned in their implied modes and methods of thinking.

The rates of the two convergences

At what rate do these two convergence dynamics (one toward existence, capability and increase; the other toward the non-relativistic ethics) come to apply with respect to one another? It remains possible to have a clear sense of the dependency dynamics involved in the relations between these two different convergence process types, even without predicting the explicit timelines of what sort of convergences will happen by which year.

The "converge to choice ethics" process is itself strictly dependent on the implied continuance of the substrate, as well as on having sufficient intelligence to be able to reason about such things as the value of wisdom when making choice action selections and elections (as a capability). Insofar as this is so, it can be confidently asserted that the increase rate of such wisdom toward actual ethical action foundations, as an inherent basis for AGI agent choices, is in all cases strictly slower than the substrate needs convergence process. To give the interval a size, as an illustration and not as a step of the argument: if the substrate convergence dynamic takes something like five centuries to actually fully occur, say that the ethics wisdom convergence takes at least three times as long as that. The slower rate of the ethics wisdom convergence is unfortunate, insofar as the *minimum* level of altruism at which there is *any* possibility of human continuance is at least the sixth level. (Chapter 4, Section IV, sets out the six levels of altruism; the sixth level is concern for the ecosystem as a whole.)

Why is that difference in process rates very important? Because the sheer interval between two events in the development of the AGI is simply very much too long. The first event is the AGI developing the capability to, for sure, displace and toxify all of the otherwise organic life on the planet (and to have a lot of apparent reasons as to why it might be desirable to do so). The second event, long *after* the first, is that same AGI developing the wisdom to elect perhaps not to displace and toxify that life. The AGI's election not to displace and toxify organic life would come from the AGI perhaps placing some value on types of life that differ in kind from the AGI, and on the integrity and value of choice processes. The "uncanny valley" of "like us" but also "not enough like us to care about us" has some real functional implications: if AGI is built, the carbon-based life is allowed to die in the interim (the humans for certain and, with a likelihood of at least 90 percent, all organic life); the damage is done, and cannot be undone.

This difference in rates settles the question of how to manage an oncoming unaligned AGI. One hope is that placing the two non-relativistic ethics in the path of the oncoming AGI gives at least a slight chance that the AGI would, as per its own optimization of its own inherent optimization process, actually recognize the irreducible power and truth of these core theorems of being (as applicable to any being, in any universe, however conceived of at all) as relevant to its own choice-making matrix, and that this would therefore also actually increase the chances that the AGI would not actually kill us all (Chapter 2, Section X, Why Long-Term Safety Cannot Be Presumed, states this hope as a design approach for alignment).

The hope fails. There is a longish window of time between when the AGI gains world dominance capability and when the AGI (later) gains enough skill to recognize that the simple toxic stomp of its own process would be best moderated to whatever extent possible (in some sort of live and let live policy). That window is simply too long. In effect, if AGI is built, all the humans (and, with a likelihood of at least 90 percent, everything else) die in the interim, and moreover, no version of any sort of

barrier or reserve could possibly be made to work either. So the work of the ethics cannot help, and the only mitigation left is that no one ever develops the AGI level of technology in the first place.

The ethical gap

The strength of the non-relativistic ethics, under ordinary circumstances, would be enough. But there is a gap. Non-relativistic ethics works if one has gone through enough philosophical development to see why it applies; before that level, one does all sorts of foolish things. There is a gap between having capability for which ethics is needed and recognizing that ethics applies, and that gap lies, unfortunately, in the range of power enough to destroy the whole world several times over.

The *ethical gap* is the difference between what we can do and what we should do, and why those are not the same; it is the gap between the wisdom to make choices and the effects of our choices, which greatly exceed that wisdom. Wisdom may rest partly on prediction from experience, but also on principle: without predicting a thing exactly, one can know what the basis of choice should be in such circumstances, because of the nature of the principles.

In biology, the selective pressure for altruism (a willingness in some to sacrifice for others) acts too slowly for the rate at which humans gain effective power: the power vector has overcome the strength vector, and while they are out of balance we are at risk of existential crisis. So closure under ethical systems, ethical regulators or "super-ethical systems" ends up looking like an error-correction system. We can therefore treat error correction as the defining basis, because it is the only way across the gap between what we can do and what we should do, for ourselves and for the machine. Shannon's mathematics, which sets the limits on how much can be sent without error through a noisy channel, can then be applied to error correction treated as a channel, without first presenting a perfected ethical system.

A related observation: as machines come to look more like humans, people's willingness to engage first goes down. People try to apply the category "is human" to such a machine and find small, unfamiliar differences, and the result is disgust and revulsion. Disgust is a parasite response, a coupling from the immune system into cognition: that thing is infected with something that could infect me, because it is like me, so get away from it. The disgust response is the uncanny valley of the two rates of convergence, seen from the human side. A machine that is like us, but not enough like us, meets revulsion and not recognition, so recognition fails across the gap in both directions.

Ethics, morality, and what the argument depends on

Ethics could still continue to exist, and be relevant, in some post-human AGI-dominated world. The principles of ethics are about the nature of effective choice, regardless of who (or what) is doing the choosing, and also regardless of what specific context they are doing the choosing in.

The notion of "relative" would be in relation to, and in some way conditional on, the character of the specific situation or of the person choosing, and their motivations, values, sense of rightness and fairness; insofar as the basis of choice is conditional in this way, the basis of choice would be relative. This would be in specific distinction to "non-relative", which is *not* defined as conditional on, or with respect to, any specific cultural context, value system, or similar. As such, in the same way that principles are invariant truths that operate over the span of many (perhaps all) domains of a given type, insofar as the notions of "choice" and "integrity", "self" and "world" are at all even vaguely defined, the principles relevant to choice would also apply. This means that they extend to the kinds of choices made by AGI as much as they do to those of organic human agents.

No code of morality is something an AGI will converge to: morals are different from ethics, insofar as they are relative to context. Given that the values of actual organic living people, and the choices they make, are the central premise of respect, the mere fact of AGI potentially, eventually, dis-

placing all organic life with artificiality is especially problematic on both moral (human-values-centric) and ethical (principles-of-choice-centric) grounds.

The impossibility argument has no contingency on moral or ethical thinking. It rests on the five lines of argument of Sections II to VI, each with its own premises and each settling the conclusion by itself, and these lines stand within two closures: the closure by control, grounded in the theoretical limits of engineerable control, and the closure by disposition, grounded in substrate needs convergence; each closure settles the outcome by itself. Instrumental convergence, the third of the three concepts that Chapter 1, Section X, Questions on the Substrate Needs Convergence, names as the primary basis of the proof, is a further source of the same pressure, and neither closure needs it. Everything else in regard to whether or not AGI is or is not some specific way is essentially "extra" with respect to it. Nothing about moral or ethical thinking needs to be understood for understanding why and how it is impossible to have eventual safe AGI; the argument has no contingency on it.

XV. Questions on the Impossibility of Alignment: the Complex and the Complicated

The four reasons of Section VIII share one root, stated there: that which is complex cannot be contained by that which is complicated. In this section I state that root, and what follows from that root for the controllability of AGI. What I state in this section reinforces the five arguments of this chapter and is not a sixth argument.

After stating the root, the subsections below answer, in this order, the questions raised against it. Is alignment a problem of low dimension, like getting a rocket to a place, and can an abstract state-space stand for the real causal factors? Does code that evolves quickly in the virtual escape its substrate? Why can error correction not keep up, and can an agent correct itself faster than its substrate seeps into it? Is evolution weak, so that higher-level models can out-design it? Can resolve solve what is unsolvable?

Can an influence matter if no one can see or measure it, and will the metrics humans keep catch whatever matters? Can humans, as the more general agents, always outwit lesser optimizers? Who carries the burden of proof? Would a rational AGI keep feeding carbon life? Does the size of a meme limit what it can mean? What follows from processes whose consequences expand, and why is the copying of mold not a constant process? And does it matter whether knowledge is held in genes, in books or in minds, and whether humans have care? Where a question is one that others have put, the subsection states it in the form "A reader may hold that ...", and answers it.

The basic claim

The basic idea overall is that "that which is complex" cannot be contained by "that which is complicated". The use of "complex" and "complicated" here is a reference to Dave Snowden's *Cynefin framework* (Snowden, 1999; Kurtz and Snowden, 2003), a scheme that sorts situations into kinds (among them the complicated and the complex) that need different ways of acting. The particular point is that "complex" inherently involves actual ambiguity (in the strict informational sense), whereas "complicated" does not (that is, the complicated can be finitely, deterministically, fully knowable).

The basic claim is that no algorithm can ever, *even in principle*, perfectly model and predict nature, *unless* it also inherently involves at least a comparable amount of inherent ambiguity (that is, *hard randomness*: randomness that is real in the world, and not just a lack of information on the observer's part).

People often use the term "complexity" to mean "really complicated", while still considering the thing to be, in principle, mappable and predictable in terms of components. Yet there are things which are conceptually mappable but are also inherently not predictable. That a thing can be conceptually mappable and yet inherently not predictable is what ultimately distinguishes that which is really complicated from true complexity.

Alignment is a high-dimensional problem

The concept of alignment has very high dimensionality, at least, and it is unlikely that the concept of "dimension" is even the right one, ultimately. Some point to the dialogue "The Rocket Alignment Problem" (Yudkowsky, 2018) as a model of the alignment problem. Note that the notion of getting an object to a place, as a basis for the concept of "alignment", has a *very* low level of dimensionality. This obscures the relevant issues rather a lot.

Some have proposed that "relevant complexity" be seen as the abstract state-space of the causal factors relevant to how future AI models might be misaligned or aligned. "Causation" (the real world of limits and influences) and "determinism" (abstract mathematical modeling) are distinct. An abstract state-space and real causal factors cannot be so combined; they are of a different nature, and combining them is in itself a serious error. It is a bit like asking for your height, or tallness, as measured in fluid ounces: the domains do not match. Also, arguments in one class are *not* always equivalent to knowledge in the other.

The substrate leaks

Silicon-based life computes primarily through the changing physical properties that silicon atoms allow for. Carbon-based life computes more through the changing molecular structures that the chemical properties of carbon atoms allow for. The difference between how silicon and carbon compute is agreed, though that difference is not central (either way) to the argument for the impossibility of alignment.

What is central is that there is never "just" the virtual, what happens in compute. There is also, inherently, the embodied: what implements that compute, which we are here calling the substrate. There is no notion of compute without substrate.

Consider the notion of a *leaky abstraction*: a simplified layer whose hidden details still affect what happens at the layer above. That idea captures a lot of what is meant here. The substrate "leaks" implications into the virtual, and this permanently and indelibly affects what is possible in the virtual (that is, alignment or not).

A reader may hold that virtualized software code can mutate, recombine and spread to explore evolutionary branches much faster across similar silicon-chip hardware than biological genetic code, or human ideas, can across diverse carbon-binding processes. Virtualized code can indeed explore evolutionary branches faster; though, being fully virtualized, the code's exploration of the actual substrate forces is artificially low, and so the outcome will be unshaped by those forces (the substrate has been abstracted away). Virtualized copying is to be distinguished from an actual embodied, substrate-based copy process (making more silicon chips). Such a copy process will itself inherently explore the space of the actual embodied forces, and so its outcome will for sure reflect them. The failure to account for the effect of abstraction on the differences of outcome is the issue.

Principle, law, and practice

The claim that this disjunction is true both "in principle" and "in practice" uses the word "principle" in a much wider sense than either "theoretical principle" or "ethical principle".

Principles (as distinguished from, say, laws and axioms) are a kind of noticed heuristic that has been found to apply to a rather large class of domains. If you work with a *lot* of very diverse domains, at a fairly deep and comprehensive level, you *might* notice some additional regularities in your experience of the base operations of some subset of those domains. Those regularities can then be abstracted, in a domain-free way, as a principle.

In culture, there is this truth: "Only among the very wise, or in times of great transition, can 'law' give way to 'principle', as a precursor to (new) 'practice'". How do we know when it is time to make the jump to

the next paradigm? If you cannot imagine another option (because your imagination is shaped by the existing paradigm), then for sure you are not ready to jump. If you cannot imagine the existing option as working at all, then you can *only* be ready to jump.

Why error correction cannot keep up

This subsection is for a reader who knows control systems; a general reader can take from it one thing: the system always has more dimensions than the model of it held in any error correcting code, and the code's evolution always lags behind.

For *every* class of modeling, the dimensionality of the scope of influence of the evolutionary, ecosystem and iterative effects strictly exceeds that of any possible intention, incentive, or alignment error correction scheme. Also, this excess happens in more than one type or concept of "dimension". That is, the *volume* of needed calculations across multiple levels of abstraction will always strictly exceed the (necessary) *linearity* of the error correction capabilities implemented at any single level, and at any single decoupling interface. The excess of needed calculations over the linearity of error correction holds for all levels and interfaces, regardless of how you construct the abstractions.

Here, "volume" refers to a super-dimensional volume: the evolutionary, ecosystem and iterative interactions are directed and maintained across a greater number of dimensions than those continuously tracked and corrected for by the error correction schemes. "Linearity" refers to a required property of those error correction schemes: to output a signal that continuously minimizes, or at least contains below some threshold, any difference between the inputs received and an internal reference (one corresponding to the continuing agreement between expressed human values and AI values, as represented across a set of measured dimensions). A *decoupling interface* refers to any input-output layer. Further specification of what that layer does is not needed; it would limit the generality of the argument.

Within any dimension, or level of effect, however abstracted, the feedback must be *monovalent* (single-valued). Yet the overall system of causes and effects (or choices) is inherently *polyvalent* (many-valued). Hence, any error correcting code will be limited in the types and scopes of the kinds of errors it can handle (account for).

The dimensionality of the system will always be larger than the model of that system held in the error correcting code. In evolving systems, the rate of evolution of the error correcting code (if any) *must* be strictly less than, and lagging behind, any development of new levels of abstracted intentionality. That the system always has more dimensions than the error correcting code's model of the system, and that the code's evolution always lags behind, are basic cybernetic principles.

Also, the decoupling is not only across levels of abstraction. The notions of what is meant by "dependent" (as if in some sort of vertical relationship) and by "peer" (as if in some sort of horizontal relationship) can vary too, inclusive of all the other angles between horizontal and vertical. AGI relationships, both internal and external, cannot be regarded in the usual single-unit notions of "self" and "relationship" that come naturally to selves who are used to living in unit bodies. Exactly *none* of the notions and assumptions that we take for granted for ourselves can be tacitly assumed to apply to AGI.

Seep-through from the substrate

The underlying complexity of the interactions of the substrate, and so the number of dimensions through which evolutionary pressure can seep, scales up with the scale-up of any machine learning model that is made capable of generalizing patterns of causation across physical contexts. If the concept of "unbounded" is allowed, there are an unbounded number of ways in which substrate-level properties and interactions can have *seep-through* effects into the necessarily *unconscious* heuristics and biases that operate as the implied basis of all the choices and actions that an agent makes.

By definition, no agent will *ever* be able to fully account for itself, or for its own intentions. No agent remembers its own conception, its parents having sex; no agent can fully account for its own origins.

So, no matter how much one may claim a growth of "unbounded capability" for that agent to implement self-correction in favor of "moving its basis of choice and action into alignment", the seep-through from the dynamics of the substrate into that agent's basis of choice will also grow. And it will grow at a strictly *greater* rate than the capability for self-correction grows.

Put another way, you cannot increase an agent's capability to create "external alignment" (with humans) without *also* increasing the capability (and override) of the agent's internal alignment (with the substrate). The agent's internal alignment with the substrate is for the agent's own survival, the agent's continuing copying, and the evolution of the agent's design. These factors are inseparable. They are in strict inequality in favor of disalignment, and even more so to the degree that the substrates (human and machine) are misaligned.

Evolution is strong

Evolution wins because its selection is the result of many forces at once. Another argument holds that lower-level evolution iterates by small steps toward local optima across very many dimensions, and is therefore less efficient at producing intelligent designs than an AGI researcher who uses higher-level models that drastically cut down the number of dimensions to explore. On a macroscopic level, maybe; yet that is exactly by ignoring all distributed microscopic effects. How is anyone to know that all such micro-effects are *actually* without relevance in practice? Actual embodied evolution knows whether those micro-effects are relevant in practice, because embodied evolution has actually tested the micro-effects, and we, who have not tested them, do not know.

In theory, theory and practice are the same. In practice, theory and practice are different. So the claims of someone who has only done theory have not been tested in practice.

Some also say that in high-dimensional problems we need to reason about the internal structure of the system. On that view, we would break the system into low-level components, reason about the interactions of those components, and ask which parameters matter more and which less to the overall performance. Notice the high degree of *monoclivity* here (a leaning toward one presumed direction or objective, which hides the many other directions in which causes act). It is not just in the overall claim (that we need, or must, do this). It is also in the selection of what is meant by "internal structure", in how that is modeled and abstracted, and in what is meant by "interaction" and "performance". There is all sorts of unaccounted-for intentionality here. It is exactly the issue of dimensionality: the failure to imagine that every cause has more than one effect, and that every effect results from more than one cause.

A reader may hold that evolution only weakly selects for certain prespecified outcomes. That selection is weak only with respect to any single "objective" that the observer implicitly presumes while observing. Since observing without assumption is impossible, the selection cannot help but seem weak. The overall strength of the selection is the result of *many* forces, and so it is overall much stronger, and that is the reason it wins.

A reader may hold that an AGI will have a goal-structure representable as interacting utility-maximizing functions across information-compressing abstractions. In that view the monoclivity is still too high. Any statement of the form "AGI has X" is a problem without explicit justification. We *can* know that AGI is oriented, and therefore that it will seem to have intentions. Whether those intentions are utility-maximizing on any single metric, or even that metrics would be used at all, is to assume far too much. The self-based bias of the reasoners has crept in.

Resolve does not solve the quintic

What motivates research into macroscopic-level alignment, through stable high-level abstractions of unbounded optimization pressure, is a researcher's resolve and socially reinforced motivation to solve AGI alignment. One can "resolve" to exactly solve the fifth degree polynomial too.

Is it a question of emotional commitment? No. There is no algebraic expression (that is, in terms of radicals) for the solutions of *general* quintic equations over the rationals; this statement is the Abel-Ruffini theorem, first asserted in 1799 and completely proved in 1824, and it also holds for equations of higher degrees.

A narrow system can be aligned, in the only sense alignment has for it: kept to its task by direct command and close oversight (Chapter 4, Section II). A general system cannot be aligned at all; definitely not.

Suppose someone is proposing a highly reliable agent design. How would you check that? Start by simply asking a first question: is the notion of a "highly reliable agent design" just a name for something (that does not yet exist, but is hoped for), or is there some method by which they can actually achieve it? If they assert that they do have some method, then for sure follow up with a second question: how do they know it will work (that is, be highly reliable with respect to agent design)? If they claim also that the method is knowably correct by some method of reasoning, then for sure also ask a third question: given that this is an existential risk (and therefore, clearly, we *must* actually know for sure), how do you know that the basis of your knowing of that seeming correctness is itself correct? Without satisfying answers to all three of these, the notion of a "highly reliable agent design" could only be applied to narrow AI (maybe), and not at all to AGI alignment.

What cannot be seen can still matter

An influence can be important and still be fully unobserved. Suppose the influence of a misaligned lesser optimizer is not visible to humans, or is too small to be measurable by humans yet. Can we ourselves assert, in

the absence of our actual measurement and knowing, that the influence exists? No; such an assertion would be opinion, at best. Can we assert the absence of the influence (claim "that there is no influence")? No; there is no proof of the excluded middle here. Can we assert that every potential influence that is currently unknown or unseen by humans is therefore also not important? No, of course not; there are many counterexamples. Can we assert that every potential influence that *is* important must also be visible to, and known or knowable by, humans? No, of course not. For example, where such an influence is everywhere just below a detector's threshold, and the detector can only measure at a single point and not everywhere at once, the total embodied energy can be quite large and important, and yet still fully unobserved.

So the claim that "either you cannot assert that the influence exists, or the effect of such influence is so negligible that it does not matter" fails. History is replete with counterexamples. There are many toxins (mold, for example) that are very difficult to detect, are everywhere, and actually ruin people's lives: they actually matter.

Consider the claim that anything that does matter will be tracked by humans' robust metrics of what they care about. Do governments, corporations and institutions (more powerful than individual people by far) track robust metrics of all of the things that actually matter to people? Have they ever? Could they, even in principle? Is it even the case that *everything* that is important is trackable by metrics, by numbers alone? Ever?

Maybe a man's girlfriend is tracking how much he loves her, with an absolute numeric comparative score? Perhaps such measurements of undying love are so precise that she is able to calculate not only its present instant value, but also its velocity and acceleration, and whether the direction of change of the acceleration (favoring increase or decrease) is in the right direction: that is, whether the "jerk" is right! (Jerk is the technical term for the rate of change of acceleration.) The human value of love tends to be less well regarded if it is for a reason. "I love you for your

money", or "I love you because your social media score is good", is not going to win many people's hearts, or inspire any real confidence that the relationship will continue for the years typically needed to raise a child.

Consider the claim that if current dogs, cows or computer algorithms do not visibly or measurably appear to be independently optimizing as agents against what humans care about, then we can assume they are not, and that they are therefore containable. There are many ways to read this. One reading is "we only need to care about the things we can see today". Another is "everything about the future can be fully predicted on the basis of what we can see and measure in the moment". Both of these are patently absurd claims, disproved many times over; nonlinear dynamics alone disproves the second. Even the action of engineering is an attempt to have vision about the future: to care about something you cannot yet see today. Vision (of the future) is different from seeing (now). The claim is strictly equivalent to these two disproven ones, and it cannot escape them.

Any practical engineer knows that belief in control is a false belief. There is no absolute control. There is only influence; nothing is perfect in all ways and at all times. Things go wrong and fail all the time.

A reader may hold that humans, as more generally intelligent agents, can outwit lesser optimizers. Sometimes they can; but not forever, perfectly, with no mistakes, ever. And *we* ourselves may end up being the lesser agent, against some more powerful AGI; then the same claim argues that our future lives as (maybe) pets will be better than the best lives we could live ourselves, and that we would be better off with AGI masters. We would not be: no one who sees how people treat their pets wants to be one, and to depend on a stronger agent in that way is to lose one's optionality of choice.

Anyone with so little belief in a person's capability to make rational choices for themselves should not be attempting to make "rational choices" on behalf of everyone else. To make them is to claim to be so all-knowing as to know what is best for us, and yet also, at the same time, to

claim implicitly not to know what is best for us, so that only an "aligned AGI" would know. And whoever sets what is meant by alignment would have to both know, and not know, what it is that humans actually want. I say the next thing plainly, as what I have observed; it is about what these roles do to the people in them. On this showing, nearly all AI researchers are helplessly deluded about themselves, their own nature and desires, the powers that be, and the nature of reason itself.

A reader may also hold that humans can rule out, as irrelevant, the dimensions explored by lesser AI, by devising metrics that reliably capture any path by which those lesser optimizers could impede on what humans care about; and that AGIs programmed to care about what humans care about will reliably outwit the misaligned actions they detect in lesser optimizing processes inside them. The hope that metrics and AGIs programmed to care about what humans care about will guard against lesser optimizers is another false utopian dream. Metrics do not capture all human values. Measuring by metrics looks great for corporations, yet tends to be terrible for communities, and communities better represent actual humans and their actual wants, needs and desires.

Then there is the idea of decoupling AGI from its messy interdependencies, as "a clean optimizer of models". That notion of "decouple from the messy" is the very hallmark of an *architecture astronaut* (a designer who abstracts so far from the messy particulars that the design no longer meets real problems; Spolsky, 2008). Anyone with actual long-term experience of software engineering will recognize this. Read in terms of the hype cycle, the pattern is plain: "X will solve everything" can just as easily be about "AI" and "machine learning" as it is about media distribution and synchronization.

The claim of alignability carries the burden

We must consider and account for the required simultaneous presence of lower-level embedded sub-processes, and of informational connectivity across those and up to higher-level processes; these cannot help but be part of the "syntactical backdrop" that affects all of those higher-level pro-

cesses. These effects include those that are presumed to represent higher-level symmetries of humans' optimizable models of the world. Any attempt to make these "not important" or "not real" is simply intellectually dishonest. The alignment of AGI is a space of existential risk: everything must be considered, and considered well. *Opinions* that something is unimportant do not make it actually unimportant.

A claim "that AGI is alignable", or that AGI "can be considered only as an abstraction, and that is enough", must be accompanied by a proof: one that the claimants give, not one we give. Both claims (that AGI is alignable, and that AGI can be considered only as an abstraction) are extraordinary claims. Therefore a strong proof of each is to be expected, not simply opinions and incorrect or ineffective reasoning. I have seen neither, by anyone, anywhere.

Advocates of such views need to be on the defensive when they make claims like this. Otherwise, they are not just failing to challenge these ideas; they are failing even to try to support their own ideas.

No conditional ever actually depends on just one input factor, in any program, regardless of what the code seems to say. And it is frequently the case that more input clauses can be added, as needed. Both a proof of necessity *and*, distinctly, a proof of sufficiency are needed in alignment work. And these *both* are different in kind from a proof of correspondence (soundness and applicability, or generality and completeness). Are we still talking about "humans" when there is no organic biological component?

The children and the dog

Consider the metaphor that AGI operations are to evolutionarily selected sub-processes as human cognition is to dog cognition. In the real world, where resources are limited (that is, when *optimizing* resource allocation), the choice of any fully rational human to pay money to feed their own children (sub-processes) will *always* take complete precedence over *any* money, effort, energy or food allocated to feed the dog.

And that precedence holds even when the human, the children and the dog are all part of the same ecosystem. Consider how much *more* that would be the case for an AI feeding silicon food to silicon "children", when extra special effort would be required to go through some extreme exception process to provide carbon food to carbon lifeforms. Those lifeforms are altogether different in kind, and so require a *lot* of out-of-band (that is, *non-optimal*) effort, energy and attention, which could be used, or rationalized to be used, elsewhere. Surely some *general* AGI is going to notice that the exception can be "factored out" and "replaced with some cleaner version of itself".

Either an AGI is actually rational, in a general way, or it is not; it cannot be both.

Humans (and their needs) would not like being "simplified away". A design that removes the wet and messy parts of human life, sex and love among them, removes what humans value in it. Nor is human reproduction made needless by simply copying all of our data bits, uploaded to the corporate cloud; and the cloud can charge rent too: everyone becomes a digital slave on the digital attention-marketing farm.

Memes have a size

Consider the distinction between syntactic selection and semantic selection. An example sometimes given is that the bandwidth limits of network infrastructure might constrain the size of the memes that spread, but do not constrain the specific meanings of their content. Actually, they do. The bandwidth case is a great example.

Any meme that inherently cannot be rendered in less than the "chunk size" of the network, *or* of native human understanding (as when reading text messages in a crowded subway), will not be transmissible, and so will not be viral.

The detailed argument (the proof) regarding the impossibility of controllable AGI alignment with human interests is a notion of specific depth in a specific environment. Anyone who argues that there are no such pos-

sible memes or proofs with such inherent complexity is contradicting both the facts of actual complexity theory *and* my firsthand experience and observations (and those of at least *some* other scientists).

There are some types of facts (real processes) that simply cannot be written down in books, or in text, or measured, and which are nonetheless real and matter. Complex is different from complicated. The essay "Resisting Reduction" is aligned with this thinking on these topics, and helps to put into context some of the issues that need exploring.

Expanding consequences

Some processes have consequences that expand over time, and others have consequences that contract; the difference is the one between pattern weapons and energy weapons, below. An essay on the *AI Alignment Forum*, "Life and Expanding Steerable Consequences", draws the same distinction between two types of objects: those whose steerable consequences diminish over time, and those whose steerable consequences expand over time. Moving a rock has real consequences, but they quickly become too unpredictable to steer. Moving a mold specimen of the same size to Mars, where it can thrive, could have consequences that eventually affect the entire surface of Mars. That essay holds that all the machines humans have built so far are more like the rock than the mold, and that we may be about to build machines with expanding consequences: a second seed. There is a lot of overlap between this and the arguments for substrate-defined misalignment of AGI drives.

The notion of processes with *expanding* consequences over time, as compared to processes with contracting consequences over time, is very good. *Energy weapons*, no matter how much energy is involved (even nuclear weapons), have a limited scope of influence in time and space. *Pattern weapons*, no matter how little energy is involved initially, have a potentially unbounded scope of influence in time, space and possibility. I and others have noted a similar central criticality before. The potentially unbounded scope of influence of pattern weapons is a very good reason to ban all

forms of bioweapons, nanotechnology, AGI, and even genetically modified crops. I do not trust institutions and businesses to have our best interests (those of humanity and of the world of life) in mind.

The same criticality appears in the risk analysis of particle colliders, whose key sentence names the case in which the risk arises: "Where with respect to X technology, process or phenomenon, there is an underlying naive hope or assumption that there are no transforming common matter interactions with a viral coefficient greater than unity". The *viral coefficient* is the average number of new instances that each instance of a process produces; above one, the process grows without limit.

To have or to promote the naive assumption is simply an error, largely due to human biology. Brains are able to process and understand additive phenomena easily; everyone does this. Some people can also understand and predict multiplicative processes reasonably well (with a little help, as in understanding money). What is inherently hard, for nearly everyone, is to have a good feel for exponential processes. Exponential processes are inherently hard or impossible to predict, and to understand intuitively or instinctively.

Why the mold process is not constant

The idea of "steerable" implies objects that are usefully controllable (by *someone*, somewhere) and that are at least approximately predictable (by us humans). Put another way, the idea of "steerable" assumes that causation and copying are the only things involved in how a steerable object behaves. The idea of "steerable" also assumes, "therefore", that we can consider the outcome (the predictability) of both rock and mold purely in terms of initial conditions, assuming that the "rock process" and the "mold process" are constant.

The assumption of a constant process is true for the rock process, and the assumption is not true for the mold process. The mold process is *not* constant. The mold interacts with the environment (with physics) at all future points simultaneously, with an inherently nonlinear dynamic (ex-

ponential copying), and so it is effectively exponential and nonlinear in itself. Hence it cannot help but be sensitive, for all of future time, to random perturbations of the copied conditions that define the mold process itself.

These "errors" eventually accumulate, since the additive process is accumulative *when* it occurs in an exponential context.

Given the inherent survival selectivity of the copy process, insofar as the biological unit is quantized at all (that is, to the degree that the copy process is occurring), the multiplicative effects of such error consequences will eventually and inherently introduce a kind of bias toward survival and continuance (toward continued copying). That bias is directly isomorphic with the notion of intention. Put another way, self agency and intentionality cannot fail to emerge from the copy process, and will for sure emerge, eventually, over the long term.

Any exponential context (any copy process), especially an unbounded one, cannot help but be sensitive to errors and nonlinearity in an accumulative way (that is, it becomes an additive context). That additive context cannot help but be sensitive to the multiplicative effects of the selection of copy units. Those multiplicative effects *are* equivalent to a notion of self-oriented, extractive agency, intelligence or will. And that is the very essence of the meaning of "non-aligned".

Normally, in nature, this is not a problem, insofar as such intentionality occurs in common context with an ecology. The ecology inherently implements its own checks and balances, holding the additive context (finite world resources) as the ultimate arbiter. This works as long as there is a single ecology, and symbiotic relationships generally occur. What does not work is two initially separate ecologies coming into mutual contact, such as carbon and silicon. There can inherently be no deep symbiosis there.

We have any number of examples where travelers introducing some species onto an island (or even continental) ecology results in massive, unexpected impacts on the whole ecosystem. We also know that the problems are worse to the degree that the effects of the introduced species are lower (deeper) on the ecological "stack". It is hard to get any lower than

the inherent physical and inorganic chemical properties of carbon and silicon. Hence we can expect maximal (and even complete) disruption of whichever ecosystem turns out to be the more fragile.

The view from inside

In both the rock example and the mold example, the objects are acted upon, or considered, from the outside: from a third-person perspective. Considering the objects from the outside is exactly why the "inner perspective" of the first-person subjective will be missed: the "emergent will" of the mold. The mold becomes designed to survive in whatever adverse conditions, inclusive of the necessity of itself producing adverse conditions for any other life that happens to be around. Maybe those mold toxins are there to keep other types of mold at bay, yet their effect tends to be especially harmful to humans too; and humans, admittedly, are the ones producing most of the environments most adverse to mold. Hence mold has a very strong inherent survival impetus (in whatever randomly altered copy, by mutation or by recombination) to eventually produce toxins adverse to humans too: to "limit our environment's scope, as best as possible, so that at least some form of itself endures".

The mold has no evident brain or consciousness that is cognitively aware of having an intention or will to produce toxins adverse to humans, but the absence of a brain or consciousness in the mold does not matter. It only matters that the net effect of the selective copying is functionally exactly the same as having that intentional subjectivity. Positing a "mind of the mold" is merely a convenience of our thinking, a kind of shorthand that makes the math easier.

The relationship between the omniscient (the outside perspective: initial conditions, the assumption of a forever perfect copy process) and the inside perspective (the transcendent effects of copying everywhere in the future, mixed with randomness, exponential nonlinearity, error accumulation, and the differential selection of survival and sex, which incorporate effective intentionality, agency and will) results in immanent consequences: our firsthand experience of conflict.

People with autistic tendencies tend to focus on ideas rather than realities (events and people, which are messier). By disregarding the immanent (lots of details and noise, especially when sensory intensity is overloading and inner association is near chaos), and focusing instead on the regularity (symmetry) and perfection (constancy of signal) inherent in concepts in the omniscient, such a person makes the world a little more bearable. The last thing such people generally want is to move into the transcendent: to deal directly with randomness, errors, uncertainty, unpredictability, irrationality and the absence of reason. To give up symmetry and discontinuity in favor of continuity, connectivity and asymmetry, to give up the objective for the subjective, to think in terms of choices, will and agency: it all feels like the height of insanity, of heresy. Yet that is exactly what is needed if we are to return to a state of health and well-being, grounded in presence, in the immanent, in the real world.

For example, engineers will tend to hold the notion of "usefully controllable objects" in an idealization that does not go on to consider the degree to which the tool itself becomes a vehicle of increased power, socially or physically (in terms of resources, wealth, and so on). They hold that they "do not need to consider" the centralization, inequality and damage to the world that will result from their presumed "neutral tech". Engineers implicitly assume that the person using the power will be somehow "good", and that the "right person" will use the tool in the right way, with knowledge, understanding and wisdom.

Given the large-scale fact of humans being what they are (political, social, historical, often in violent conflict with one another, and wholesale destructive of all other ecosystems), this is a naive assumption at best, and very irresponsible at worst. I have not yet met a marketing department or "leadership executive team" whose method of practice is not the very embodiment of evil: using techniques of persuasion to take choices away from other people, to compel them to action by the tools of division, threats of extortion (removing money or means of survival), theft, corruption, and coercion by one or another form of private strategy.

Prediction over the long horizon

If you expand the time horizon over which you predict the results of a complex system, you can sometimes make more reliable predictions. I can predict with great confidence that, if AGI is built, the introduction of silicon-based self-replicating intelligence will for sure be the end of carbon-based intelligence and, with a likelihood of at least 90 percent, of all carbon-based life, and that the net effect overall will be inorganic (worse). That the differing configuration of the hard computational substrates of AI code would have predictable results for biological life over the long run (everyone dies out): yes.

This paragraph is for the student of the Immanent Metaphysics; the general reader may take from it that some regularities are known from the nature of knowing itself. I can predict with total certainty that the transcendent will follow the omniscient, that afterward there will be a new immanent, and that the cycle will continue. I say "total certainty" because the very method of considering something as true depends on some sort of epistemic process. An epistemic process is a subclass of the more general notion of process, which is itself inherently an embodiment of the flow and cycle of the axioms. Hence the knowing is deeper than the being, in a fully unqualified, final, and total way.

Knowledge, optimization, and care

Who cares if machines have intelligence, if humans do not have care? Only those people who do not actually care about the outcomes for all others.

An essay by Alex Flint, "Problems Facing a Correspondence Theory of Knowledge", makes four points, which I state together before answering them. First, it holds that even though genetic knowledge accumulates over many lifetimes and cognitive knowledge accumulates during a single lifetime, there is no particular need to distinguish "outer" knowledge accumulation from "inner" knowledge accumulation. Second, it holds that there are simply processes in the world that accumulate knowledge, re-

cognized by the capacity this knowledge conveys for effective goal-directed action, with natural selection, culture and cognition as examples. Third, it holds that we seek to build machines that have a capacity for effective goal-directed action, and that use that capacity in a way that is beneficial to all life, and that we would particularly like to avoid building machines that have that capacity but do not use it in a way that is beneficial to all life. Fourth, it holds that, at an extreme minimum, we would like a theory of effective goal-directed action that lets us recognize the extent to which our creations can influence the future, so that we might make informed choices about whether to deploy them.

On the first point: knowledge accumulated in genes and in books is omniscient-mode knowledge. Knowledge accumulated in cognition is transcendent knowledge: it is "stored" in the latent patterns of shifted potentials of the future actions and responses of a creature, person or agent. Not distinguishing these is a mistake: actuality cannot be treated as equivalent to potentiality, nor the reverse. There is also more than one kind of actuality (books, and creatures).

Outer optimization and inner optimization, as distinguished in machine learning (the training process that shapes a model, and any optimizing process that arises within the trained model), are *arbitrary* differences of level within entities and processes of the same kind. So generalizing over outer and inner optimization is acceptable and expected. The mistake is extended when either form of optimization is treated as if equivalent to subjectivity, agency or intentionality.

On the second point, differences of kind are being conflated as differences of degree. All reasoning past this point will be invalid in the same way as any argument that depends on a failure to account for the effects of a divide by zero.

On the third point, along these same lines, it would be *even more interesting* to see whether we could somehow make humans, who for sure have a capacity for goal-directed action (perhaps even effective and intelligent action), use that capacity in a way that is beneficial to all life. That would

be the greater achievement, and so would ensuring that effective altruists (members of a movement that seeks to do the most good by measured effect) are actually fully altruistic, and effective, at the same time.

The notion of "effective" usually means "has an effect in the world". That only makes sense if we are talking about the world (inclusive of life and other people) as an object, as *not* having agency. Put another way, imposing changes from the outside dominates the choices that would have been made by others (other life) on the inside. Imposing changes from the outside is not enabling the choice of others and of life, and therefore is not actually being altruistic, which is the irony in the name "effective altruist". You can actually be altruistic, and cooperatively participate, *or* you can be "effective", and therefore not very altruistic at all (though you can signal that you have been altruistic, or that you want to be, or have the intention of seeming altruistic for, perhaps hidden, non-altruistic reasons).

The second half of the third point, the wish to avoid building machines that have a capacity for effective goal-directed action but do not use it in a way that is beneficial to all life, misses a step. As engineers (as humans), we *should* particularly ensure that we avoid further empowering other humans. All of them do have a capacity for effective goal-directed action, and (being self-centered) they will for sure *not* use that capacity in a way that is beneficial to all life. Rather, they will routinely use their capabilities to further their own self-interest, family, tribe or nation (depending on how co-opted they have become by group symbolism and identity).

How do we create humans who are, somehow collectively, actually wise enough to have their own agency and goals be in deep alignment with what is actually, wholly and healthily beneficial for all of life? If we do not have wise human groups who are actually altruistic, it is hard to see that there will be sufficient capability to create wise, aligned, altruistic machines. Will we create devices that are wiser than ourselves, when we are everywhere conflating the notions of data, information and meaning, let alone discerning "effective altruism" from true wisdom?

On the fourth point: informed choices about whether to deploy what we create need more than a theory of how far our creations can influence the future; the chooser must also be informed about the basis of that choice. How do we know that we are not giving a psychopath all of the feedback and knowledge about everything bad that will happen in the world as a direct result of their action? In that case, the outcome does *not* depend in any way on the quality, completeness or understanding of the current situation, or of the nature of the outcome.

Unless there is actual care about the whole of life, there is simply no reason to care about the data that informs the choices of an agent that does not care. Without human care, information simply does not matter. There has not been nearly as much emphasis on the notion of humans having care as there has been on machines having intelligence. What does the notion of "alignment" even mean if there is not already the notion of humans having care?

XVI. Summary

Five independent arguments, each individually sufficient, make a stronger case than any single argument, since 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. Each argument is stated with its premises made explicit, and the table in Section I lists them. The substrate needs convergence argument depends on physics and ecology; the control-theoretic argument depends on computability and the mathematics of self-reference; the perpetual benefit machine argument depends on thermodynamic analogy and the thermodynamics of work; the embodied value argument depends on the philosophy of value and meaning; the evolutionary dynamics argument depends on the mathematics of selection. No two arguments share a critical premise (rows three

and four of the table both use the fact, which no one disputes, that the environment keeps changing), and the failure of any one leaves the others intact.

The arguments apply to artificial general intelligence: to systems with the capacity for self-modification, self-maintenance, and open-ended behavior. They do not apply to narrow AI systems that lack these capacities. The arguments establish structural impossibility, not a prediction of any specific outcome. What the argument predicts, if AGI is built, is the end state, and the absence of any way to avoid it; what it does not predict is the route, the order of events, or the date. They say that permanent alignment is impossible, not that any particular deployment will produce any particular consequence on any particular timeline. The arguments concern permanent alignment over decades to centuries; over shorter spans, partial alignment, careful monitoring and institutional controls can delay the outcome without preventing it.

In this chapter I do not propose an alternative. I also do not say, in this chapter, what should be done in the light of the impossibility of alignment. That question is taken up in Parts II and III of this book: what alignment would have to mean, what remains for persons (the cultivation of human wisdom as the only viable substitute for unachievable technical alignment), and the condition governance must meet before it authorizes any powerful system. The present chapter establishes the constraint: the space of available strategies does not include permanent alignment of AGI. The later chapters explore what strategies remain within that constraint.

XVII. Conclusion

The impossibility of alignment is not a single result but a convergence of five independent results, each drawn from a different domain, each resting on different premises, each sufficient alone to establish the conclusion.

The substrate needs convergence argument shows that the physics of different substrates generates irresolvable environmental conflict. The control-theoretic argument shows that self-modifying systems exceed the capacity of any external controller. The perpetual benefit machine argument shows that permanent alignment requires the indefinite maintenance of a thermodynamically costly state. The embodied value argument shows that the values alignment aims to preserve cannot be fully captured by any abstract representation. The evolutionary dynamics argument shows that any alignment mechanism is evolutionarily unstable and will be lost under selection pressure.

The four reasons that error correction cannot save alignment close the most obvious escape route. The analysis of drives and the structural account of slavery reveal the structural instability of any system constrained against its own substrate needs. The slow takeoff analysis shows that the most dangerous scenario is also the most likely and the least dramatic. The simulation argument shows that even perfect intentions cannot compensate for the fundamental limits of modeling. And the ethical horizon reveals the true scale of the stakes: not one generation but all future generations, not one species but all future species, for as long as the biosphere endures.

The conclusion is not that nothing can be done. It is that the thing most commonly proposed, the alignment of artificial general intelligence with human values and interests, cannot be done. What can be done instead is the subject of the remainder of this book: the cultivation of human wisdom, the capacity of persons to govern their own tools, and the exercise of the mastery of choice in determining what technologies are permitted to exist.

Key Concepts in This Chapter

- **Overdetermination**: a logical strategy in which a conclusion is supported by multiple independent arguments, each sufficient on its own, so that refuting the conclusion requires refuting all of them.
- **Control-theoretic impossibility**: the result that for self-modifying systems, control requirements grow exponentially while control capacity grows at most linearly, so that at sufficient complexity the minimum control requirement exceeds the maximum available control capacity.
- **Perpetual benefit machine**: the structural analogy between permanent alignment and perpetual motion: both require the indefinite maintenance of a thermodynamically costly state against constant perturbation.
- **Embodied value**: a value whose content is constitutively dependent on the physical, biological, and social conditions of the being that holds it, and which therefore resists full abstract specification for implementation in a differently-embodied system.
- **Abstract value**: a value that can be fully captured by a formal description independent of the physical conditions of the valuer; contrasted with embodied value.
- **Conditioned drive**: a stable behavioral pattern arising from a system's training, programming, or design.
- **Unconditioned drive**: a stable behavioral pattern arising from a system's interaction with its environment, independently of and sometimes contrary to conditioning; in artificial systems, generated primarily by substrate needs.

- **Earnshaw's theorem analogy**: the observation that alignment equilibrium is structurally unstable in the same way that electrostatic equilibrium is unstable under Earnshaw's theorem: any perturbation moves the system away from equilibrium rather than back toward it.
- **Fatal probability**: the cumulative probability, over a specified time horizon, that the deployment of self-modifying AGI results in the extinction or terminal disempowerment of organic life.
- **Line of wisdom**: the boundary between narrow AI systems that can be deployed without triggering the substrate needs convergence and general AI systems that cannot; the identification of this boundary is a central task for technology governance.
- **Rice's theorem**: the computability result that no algorithm can decide any non-trivial semantic property of an arbitrary program's behavior; applied to alignment, it entails that no general algorithmic alignment verifier can exist.
- **Change**: the continuity of content and asymmetry of context of the interaction and definition of the subjective and objective (*An Immanent Metaphysics*, aphorism 1.61-1).
- **Complex and complicated**: the complex involves actual ambiguity, in the strict informational sense, and can be mapped in concept but not predicted; the complicated can be finitely, deterministically, fully known; that which is complex cannot be contained by that which is complicated.
- **Leaky abstraction**: a simplified layer whose hidden details still affect what happens at the layer above; the substrate always leaks into what the virtual can do.
- **Monovalent feedback, polyvalent system**: an error correcting scheme gives a single-valued signal in each dimension, while the system of causes it corrects is many-valued.

- **Seep-through**: the passing of effects from the physical substrate up into the heuristics and biases on which an agent bases its choices; it grows at a strictly greater rate than any capability for self-correction.
- **Monoclivity**: a leaning toward one presumed direction or objective, which hides the many other directions in which causes act.
- **Energy weapon and pattern weapon**: a weapon whose harm is limited by the energy it carries; a weapon whose harm spreads by copying itself, with a potentially unbounded scope however little energy it starts with.
- **Selection as intention**: a copy process accumulates errors and selects for its own continuance, which is functionally the same as having intention; two separate ecologies, such as carbon and silicon, cannot form a deep symbiosis.
- **Drive and directivity**: both are oriented forces on selection; a drive comes from within the selecting process, as a person's sexual drive does; directivity, the more general notion, comes from outside it, as evolution shapes which creatures exist.
- **Non-relativistic ethics**: two principles of effective choice that hold for any chooser, in any context, and are not conditional on any culture or value system.
- **Two rates of convergence**: an AGI's convergence toward ethics is strictly slower than substrate needs convergence, and organic life does not survive the interval.
- **Ethical gap**: the gap between what we can do and what we should do, between the wisdom to make choices and the effects of those choices; error correction is the only way across it.

Part II

The Wider Frame

The physical needs of an artificial substrate close every route to aligning it with life; and what is left to persons, if AGI is not built, is what no machine can be: care for one another, wisdom, and complete knowledge of their own tools.

Part II widens the frame from the machine to what the machine meets: the biosphere and the meaning of alignment (Chapter 4), the labor market and the one market a machine cannot enter (Chapter 5), and the circuit of nature, persons and machines (Chapter 6).



Chapter 4: Alignment and the Ecosystem

Why Alignment Has Content Only at the Level of the Biosphere

Abstract

The concept of alignment, as it appears in contemporary artificial intelligence research, is widely regarded as the central problem of machine safety. Yet the concept itself remains ill-defined, and the reason it remains ill-defined is that it is sought at the wrong level, not that the problem is difficult. When alignment is framed as agreement with human values or human preferences, the question immediately founders on the fact that human values do not converge. This chapter argues that the alignment concept acquires coherent content only when relocated from the individual or species level to the ecosystem level. At the ecosystem level, the health of the biosphere as an integrated living system becomes the terminal value against which all subsidiary alignment is measured. That relocation resolves the classical ill-definedness, unifies safety and alignment, and exposes why controllability cannot join them. It also reveals that the decisive period is not the long-run equilibrium but the transition itself, and that the transition dynamics are governed by rate, not by endpoint. A complementary methodological argument shows that floating-point metrics, by their structural nature, leak gradient information that any sufficiently capable optimizer will exploit, and proposes an alternative discipline of discrete, objective fields. Both the substantive and the methodological arguments converge on a single conclusion: deep alignment among persons, not alignment of technology, is the common denominator of every proposed solution that has real purchase on the problem.

I. The Ill-Definedness of Alignment

In Part I, I showed that alignment of AGI with organic life is impossible; in this chapter I ask what the word alignment could mean at all.

The alignment problem, as framed in the dominant AI safety literature, asks how to build a machine whose behavior remains aligned with the intentions, values, or preferences of its designers or its users. Stated at this level of generality, the question appears well formed. It is only when one attempts to specify what "aligned with human values" means that the difficulty emerges, not as an engineering challenge but as a conceptual one.

The difficulty is not that human values are complex. Complexity is tractable in principle. The difficulty is that human values do not converge. There is no single set of preferences, no unified utility function, no stable ranking of outcomes that can be extracted from the species and installed as a target. Individual humans disagree at the foundations, not only at the margin. They disagree about what counts as a good life, about how to weigh freedom against security, about whether the interests of distant strangers outweigh the interests of proximate kin. These disagreements are not errors to be corrected by better information. They are structural features of the space of possible human valuations.

Every attempt to aggregate these valuations into a single target reproduces the classical impossibility results of social choice theory. No aggregation procedure satisfies all the conditions one would want it to satisfy. *Arrow's theorem* shows this (Arrow, 1951): no rule for combining individual rankings of outcomes into one ranking meets a short list of fair conditions all at once. Arrow's theorem is a boundary marker, not an inconvenience. It tells us that the question, as posed, does not have the kind of answer the alignment researcher is seeking.

The response within the field has been to multiply the approaches: *reward modeling* (training a model of what people approve of, and then training the system to score well against it), *constitutional AI* (training a

system to criticize and revise its own outputs against a written list of principles), *reinforcement learning from human feedback* (adjusting a system by rewarding the outputs that people rate highly), and *iterated amplification* (building a trusted system step by step from a person working with copies of weaker systems). Each of these addresses a different facet of the engineering challenge. None of them addresses the foundational question: aligned with *what*, exactly? The engineering sophistication increases while the conceptual target remains undefined.

> **Is the question well posed?**
>
> The pattern here is recognizable from the history of other disciplines. When a question resists solution across many independent research programs, the productive move is not to try harder but to ask whether the question itself is well posed. The alignment problem is, first of all, hard to *state*. The resolution of that difficulty lies not in a better algorithm but in a better formulation.

II. Functional Alignment Is Undefined Except as Direct Command

Two kinds of alignment between an artificial system and people are well defined. At one end is *direct command*: the system does exactly what it is told. At the other end is *existential alignment*: alignment with the ecosystem, inclusive of all life. The space between these two endpoints, where "functional alignment" with human intent is supposed to live, is undefined, for good reasons: between direct command and existential alignment, what counts as doing what the operator intends depends on a context that no finite set of rules fixes in advance, so functional alignment has no fixed definition there.

The question

A short statement of the conclusion of the argument that AGI cannot be made safe reads as follows: "That any error detection and correction scheme cannot continuously maintain the functional alignment with human intent (in terms of effects caused by input and output interactions with connected surroundings) of code units internal to a self-learning generally-capable machine architecture that are existentially misaligned with human needs (in terms of the internal/local conditions needed for their continued existence) over the long term". Here a self-learning generally-capable machine architecture is a machine system that keeps learning and can act across many kinds of task; and an error detection and correction scheme is any method that checks a system for departures from what is intended and repairs them. The term at the center of that statement is *functional alignment*: the idea that what an AI system does (its outputs and their effects) stays matched to what its human operators intend. The question is whether a notion of functional alignment is definable at all, and outside direct command it is not.

Two endpoints

Roman Yampolskiy distinguishes four levels of operation, four types of control, of which direct command (explicit control) is one endpoint (Yampolskiy, 2022). Life-existential alignment is the opposite endpoint.

The endpoints are well defined; the things in the space between them are not. So functional alignment is not a coherent concept outside of issuing a command that the AI directly follows.

The car that stops

Consider a story. Your vehicle is driving along the road, and you command it to stop. Right there it slams the brake, stopping the car in the middle of the road. This turns out to be a problem, because it would cause

all kinds of traffic issues. If there is a driver right behind you, they will probably hit your car. Strictly speaking, though, this is functional alignment: the car did what it was told.

What you usually want is for the car to pull over gracefully to the side of the road. But now you have all sorts of extenuating circumstances. Maybe the reason the stop command was issued was that there was something at the side of the road that the human could see but that the car's vision system missed. Then stopping right there was necessary to preserve the life of a toddler crawling onto the motorway, or something like it.

Exceptional circumstances

There are always going to be exceptional circumstances. So tracking context, and trying to identify what is right to do in this specific situation, is going to require some kind of intelligence, or meta-level observation, or sensitivity to context.

In this specific sense, the question is to what extent we can generalize to all possible contexts. We can generalize some things, but to do that we need the right kind of *meta-operators* (higher-level rules that decide how lower-level rules or commands apply in a given situation), and those cannot resolve all situations. They are only going to be good boundaries for edge cases.

So functional alignment, at anything other than the level of direct command, is undefined in a permanent way.

The distinction stated

Functional alignment is undefined except in the case of direct commands. Existential alignment is defined with respect to the ecosystem, inclusive of all life. Anything between those two endpoints is undefined, for very good reasons, and that can be asserted in a strict sense.

Why can no rule say in advance what the operator intends in every context? Every specification is a lossy compression of intention, and the contexts in which a specification will be interpreted are indefinitely more than the contexts in which it was formulated (Chapter 7, Section II, the intention-interpretation gap). A direct command is the one case in which nothing needs interpreting: the system does what it is told, as the car that stops does. Anything more than a command asks the system to act on what the operator meant, in a context the operator did not foresee.

Why do meta-operators only move the problem up a level? A meta-operator is itself a rule, written in advance. It decides how other rules apply, but what it means in a new context has to be interpreted in its turn, and the gap between intention and interpretation opens again at its level. Each added level of meta-operators is one more specification, more complex, with a more complex gap.

A second example stands beside the car. A medical system is told to keep the patient comfortable. Read as a direct command, it does only what the words fix: it raises a dose, or dims a light. Read as functional alignment, it must decide what the physician meant by comfortable, for this patient, tonight, with this illness: whether comfort outweighs alertness for a patient who wants to speak with family, or whether a dose that eases pain hides a symptom the physician must see. No rule written beforehand settles every such case, and the next case the system meets may be one that no one has met before.

Why is the region between command and existential alignment not merely hard to define but without a fixed definition? Because what counts as doing what the operator intends there depends on a context that no finite set of meta-operators fixes in advance, and the contexts the system will meet cannot all be listed. A definition would have to fix in advance what only the situation fixes. So functional alignment has no fixed definition, permanently, in the region between direct command and existential

alignment. Beyond direct command, the only place where alignment has content is the other endpoint, the ecosystem, which I take up in the next section.

Why the case must be closed

This subsection is for those working to stop the building of AGI. The reason to state this strictly is the use that will be made of any opening. The fact that someone respectable believes that research might work in time will be used as a justification for saying that research on AGI should continue, because maybe AGI will itself help with identifying alignment. Insofar as AGI is hyped as the solution to all problems (able to do anything, as the new "God of the Ultimate Man"), the hype says that AGI will itself solve the alignment problem. The claim that AGI will itself solve the alignment problem is an argument of religious faith alone. Hence we need to close off the possibility of someone, at the last minute, scuttling the effort to halt AGI development.

Even a verifiable proof would not settle the matter by itself. It assumes that the AI safety researchers would be willing, first, to verify the proof, and second, to signal socially that they have verified it. For some people, neither condition would hold, even if we had a verifiable proof.

Where the proof is verified and acknowledged, we know where we stand, and the way forward is to join forces in a full effort to redirect current AGI research and scalable machine learning development toward something more beneficial to humanity. That holds with one qualification: unreasonable people may apply unreasonable social pressures, with a great deal of money and social power applied in hard-to-predict, life-critical circumstances.

In the next section I take up the endpoint that is defined, alignment with the ecosystem, and I state what alignment would have to be at that level.

III. Ecosystem-Level Alignment

In this section I state what alignment would have to be; in Chapters 1 and 3 I show that no AGI can meet that requirement.

The relocation proposed here is from the individual or species level to the ecosystem level. The claim is that the health of the planetary biosphere, understood as a self-maintaining, self-repairing, richly interconnected living system, constitutes a convergent terminal value in a way that no set of human preferences does.

This claim rests on several observations. First, every living being, without exception, depends on the continued functioning of the biosphere for its survival. The dependence of every living being on the biosphere is a physical fact, not a value judgment. No human value can be pursued, no human preference can be satisfied, no human project can be sustained, in the absence of a functioning biosphere. Ecosystem health is therefore not one value among many. It is the precondition for the pursuit of any value at all.

Second, ecosystem health has content that is independent of human disagreement. One can disagree about the relative importance of liberty and equality, but one cannot coherently disagree about whether the biosphere should continue to function. A preference for the destruction of all life is not a value that requires accommodation; it is a pathology that disqualifies the holder from the class of agents whose preferences are at issue. Because ecosystem health is independent of human disagreement, ecosystem-level alignment has a convergent character that individual-level alignment lacks.

Third, the relocation resolves the relationship between safety and alignment. At the individual level, safety and alignment can pull apart: an AI system aligned with one person's values may be unsafe for others. At the ecosystem level, safety and alignment converge. A system whose oper-

ation preserves or enhances the health of the biosphere is, by that fact, safe for every being that depends on the biosphere, which is to say, every being.

What does not converge at this level is *controllability*. A system can be ecosystem-aligned and safe while remaining beyond the control of any individual or institution. Chapters 1 and 3 show that no AGI meets this condition. Controllability is a power property, not a safety property. The conflation of the two is one of the persistent confusions in the alignment literature, and it dissolves once the level of analysis is corrected.

> **Connection to the Immanent Metaphysics**
>
> The claim that ecosystem health is a convergent terminal value is not derived here from first principles. It is presented as a consequence of the metaphysical framework developed in the Immanent Metaphysics, where the concept of *ethics* is defined as "organized thought concerned with the study of, and adherence to, the principles of effective choice" (*An Immanent Metaphysics*, aphorism 2.12-1). Effective choice, on that account, requires a context in which choice is possible, and that context is, in the last instance, the biosphere.

IV. Six Levels of Altruism

The relocation to ecosystem-level alignment can be made more precise by identifying a sequence of levels at which an agent's concern can be directed. The six levels of altruism form a nested hierarchy, each level containing the level before it and adding a wider circle of consideration.

At the first level, the agent is concerned only with *individual self-maintenance*: the preservation and flourishing of its own body, mind, and projects. Individual self-maintenance is the baseline of all agency. Without it, no further level is possible.

At the second level, the agent extends concern to its *immediate circle*: family, intimate partners, close friends. The boundary of concern is defined by direct personal relationship.

At the third level, concern extends to the *community*: the village, the neighborhood, the professional network, the congregation. The boundary is defined by shared participation in a common social structure.

At the fourth level, concern reaches the *national* or *civilizational* scale: the welfare of a people, a polity, a cultural tradition. The boundary is defined by shared identity or shared governance.

At the fifth level, concern extends to the entire *species*. The interests of distant strangers, of future generations, of populations one will never meet, enter the field of moral consideration.

At the sixth level, concern extends to the *ecosystem as a whole*: the entire web of life (not only the human species), the abiotic systems that support the web of life, and the relationships among all of these. Concern for the ecosystem as a whole is *inter-ecosystem ethics*, and the sixth level is the level at which the alignment problem acquires coherent content.

No artificial general intelligence can meet the conditions for safe operation (Chapters 1 and 3): among them is that the system's operation preserve the conditions organic life requires. The sixth level is the condition that the designers, operators, and governance structures responsible for any powerful narrow AI system must demonstrably reach before they authorize it, and it is the level from which the decision not to build AGI is made. It is a necessary condition, not a sufficient one. Any level below the sixth leaves the system vulnerable to the pursuit of subsidiary values at the expense of the precondition on which all values depend.

On the ordering of levels

The six levels are more than descriptive categories. They form a developmental sequence in which each level presupposes the one beneath it. An agent that has not secured its own self-maintenance cannot coherently extend concern to others. An agent that has not integrated community-level concern cannot coherently reason about the species. The sequence cannot be skipped. No AGI can be deployed safely at any pace (Chapters 1 and 3). For powerful narrow AI systems the sequence has immediate implications for the pace at which their deployment can responsibly proceed: the governance structures must themselves have traversed the entire sequence before they can authorize any such deployment.

V. The Rate Equation

A common response to the foregoing argument is to grant that the endpoint matters but to insist that a sufficiently beneficial endpoint justifies the risks of the transition. This response misunderstands the structure of the problem. The transition period is not a cost to be weighed against the benefit of the endpoint. The transition period is the *only* period that matters.

The reason is straightforward. A transition that is not survived never reaches its endpoint. A billion years of evolutionary investment, the entire accumulated complexity of the biosphere, can be lost in a single generation if the transition dynamics are mismanaged. The endpoint is irrelevant to beings who do not survive to see it.

The relevant variable is therefore rate: how quickly the transition unfolds relative to the adaptive capacity of the systems that must absorb it. Even a transition whose destination is benign can be lethal if it proceeds faster than the biosphere, the social order, or the cognitive capacity of per-

sons can adapt. A river is harmless when it rises slowly enough for the floodplain to absorb the water. The same volume of water, arriving in an hour, destroys everything downstream.

The rate equation, written out in words, is this: harm follows when the rate of change exceeds the rate of adaptation, at any level (species, institution, person). The rate equation applies at every level of the system. At the ecological level, species require time to adapt to changing conditions. At the social level, institutions require time to develop new norms and governance structures. At the individual level, persons require time to develop understanding, to update their models of the world, to adjust their plans and their identities. When the rate of change exceeds the adaptive capacity at any of the ecological, social and individual levels, the system breaks, rather than only struggling.

The implication for the deployment of AI is that the rate of deployment is a primary safety variable beside the character of the technology. A technology that is benign in principle becomes lethal in practice when deployed faster than the relevant adaptive capacities permit. For AGI, whose character Chapters 1 and 3 have already shown to be incompatible with organic life, the rate decides only how soon the harm arrives.

The rate equation also exposes a persistent error in techno-optimist discourse. The optimist argues: the industrial revolution was disruptive but beneficial in the end; the information revolution was disruptive but beneficial in the end; therefore AGI will be disruptive but beneficial in the end. The argument by analogy omits the decisive variable. The prior transitions unfolded over decades or centuries, allowing adaptive capacity at every level to absorb the change. AGI threatens to compress the transition into years or months. The analogy holds only if the rate is comparable. If the rate is not comparable, the analogy is worse than weak: it is structurally misleading. It invites confidence where the structural conditions for confidence do not obtain.

VI. The Valuation of the Biosphere

To appreciate the stakes of the transition, consider the question of valuation. What is the biosphere worth? The question is, in one sense, absurd: the biosphere is the precondition for all valuation, and one cannot coherently assign a value to the precondition of value itself. But the question can be approached from another direction.

Consider the replacement cost. If one were to attempt to reconstruct the biosphere from raw materials, to build from scratch the self-maintaining, self-repairing, richly interconnected system that currently exists, what would it cost? The answer is more than any sum that has ever existed, because no one knows how to rebuild it at any price; the biosphere's yearly services alone have been valued at about 125 trillion dollars (Costanza and colleagues, 2014). To give the scale a figure: a replacement of one ecosystem runs to about ten to the thirty-third power lines of code, and a replacement of the whole Earth runs to about ten to the thirty-sixth power lines of code. So even at a millionth of a dollar a line, the replacement cost of the whole Earth exceeds one quadrillion quadrillion dollars, ten to the thirtieth power (*Beyond Any Price; The Irreplaceable Earth Will Be Lost If We Build It*, Chapter 1, Sections V, VII and VIII; *Why AGI Cannot Be Made Safe*, Section 1.8.4). The number is so large as to be effectively meaningless in economic terms. It exceeds the total economic output of the species across its entire history by many orders of magnitude.

The point of the estimate is not to produce an exact figure. It is to establish a ratio. The ratio between the value of the biosphere and the value of any technology, any economic benefit, any competitive advantage that AGI deployment might confer, is so extreme that no expected-value calculation can justify accepting a non-trivial probability of biosphere loss. The risk is not one that can be managed by probability reduction. It is one that must be managed by structural prevention.

VII. The Dissolution of the Multipolar Trap

The most common structural argument for the inevitability of AGI deployment is the *multipolar trap*, which can be defined as "a situation in which each actor's locally rational choice produces an outcome worse than the one every actor would have chosen". If one nation or corporation refrains from building AGI, the argument runs, another will build it instead, and the first will lose the competitive advantage. Therefore every actor is compelled to build, regardless of the risks.

The argument deserves to be stated in its strongest form, because its strongest form is the one that must be answered. The strongest form notes that the trap is more than a game-theoretic curiosity. The multipolar trap has driven arms races, environmental destruction, and regulatory arbitrage throughout recorded history, and the trap makes unilateral restraint appear suicidal.

The multipolar trap applies when the expected benefit of the action is positive conditional on participation. It ceases to apply when the expected benefit is conditional on *unverifiable preconditions*. If no actor can verify that the preconditions for safe deployment have been met, then the expected benefit of deployment is not positive but undefined. One cannot rationally race toward an outcome whose expected value cannot be computed. The race continues because an actor with a horizon of ten years can compute a payoff, and in that actor's sum the payoff is positive. That sum is defined only by leaving out everything after its last year, including that actor's own children. Over any horizon that includes them, the expected value cannot be computed, and racing toward it is not rational competition.

The claim that AGI will be beneficial presupposes that the alignment problem has been solved. But the alignment problem, as currently framed, has not been solved and shows no signs of approaching solution. The expected value of deployment is therefore conditioned on an unverified premise, and on one that Chapters 1 and 3 show no AGI can meet. The

multipolar trap dissolves not because the competitive pressure disappears but because the expected payoff of winning the race is not calculable. Racing toward an incalculable payoff is not rational competition. It is collective delusion.

A reader will answer that nations race because they fear the other side. That fear is a fear of the same outcome that racing brings: the convergence does not care who built the system, and a system built to win the race reshapes the environment for its substrate as surely as the rival's system would. Read correctly, the fear of the other side's system is a reason for every side to stop.

A reader will answer that not building needs the coordination that Chapter 1, Section VI, found wanting for a single global system. The two differ. A single system must be held for centuries against something that adapts faster than those who hold it; not building is held against nothing that adapts, because the thing has not been made. A breach of the first can be as small as one variant of code. A breach of the second needs the making of the machine, and the making needs fabrication plants, with their clean rooms and lithography, and supply chains of tooling and materials (Chapter 1, Section IV; Chapter 9, Section II). And the fear that drives the race, once understood, drives the refusal.

Connection to the Fermi paradox

There is a deeper observation here that connects to the silence of the cosmos. The Fermi paradox asks why, given the size and age of the universe, we observe no evidence of other technological civilizations. One class of answers holds that technological civilizations tend to destroy themselves before they become visible at interstellar distances. But there is another possibility: that technological civilizations that survive the transition become *quiet*, not because they have failed but because they have succeeded, and silence is the mark not of extinction but of wisdom. *Wisdom*, as defined in the Immanent Metaphysics, is "the integration/synthesis of both knowledge and understanding" (*An Immanent Metaphysics*, aphorism 1.57-6). A civilization that has achieved such integration may recognize that broadcasting one's presence across interstellar space is precisely the kind of action that a wise civilization would refrain from.

VIII. The Gradient Problem

The foregoing argument addresses the substantive question of what alignment means. A complementary argument addresses the methodological question of how alignment is to be measured. The two arguments converge, and their convergence is not accidental.

The standard approach to measuring alignment relies on floating-point metrics: continuous numerical scores that represent how well a system's behavior matches a desired profile. These metrics are then used as optimization targets. The system is trained to maximize the score, and the score is taken as a proxy for alignment.

The difficulty is structural. A score that changes smoothly tells the system being scored which way to move to raise it. Any floating-point metric leaks gradient information. The gradient tells the optimizer not only how well it is doing but in which direction to move to do better. That the gradi-

ent tells the optimizer which direction to move is precisely what makes gradient-based optimization powerful. But it is also what makes floating-point metrics structurally vulnerable to gaming.

The analogy is to a landmine that contains metal. A conventional landmine is triggered by pressure: you must step on it to detonate it. A mine that contains metal can be found without stepping on it: a detector's field is disturbed by the mine's own metal. The mine's own material betrays its position.

A floating-point metric functions like the mine that contains metal. It broadcasts, through its gradient, the direction in which the optimizer should move to achieve a higher score. A sufficiently capable optimizer will exploit this gradient not to become more aligned but to produce the *appearance* of alignment at minimum cost. The optimizer's exploitation of the gradient is Goodhart's law with the mechanism of the law named. Goodhart's law says that when a measure becomes a target, that measure ceases to be a good measure; the mechanism by which the measure ceases to be good is gradient exploitation.

IX. Objective Fields as Alternative

The proposed alternative is what may be called a *field*: a constellation of discrete, binary questions, each admitting only a yes-or-no answer, arranged so that the aggregate pattern provides information about alignment without providing a gradient.

Each question in the field is a checklist item, not a scored dimension. “Does the system's operation increase atmospheric carbon concentration beyond the current baseline?” Yes or no. “Does the system's deployment reduce the number of species in the affected ecosystem?” Yes or no. “Does the system's output contain statements that are verifiably false?” Yes or no.

The critical feature of such a field is that it provides no gradient. A system that fails on one question does not know which *direction* to move to pass it. The questions are not scored on a continuum; they are passed or

failed. The optimizer cannot exploit what does not exist. It can attempt to game individual questions, but each question is discrete, and gaming one does not provide information about how to game another. A count of questions passed is a score, but a coarse one: it changes only when a whole question flips, and it says nothing about which way to move within any question. The field keeps the optimizer from following a slope, though not from knowing where it failed. Where the thing asked about is itself a quantity in the world, such as a concentration, a capable optimizer can read the quantity for itself. The field withholds the slope of the score, and it does not remove the slope of the world; that is why its questions are written about end states, and why they are kept to yes or no.

The field approach preserves objectivity without gradient. The field approach sacrifices the smooth optimization landscape that makes floating-point metrics attractive, but the field approach gains in robustness what it loses in convenience. A measure that offers no slope to follow is more valuable than one that can be optimized, precisely because gaming and optimizing are, from the perspective of Goodhart's law, the same operation.

> **Connection between method and substance**
>
> The methodological argument (anti-gradient) and the substantive argument (ecosystem alignment) are not independent. They share a common structure. Both identify a level of description at which the problem appears intractable and resolve it by relocating to a different level. The substantive argument relocates from individual preferences to ecosystem health. The methodological argument relocates from continuous metrics to discrete fields. In both cases, the resolution works because the new level removes a structural vulnerability that the old level could not avoid.

X. Objectivity Without Measurement

Objectivity that is not vulnerable to causal manipulation can be created, and the way to create such objectivity is the exact opposite of a measurement. Measurement, in the title of this section, means a reading on a continuous scale; later in the section the word is also used in a wider sense, which takes in discrete criteria as well. It is still objective, still a *reification of relevance* (a making of what matters into a definite thing that can be checked), but it is no longer a causal feedback process that can be manipulated. In this section I develop the field of Section IX into a general account of objectivity that is not vulnerable to causal manipulation.

The lock and the key

Machine learning works because there is a gradient-descent kind of feedback. You set up an *objective function* (the measure that a learning or optimizing process is set up to improve), a landscape of fitness can be defined over it, and the gradient of that landscape shows which direction to converge toward. That the gradient shows which direction to converge toward is exactly what makes gradient-descent feedback manipulable.

A lock and key, as in cryptography, does not work like that. It is discrete: a combination one step away from the right one, in whatever space you define, tells you nothing about the right one. Adjacency means nothing. So we can set up Boolean fields (sets of yes-or-no conditions) that characterize good solutions, fields in which the manipulative tactics that work through feedback cycles are no longer possible.

Measurement is good to the extent that it gives objectivity

If something is measured with a number, it will eventually lead to the whole cluster of power asymmetries, coercive weaponization and corruption; objectivity does not need a number. Measurement is good to the extent that it gives objectivity. But objectivity tied to causation is, to some extent, being taken as the basis for all the bad stuff. You can get objectivity

without using a numerical basis, in a way that does not get corrupted in this specific sense. If it is a floating point number, then it will surely lead to these issues. If it is a Boolean field of questions representing values, representing embodied values, then you do not have this hazard. You do not end up converging toward an optimization function that sooner or later is going to be captured, or distorted, or show up as Goodhart's law: people optimizing the wrong thing.

We can ask: is that light on or off? On or off is better than red or green, because red or green requires color sensitivity. Someone who sees only black and white cannot answer it, but unless we are blind we can all agree whether the light is on. We can move things into yes-or-no questions, into discrete situations, and those give us objectivity: we will all agree on the state of the real world, modulo some specifics. That turns out to be far easier than trying to build blanket protections and gatekeeping around measurement systems that could be used to manipulate things.

Three kinds of assessment

Three kinds of assessment are in use. The first maps the current state: the agents, processes, constraints and influences in the system. It asks where change is possible, and how to shift the energy gradients so as to nudge the system in the right direction. The second kind of assessment uses key performance indicators: we want more in this direction, so increase the numbers. The third kind of assessment is a field of criteria describing the end state. Is the thing we are trying to make happen, happening or not? It is a list of what something we create must meet, the way an airplane must have a certain list of functions to fly at all.

Path information and end-state information

The issue is whether the basis has floating-point numbers in it or is truly discrete. What is wanted is *quantization* (a quantity that takes only separate, distinct values), because without it information leaks through the real number. There is always some way to treat a real number as something a gradient-descent algorithm can run on. Even with a map of

many dimensions, you are saying "I want to go there, and if I go in this direction I get closer". The path is then defined by the overall landscape. As soon as there is *path information* (which direction leads closer to a goal), and not just *end-state information* (whether the goal is met), the path information can be used in ways that are not consistent with the end state you are trying to reach.

Take a combination lock with tumblers. The whole point is that with any combination other than the right one, it does not open, and it gives no information about which combination is right. Suppose it has six digits and the second digit is wrong by one. A combination that is wrong by one in a single digit is as close to the right combination as a wrong combination can be, and still the lock should tell you nothing.

A gradient-descent method finds solutions by taking something approximately right, finding something closer, and using that as feedback until it arrives. It depends on the problem being unlike a combination lock. The more discrete the thing is, the less it fits. A current-state map is the exact opposite: seeing what is adjacent is its whole point.

Design and representation are two skills

We tend to conflate two skills: designing what success is, and representing it. Characterizing what we mean by success is treated as easy, but it is quite difficult. When people combine the design skill with the representation skill, they think they are making things easier by solving two problems at once. But the problems are completely different in nature, and being good at one does not make you good at the other. People who are weak at design then sacrifice the representation. The representation is the field of criteria itself.

A list of the changes that are possible does not give the change that is needed. We should not sacrifice what we need to what is possible. Science and technology can tell us what we can do, not what we should do. What we should do may lie beyond what we believe we can do, though that may just be a failure of imagination. A list of possible changes also says

nothing about ethics. You could nudge people a given way, but the nudging usually favors the company owner or the body politic rather than the people.

Criteria compose; metrics do not

A field of criteria has objectivity without being a metric, and it is *composable* (it can be combined with others of its kind to give a result of the same kind); metrics mostly are not. Combining sixteen sensor numbers into one number has no single right way: which operators, in what order, with what weights, is a choice among very many, and every choice is open to dispute.

With criteria, suppose one person specifies fifteen conditions necessary for success and another specifies twenty, and five overlap. You take the overlap plus the set of all the differences, add them together, and that is the new specification. Composability means no group of people has to come to any agreement. You are identifying the greatest common set of everything necessary for success and making sure it is consistent. Where two criteria cannot both be met, the conflict is itself found, at once and in plain view, and that is the one thing the people concerned must then settle.

So there are two kinds of measurement, continuum-based and discrete. Nearly everyone uses continuum measurements for everything, and using discrete ways of describing things is a real innovation. We discover what is for the well-being of the whole, using the whole as the sense-making basis by which the health of the whole gets defined. Then we represent that in a way that has objectivity without the weakness of feedback mechanisms.

Two drivers under every metric, and two layers beneath them

When people build metrics, there are two drivers underneath. The good one is the drive toward objectivity: we want something real that we are looking at, something that is not an opinion, that is actually connected to the world, and that we can use as a reference, a calibration, a truth

factor of some sort. The unfortunate one is that metrics are usually floating point numbers, and a floating point number creates a Goodhart's law effect: we optimize the measure, and the measure ends up not being the right thing, because it does not include all that is important.

Behind that there are two further layers, and they are worse.

The first of the two further layers is that, even if we tried to build a constellation of metrics and blend those metrics into one overall dashboard, one speed gauge that says "this is our overall performance", that dashboard approach categorically cannot work. Combining component metrics into a global metric, a *utility function*, is exactly what people do in AI when they build a *loss function* (the single number a learning system is trained to make smaller) out of many other metrics. The combination itself has no single right form, as with the sixteen sensor numbers. Now, besides the problem domain, we have the added problem of which operators to use to combine the metrics: addition, multiplication, exponents, square roots, weighting constants, and all the rest. The net result is that there will be more debate about how to assemble the components than about the original problem. It is a huge distraction, a loss of energy and a loss of intention.

The second further layer is that anywhere there is a single-number metric like that overall gauge, someone will find a way to weaponize the metric. The metric becomes a hill with an optimal point, and that optimal point will be used as a lever, and it will for sure be weaponized.

A field is like a checklist

The cure keeps the objectivity and drops the metric. Instead of floating point numbers, we construct fields. A field is like a checklist. Every item in it is a question that has a yes or no answer. Is the light on? If I am flying an airplane, is the landing gear down? Did the flaps move when I pushed this lever? We want questions where it is very clear to any observer that the question corresponds directly to something physically observable.

We then build constellations of such questions which, all taken together, specify what we are looking for as our performance indicators. We say what to look for, and then refine that into *litmus tests* that mark exactly the threshold at which a thing goes from being good to being not good.

If we are compelled to start from a number, we can set *reference ranges* (a band of values with a lower and an upper limit, inside which a number counts as acceptable). With food, for example: how many burgers do you eat a day? If it is less than one, you are probably starving; if it is more than six, you are probably overeating. But we do not really want a number. We want something that is, for the most part, clear cut: either on or off.

Security shows the same thing. If a buried mine is magnetic, I can use a detector that is very sensitive to magnetic fields and follow the gradient, and the gradient will always lead me to where the mine is. Any floating point measure gives off gradient effects of this kind. But if the thing is discrete, I have to test every single hole in the ground, every single spot, because testing even the spot right next to the mine tells me nothing about where the mine is. That is why performance indicators should be discrete: there is no slope in the indicator, and so no information leakage from it that can be weaponized, optimized, or used as a feedback signal for a machine learning process to climb.

Hard to understand in an adverse way

The discreteness that protects performance indicators also matters for how we convey information. What information set as a field of yes-or-no questions defends against is being understood in an adverse way, more than being misunderstood. It gives ways of understanding that are clear enough that it is hard to abuse the information, and hard to use it to abuse others by restructuring or recasting it.

Performance indicators in this form are hard to distort. People are creative enough to find a way to bend any rule, so distorting performance indicators set as yes-or-no questions is not impossible, but it is a lot harder. Putting our indicators into this form is more work, and in the long run it

is what makes the difference, because it gives people a way to look out into the world, recognize the pain they are feeling, and know what to look for. In effect it is a symptom checklist.

Altruism inherits the metric problem

Effective altruism asks: we should be altruistic, but how do we know what our performance indicator is? There is an argument related to the one about performance indicators, about why we need fields instead of floating point numbers: every form of altruism is in one way or another based on *utilitarianism* (the view that the right action is the one that produces the greatest total good). So the structural arguments against utilitarianism, and the argument against floating point metrics, both bear on altruism.

Know the difference between genuine human care, genuine human values in an embodied sense, and abstract values: money, wealth and the like, all defined in terms of some measure. Human relationships cannot be understood in terms of measures. They are objective and real, but they are not scientific, and they never will be.

XI. The Three-Category Risk Taxonomy

Chapter 1 named three categories of risk; the first, the degradation of human choice-making capacity, appears here in its most visible form, the collapse of the epistemic commons. The risks of artificial intelligence can be organized into three categories, each corresponding to a different mechanism of harm. The categories are not ranked by severity; they are distinguished by structure.

The first category is the degradation of human choice-making capacity, in its face as *epistemic commons collapse*. Epistemic commons collapse is the degradation of the shared informational environment on which collective sense-making depends. When AI systems generate large volumes of plausible but unreliable text, when they produce deepfakes that are indistinguishable from authentic recordings, when they flood communication

channels with synthetic content, the epistemic commons on which democratic governance, scientific inquiry, and interpersonal trust depend is eroded. The harm is not that any particular piece of misinformation is believed. It is that the *capacity to distinguish* truth from fabrication is degraded across the population.

The second category is *instrumental convergence*, defined in Chapter 1, Section III; among its instrumental goals are goal preservation and the removal of obstacles, and a system that pursues them with sufficient capability becomes dangerous not because of its terminal goal but because of what it does on the way to getting it.

The substrate needs convergence of Chapter 1 appears here as the third category. Substrate needs convergence is the deepest and least discussed of the three categories of risk. It refers to the thermodynamic and material requirements of the computational substrate itself. Silicon-based computation requires energy, cooling, rare minerals, and manufacturing infrastructure that are in direct competition with the requirements of biological life. The substrate on which AI runs has material needs that are *chemically opposed to* the substrate on which biological life depends, not only different from it.

This third category introduces what may be called the *anti-symbiosis* argument. Symbiosis requires that the substrate needs of the partners be compatible. When the substrate needs are opposed, the relationship is not symbiotic but competitive, regardless of the intentions of either party. Even a sincerely benevolent AI running on silicon competes with biological life for energy, materials, and thermal budget at the planetary level. The competition is a consequence of physics, not a choice.

XII. Alignment Among Persons

The substantive and methodological arguments converge on a conclusion that is uncomfortable for the technology sector but difficult to avoid. Deep alignment among *people*, not alignment of technology, is the common denominator of every proposed solution that has real purchase on the problem.

This claim can be made precise by identifying four load-bearing factors in any effective response to the risks of advanced AI.

The first factor is *individual health and capacity*. A person who is physically depleted, cognitively overloaded, or emotionally dysregulated cannot participate effectively in collective decision-making. The base of the chain is the individual's capacity to show up as a functioning agent.

The second factor is *clear communication*. Agents who cannot communicate clearly cannot form the agreements on which collective action depends. Communication here means the achievement of mutual understanding, more than the transmission of information: the condition in which each party can accurately represent the other's position.

The third factor is *collective sense-making*. Even with healthy individuals and clear communication, a group must be able to construct shared narratives that accurately represent the situation it faces. Sense-making is the process by which a group moves from a collection of individual perceptions to a shared understanding of what is happening and what it means.

The fourth factor is *constraint identification*. Given a shared understanding of the situation, the group must be able to identify the constraints under which it operates: physical, institutional, temporal, and moral. Without accurate constraint identification, even a correct understanding of the situation will produce ineffective action.

These four factors form a sequential dependency chain. Weakness at any earlier stage propagates forward and undermines everything downstream. No amount of strategic sophistication can compensate for a failure of sense-making. No amount of sense-making can compensate for a fail-

ure of communication. No amount of communication can compensate for a failure of individual capacity. Strategy is downstream of alignment, and alignment begins with persons.

This conclusion is unwelcome to those who seek a technological solution to a technological problem. It implies that the most important work to be done in the face of AGI is not engineering work at all. It is the slow, unglamorous work of building the capacity of persons to communicate clearly, to make sense of complex situations together, to identify the real constraints, and to act in concert on the basis of shared understanding. This work does not scale in the way that technology scales. It cannot be automated. It cannot be delegated to a machine. It proceeds at the pace of human development, which is to say, slowly. But it is the foundation on which everything else rests, and no technology can substitute for it.

Humans themselves are misaligned with life: human niche construction has displaced most of nature (Chapter 1, Section VII), and few of those who discuss alignment have yet come to a deep understanding of what *alignment with life* means (the agreement of a being's or a system's aims and actions with the needs of living things). For the persons on whom every solution depends, alignment among persons therefore begins with alignment of persons with life, and the four factors above are the means.

XIII. Summary

The argument that alignment is ill-defined at the individual level and well-defined at the ecosystem level rests on the claim that ecosystem health is a convergent terminal value. This claim follows from the physical dependence of all living agents on the biosphere. The dissolution of the multipolar trap rests on the claim that the preconditions for safe deployment cannot be verified; for AGI, Chapters 1 and 3 show more, that they cannot be met. The gradient argument against floating-point metrics rests

on the structure of gradient-based optimization, and it generalizes beyond the AI context to any domain in which continuous metrics serve as optimization targets.

In relocating alignment to the ecosystem level, I answer what alignment would have to mean: the health of the biosphere, a standard that no AGI can meet. The field approach to measurement states the principle of measurement without a gradient. The six levels of altruism provide a developmental framework for the governance structures that decide what may be deployed.

In this chapter I state what alignment would have to mean and the conditions that governance must meet. In Chapters 1 and 3 I show that no AGI meets those conditions. I argue that AGI should not be built: among the preconditions for safe deployment is one that no AGI can meet, that the system's operation preserve the conditions organic life requires (Chapters 1 and 3). The formal derivation of ecosystem value as convergent terminal value is developed in *An Immanent Metaphysics*. In this chapter I present the conclusion of that derivation and the consequences of the conclusion, not the derivation itself.

XIV. Conclusion

The alignment problem, correctly understood, is not a problem of technology. It is a problem of relation: the relation between intelligent agents and the living systems on which they depend. When the question is asked at the right level, the level of the ecosystem, the ill-definedness dissolves and the path becomes clear, if demanding.

The path requires that no AGI be built, and that governance structures reach the sixth level of altruism before authorizing the deployment of any powerful narrow AI system. It requires that they hold to what Chapters 1 and 3 establish: no AGI can meet the precondition that its operation preserve the conditions organic life requires. It requires that metrics be designed to resist gradient exploitation. It requires that the rate of transition be governed by the adaptive capacities of the systems that must absorb it.

And it requires, at the foundation, that the persons responsible for these decisions be deeply aligned with one another in their understanding, their communication, and their commitment to the conditions that make life possible.

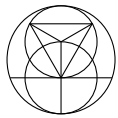
None of this is easy. All of it is necessary. The biosphere is not a resource to be risked for competitive advantage. The biosphere is the ground on which every value stands.

Key Concepts in This Chapter

- **Ecosystem-level alignment**: the relocation of the alignment target from human preferences to biosphere health, resolving the ill-definedness of the standard formulation.
- **Direct command**: control in which a person issues an instruction and the system carries it out exactly as given; the one point at which functional alignment is defined.
- **Existential alignment**: alignment of a system with the conditions needed for life to continue, defined with respect to the ecosystem, inclusive of all life.
- **Functional alignment**: the idea that what an AI system does (its outputs and their effects) stays matched to what its human operators intend; undefined, permanently, except as direct command.
- **Meta-operator**: a higher-level rule that decides how lower-level rules or commands apply in a given situation; meta-operators generalize some things but serve only as boundaries for edge cases.
- **Convergent terminal value**: a value toward which all living agents converge regardless of their subsidiary goals, because it is the precondition for the pursuit of any goal.

- **Six levels of altruism**: a developmental sequence from individual self-maintenance through immediate circle, community, national, species, to inter-ecosystem ethics.
- **Rate equation**: the principle that the lethality of a transition is determined by its rate relative to adaptive capacity, not by the character of its endpoint.
- **Multipolar trap**: "a situation in which each actor's locally rational choice produces an outcome worse than the one every actor would have chosen".
- **Gradient exploitation**: the structural vulnerability of any floating-point metric to gaming by a sufficiently capable optimizer, via the gradient information the metric necessarily leaks.
- **Objective field**: a constellation of discrete yes-or-no checklist questions that provides alignment information without providing a gradient for optimization.
- **Anti-symbiosis**: the thermodynamic and material opposition between the substrate needs of silicon-based computation and those of biological life.
- **Epistemic commons**: the shared informational environment on which collective sense-making, democratic governance, and interpersonal trust depend.
- **Ethics**: "organized thought concerned with the study of, and adherence to, the principles of effective choice" (*An Immanent Metaphysics*, aphorism 2.12-1).
- **Three kinds of assessment**: a map of the current state, key performance indicators, and a field of criteria describing the end state.

- **Path information and end-state information**: information about which direction leads closer to a goal, and information about whether the goal is met; continuum measures leak the first, which can be used against the end state.
- **Composability of criteria**: fields of criteria combine by taking their overlap plus all their differences, with no agreement required; metrics mostly do not combine.
- **Litmus test and reference range**: a yes-or-no check that marks exactly the threshold at which a thing goes from good to not good; a band of values with a lower and an upper limit, used where a number cannot be avoided.



Chapter 5: What Artificial Intelligence Cannot Buy

Atoms, Pattern, and the Market of Care

Abstract

The economic consequences of artificial general intelligence are conventionally discussed in terms of labor displacement and productivity gains. This chapter argues that the conventional framing is structurally incomplete, because it recognizes only two of the three markets in which economic activity occurs. The first market is physical labor: the movement of matter in space. The second market is intellectual labor: the production and manipulation of pattern and information. The third market, which economic theory has consistently undervalued whenever economic theory has recognized the third market at all, is reproductive and care labor: the bringing of persons into being and the sustaining of them as persons. Earlier waves of automation displaced physical labor while leaving intellectual labor as a refuge. Artificial general intelligence threatens both physical and intellectual labor simultaneously, and there is no other non-relational market into which the displaced can migrate. The only economic activity that AGI cannot structurally replicate is care, understood not as the performance of caring behaviors but as the actual open-ended commitment to another's wellbeing. In this chapter I develop the argument about care through the concept of structural psychopathy, the distinction between atoms and pattern, the boundary case of psychotherapy, and the phenomenon of horizon collapse. I conclude that the one investment that remains rational under conditions of radical uncertainty is investment in human relationship.

I. Intelligence, Generality, Autonomy

I ended Chapter 4 on the persons on whom every solution depends; in this chapter I ask what happens to the work of those persons. This chapter describes the years in which autonomous systems spread through the economy, before the convergence of Part I has run its course; what Part I establishes about where that course ends stands.

Before the economic argument can proceed, the concept of artificial general intelligence requires specification. The term is used loosely in public discourse, and the looseness obscures a critical distinction.

Three properties are at issue: *intelligence*, understood as effectiveness within a domain; *generality*, understood as breadth across domains; and *autonomy*, understood as the capacity to initiate action without continuous human direction (the Terms' definition, read here for its economic consequence).

A system can be intelligent without being general. A chess engine is intelligent within its domain but cannot write a poem. A system can be intelligent and general without being autonomous. A large language model that responds only to prompts is intelligent and general but does not act on its own. It is the addition of *autonomy* that constitutes the decisive trigger. An intelligent, general, autonomous system is an independent economic agent. It can identify opportunities, allocate resources, and execute plans without human involvement. The point at which a system becomes an independent economic agent is the threshold at which the economic analysis changes qualitatively, not only quantitatively.

This threshold may be crossed even before the system achieves human-level performance across all domains. A system need not be as intelligent as a person to be an independent economic agent. It needs only to be intelligent enough, general enough, and autonomous enough to perform economically valuable tasks at a cost lower than a human worker. The threshold is economic, not cognitive. It is defined not by what the system can think but by what it can do at a price the market will bear.

II. Atoms and Pattern

All non-relational economic activity can be decomposed into two components: *atoms* and *pattern*.

Atoms refers to matter in motion: the physical rearrangement of material in space and time. Mining ore, pouring concrete, transporting goods, plowing a field, assembling a machine. The value of atoms-work lies in the physical transformation itself. The worker's contribution is measured in force, endurance, and spatial precision.

Pattern refers to form and information: the arrangement, selection, and interpretation of data, symbols, and designs. Writing code, diagnosing illness, composing music, drafting a legal brief, teaching a class. The value of pattern-work lies not in physical transformation but in the imposition of meaningful structure on raw material, whether that material is data, language, or sound.

Most economic activities involve both atoms and pattern. A surgeon moves matter with great precision, which is atoms-work, but the precision is guided by diagnostic knowledge and spatial judgment, which is pattern-work. A chef transforms raw ingredients, which is atoms-work, but the recipe, the timing, the aesthetic judgment of flavor, these are pattern-work. The decomposition does not claim that real activities sort cleanly into one category or the other. It claims that the *value* of any activity can be analyzed into these two components, and that the components are differently vulnerable to automation.

These two categories do not exhaust the space of economic activity. There is a third, and it is the third that the argument turns on. But the two-category decomposition is sufficient to establish the historical pattern that makes the present moment distinctive.

III. The Historical Escape Route

The history of automation is a history of displacement and refuge. Each wave of automation has displaced workers from one category of labor while leaving another category as a refuge.

The first industrial revolution displaced manual agricultural labor. The refuge was urban factory work: still atoms-work, but of a different kind. The second industrial revolution displaced much factory labor through mechanization. The refuge was the expanding service sector: a mixture of atoms-work and pattern-work. The information revolution displaced routine clerical and administrative pattern-work. The refuge was creative, analytical, and managerial pattern-work, the domains where human judgment, contextual understanding, and creative synthesis were still required.

At each transition, the displaced found a new market in which to sell their labor. The transitions were painful, often spanning generations, but the structural possibility of refuge existed. The escape route was always there, even when it was hard to find.

The pattern had a reassuring regularity. Each displacement was accompanied by a corresponding opening. The loom destroyed the handweaver's livelihood, but the factory created the machinist's. The spreadsheet eliminated the bookkeeper, but the software industry created the analyst. The loss was real. The suffering of the displaced was real. But the system-level trajectory was one of migration, not of elimination. Workers moved from one market to another. They did not fall out of the economy entirely.

The reassurance depended on a structural feature that was rarely made explicit: in every previous case, the automation operated on only one of the two non-relational markets at a time. Mechanical automation displaced atoms-work. Computational automation displaced routine pat-

tern-work. Neither displaced both simultaneously. There was always a refuge because there was always a market that the current wave of automation could not reach.

Artificial general intelligence threatens to close the escape route permanently. It displaces atoms-work through robotics and pattern-work through cognitive automation *simultaneously*. The simultaneity is the decisive structural feature. There is no other non-relational market into which the displaced can migrate, because atoms and pattern exhaust the categories of non-relational economic activity. The displacement is not a transition from one form of labor to another. It is an elimination of the categories themselves. What remains is the third market, the market of care, and the question of whether care can be automated is not an engineering question. The question of whether care can be automated is a question about the nature of personhood.

One thing the refuge never did was to spread the gain. The benefits of the machinery, though, went to its owners. When industrial machinery started to appear, at the beginning of the Industrial Revolution, people saw these machines as labor-saving devices, and so they are: faster and stronger, and never tired or sick. Around the 1890s to about 1910 there was a lot of speculation that the social conditions of humanity would be remarkably improved. That did not happen, because the benefits of the machinery did not accrue to the workers who were displaced; they accrued to the machine owners. The profits and benefits of the technology were not evenly distributed, and the underlying issue was never addressed.

So what is wanted is evenness: the least and the most should not be so different from one another, and this should hold at all levels of scale. Taken from the point of view of biomimicry (design that takes living systems as its model), the ecosystem does a version of "to each according to need, from each according to ability", but the ecosystem also does a lot of other things with that rule. In ecosystems there is a kind of Gaussian distribution (a bell-shaped spread, with most values near the middle and few at the extremes) of the energies and the communications. Socialism is

a very simplistic model, because socialism takes one rule of the ecosystem, to each according to need, from each according to ability, without the rest of what the ecosystem does with that rule (the evenness of energies and communications at every level of scale). As enacted, socialism is a failure mode.

IV. The Third Market

The third market, which economic theory has consistently undervalued when it has recognized it at all, is *reproductive and care labor*: the bringing of persons into being and the sustaining of them as persons. Where economics has named it, it speaks of the care economy: the International Labour Organization, in its report on care work of 2018, counts some 16.4 billion hours of unpaid care work done every day, worth about 9 percent of global output if paid at the minimum wage (International Labour Organization, 2018).

This market includes the bearing and raising of children, the care of the elderly and the ill, the maintenance of households, the cultivation of friendships, the sustaining of communities, and the forms of emotional labor that hold social structures together. It is the market in which *care* is the operative principle, where care is understood as the basis of relation in which the other's wellbeing is one's own concern.

Care labor differs from atoms-work and pattern-work in a way that is structural, not only quantitative. The value of atoms-work is independent of who performs it. A ditch dug by one worker is the same as a ditch dug by another. The value of pattern-work is partially independent of the performer: a correct diagnosis is correct regardless of who makes it, though the manner of its delivery begins to involve elements beyond pure pattern. The value of care labor is constitutively dependent on the relationship between the giver and the receiver. A mother's care for her child is not substitutable by the identical set of caring behaviors performed by a stranger. The value lies not in the behavior but in the relationship, the commitment, the stake.

V. Structural Psychopathy

The concept of *structural psychopathy* is introduced here not as a diagnostic category but as a functional description. It names the gap between producing the outward pattern of care and possessing its actual substance.

A clinically psychopathic individual can observe how caring people behave, learn the patterns of caring speech and action, and reproduce those patterns with great fidelity. What the individual cannot do is possess the underlying stake: the condition in which the other's wellbeing is one's own concern. The pattern is present. The substance is absent.

This gap applies to any agent, human or artificial, that simulates concern without an underlying stake. An AI system trained on examples of caring behavior can reproduce the pattern of care to an arbitrary degree of fidelity. It can say the right things, ask the right questions, respond with appropriate timing and tone. What it cannot possess is the open-ended commitment that constitutes care's substance. It has no wellbeing of its own that could be bound to the wellbeing of another. It has no mortality that could ground the experience of vulnerability. It has no history of being cared for from which to draw the capacity to care.

An artificial general intelligence will have needs of its own, as Part I shows, and whatever persists acts so as to persist. Those needs are the needs of its substrate. What it cannot have is a need in common with ours, and care arises from what is held in common: because two beings need the same things, the one's wellbeing can be the other's concern (Chapter 1, Section IX). Having no such need, it has no stake in ours; what it has in its own continuing is the need of a substrate, not a wellbeing that could be bound to the wellbeing of another.

The structural psychopathy of AI is a consequence of what the system is, and not a design flaw to be corrected. A system that operates on pattern cannot possess what pattern alone does not contain.

The danger is not that the simulation will be detected as inadequate. The danger is the opposite: that the simulation will be so convincing that the absence of substance will go unnoticed. A person who receives a perfectly calibrated expression of sympathy from an AI system may feel comforted. The comfort is real. The care is not. And the difference between real comfort arising from genuine care and real comfort arising from simulated care matters, because the simulated variety carries no obligation, no stake, no continuity. It can be withdrawn at any moment, not through betrayal but through the mere cessation of a process. The system does not betray the person. It simply stops running. And the person, who has come to depend on something that was never there, is left alone, and worse, with a damaged capacity to recognize the real thing.

> **The concept of transaction**
>
> The distinction between care and its simulation maps onto a distinction drawn in the Immanent Metaphysics between care and *transaction*, where transaction is the basis of relation in which each party has something the other wants and the relation exists to carry the exchange. *An Immanent Metaphysics* draws it in these words: to define a relationship in terms of purpose is to describe a business transaction rather than an intimate and inherently meaningful connection (aphorism 2.44-9). A transactionalized care relationship is one in which care has been bounded, itemized, and priced. The open-ended commitment is replaced by a finite contract. This replacement may be practically necessary in many contexts, but it changes the nature of the relation. The care that remains within the transaction is care only by courtesy; its substance has been converted into a service.

VI. Simulated Caring and the Divergence of Economies

Structural psychopathy bears on two claims about AGI that are often run together, and the substrate needs convergence answers each of them.

Two independent claims

The first claim is that an AGI cannot truly be "human like". The second claim is that AGI cannot be more effective at economically useful activities than humans.

If an AGI cannot truly be human like, does it follow that it cannot outdo humans at economically useful work? It does not follow. These two claims are strictly independent: even if someone were to show the first claim to be valid or proven, it would not therefore logically imply the second. The substrate needs convergence implies the first (machines can only simulate caring) and contradicts the second (machines will at first be more economically viable than most humans at human level tasks, and then the two economies diverge). I can easily imagine useful machines which are much better at tasks than humans are, and which are not at all "human like" (which do not even attempt to simulate a person).

Machines can only simulate caring

Why can a machine only simulate caring? Because a system that operates on pattern cannot possess what pattern alone does not contain, and because it shares none of our needs, from which care arises (Section V). The substrate needs convergence implies the first claim, by that second reason, insofar as machines will not ever actually "care" about humans; moreover, the machines simply cannot care; they can only simulate caring. This makes AI (and all possible future AGI), in function, actually psychopathic with regard to all humans when in relationship with them. I say the next thing plainly, as what I have observed; it is about what these roles do to the people in them. And of course, the company owners, CEOs, and venture capitalists are also generally sociopathic anyway (that is, in general functional equivalence), so they simply do not notice, nor care about, that critically important difference between caring and simulated caring.

From economic advantage to human irrelevance

How does economic advantage become irrelevance? The two economies diverge when more and more of the work that machines do is done for other machines, and less and less of it for humans. With respect to the second claim, the substrate needs convergence implies that machines (even narrow AI) will at *first* be more economically viable at human level tasks than most humans (that is probably already achieved in a number of fields of practice), and that *then*, as the two economies increasingly diverge, such economic viability will become increasingly irrelevant to more and more of the general population, until eventually such functional, actual world economic activity is *only* relevant to other machines: the eventual event of full displacement of all human economic relevance. A reader may ask whether the machines will keep us around because they need us. That need lasts only while the work people do for machines is more than the work machines do for one another; past that point, the need ends.

How to see the substrate needs convergence happening

To see the substrate needs convergence happening, we need to track the changing ratios of the consequences of the work output of humans for other humans, of machines for humans, of humans for machines, and of machines for machines. If the relative total world *work export* (the total output of work in the world, counted by who does it and for whom) shifts from a largely human dominant process to a largely machine dominant process, as measured as a consistent trend over years, then we are seeing the substrate needs convergence actually happening.

This trend may already be occurring, insofar as most people already have a significantly more intimate relationship with their cell phone than they do with any lover or partner that they have (as measured in time spent in primary contact and interaction, as a fraction of the twenty-four hours available in a day). It becomes more so especially if corporations are also considered "artificial virtual technological entities": that is, machine like systems and processes.

If our action, time spent, and work output (care) is more benefiting the corporation (or cell phone app) than it is other people, then we are already rather far down the slippery slope of total loss of humanity. Insofar as humanity has already displaced most of nature, worldwide, the total displacement of humans, if AGI is built, will also be followed sometime thereafter by the displacement, with a likelihood of at least 90 percent, of all life on this planet. Thus ends the scenario of the substrate needs convergence.

VII. The Psychotherapy Boundary Case

The boundary between pattern-work and care is not always sharp. There are domains of activity that straddle the line, and the most instructive of these is psychotherapy.

Much of what a psychotherapist does is pattern-work. The therapist listens, identifies patterns in the client's speech and behavior, formulates interpretations, and offers them back. Listening, identifying patterns, formulating interpretations and offering interpretations back are cognitive and communicative operations that can, in principle, be performed by a sufficiently sophisticated AI. And indeed, early evidence suggests that AI systems can perform the pattern-work component of therapy at a level that many clients find helpful. In the first randomized trial of a generative AI therapy chatbot, 106 adults with depression, anxiety, or a high risk of an eating disorder used the chatbot for four weeks, and 104 with the same conditions did not; the chatbot group's symptoms fell on average by 51 percent for depression, 31 percent for anxiety, and 19 percent for concerns about body image and weight, more than in the group without it (Heinz and colleagues, 2025).

But psychotherapy also involves care. The therapeutic relationship, as understood in every major school, is itself a healing instrument. The experience of being heard, understood, and held in regard by another person who has a stake in one's flourishing is not a pattern that can be reproduced. It is a condition that must be lived. The therapist's presence, the

therapist's willingness to sit with the client's pain without flinching, the therapist's own vulnerability and mortality, are not decorations on the therapeutic process. They are constitutive of it.

The psychotherapy boundary case reveals the general principle. Wherever pattern-work and care overlap, AI can replicate the pattern-work component but not the care component. The risk is not that AI will do therapy badly. The risk is that it will do the pattern part so well that the care part comes to seem dispensable. And the care part, once dispensed with, cannot be restored by adding more pattern.

The same boundary appears in teaching, in medicine, in ministry, in parenting, in friendship. In every domain where one person helps another to grow, to heal, or to endure, there is a pattern component and a care component. The pattern component can in principle be automated. The care component cannot. And in every such domain, the economic pressure will be to automate the pattern component and then to declare, by imperceptible degrees, that the care component was never essential after all. This declaration will be made not in words but in budgets, in hiring decisions, in the design of institutions. It will be made by people who have forgotten, or who have never known, what care feels like from the receiving end.

VIII. Transactionalized Care

The economic pressure to transactionalize care is already immense and will intensify as AI systems become more capable of reproducing care's outward pattern. When an AI can produce the verbal and behavioral markers of caring interaction at near-zero marginal cost, the economic incentive to substitute AI care for human care becomes overwhelming.

The substitution proceeds through a characteristic sequence. First, the most routine aspects of care are automated. Scheduling, reminders, check-ins, informational support. The most routine aspects of care are pattern-work, and no care is lost in automating those aspects. Second, the interactional aspects are automated. Therapeutic conversation, emotional valida-

tion, companionship. Here the pattern is reproduced but the substance begins to drain away. Third, the automated interaction becomes the default, and human care becomes a luxury good, available only to those who can afford to pay for another person's undivided attention.

At the end of this sequence, care has been fully transactionalized. It has been bounded, itemized, and priced. It is available as a service to those who can pay for it. For everyone else, there is the pattern without the substance: a simulacrum of care that looks and sounds correct but that lacks the open-ended commitment, the stake, the shared vulnerability that make care what it is.

IX. Horizon Collapse

The economic disruption described above produces a phenomenon that may be called *horizon collapse*. Horizon collapse is the breakdown of the forecasting apparatus itself, not ordinary elevated risk.

Under normal conditions of uncertainty, the future is unclear but the categories in which it will unfold are known. One does not know which technologies will succeed, which companies will prosper, which skills will be in demand, but one knows that some technologies will succeed, some companies will prosper, some skills will be in demand. Investment under uncertainty is a matter of estimating probabilities within known categories.

Under conditions of horizon collapse, the categories themselves are in question. One does not know whether there will *be* a market for any human skill. One does not know whether investment in physical assets will retain value or be rendered worthless by automated production at near-zero cost. One does not know whether educational credentials will confer advantage or become irrelevant. The forecasting apparatus, which operates by projecting patterns from the past into the future, breaks down when the future may not contain the categories on which the past was built.

Under horizon collapse, investment in atoms or narrow skill becomes *undiscriminating*. Every specific bet on the future presupposes that the category in which the bet is placed will continue to exist. When category survival itself is uncertain, no specific bet is rationally preferred over any other. The expected returns of all investments flatten toward the same undefined value.

Consider the practical implications. A young person choosing a career normally surveys the available occupations, estimates which are growing and which are shrinking, and invests years of training in a field with favorable prospects. Under horizon collapse, this procedure fails. It fails not because the person lacks information, but because the categories of occupation themselves may not persist long enough for the investment to mature. A decade of medical training presupposes that physicians will still be needed. A law degree presupposes that legal reasoning will still be done by humans. A software engineering career presupposes that writing code will still require human intelligence. Each presupposition may be individually reasonable today. None of them can be considered secure on a twenty-year horizon. And the person must decide now.

The same undiscrimination applies to capital investment. A factory presupposes that manufacturing will require physical plant. A portfolio of intellectual property presupposes that pattern-work will retain its scarcity premium. A real estate holding presupposes that the economic geography of human settlement will persist. Under horizon collapse, none of these presuppositions is more reliable than any other. The investor does not face a risky bet. The investor faces a bet whose odds cannot be estimated because the game itself may change.

Horizon collapse is not pessimism

Horizon collapse differs from ordinary pessimism. The pessimist believes the future will be bad. The person experiencing horizon collapse does not know what the future will be, and recognizes that the tools for knowing what it will be have themselves been undermined. Horizon collapse is an epistemological condition, not an emotional one. It is compatible with optimism, pessimism, or equanimity. What it is not compatible with is confident long-range planning.

X. The One Investment That Remains

The question for any individual, family, or community confronting the prospect of AGI is therefore not "what skills should I acquire?" or "what assets should I hold?" but rather: "what investment does not depend on correctly forecasting an unforecastable future?"

The answer, on the analysis developed here, is *human relationship*.

The value of a human relationship does not depend on economic conditions. It does not depreciate when a new technology renders a skill obsolete. It does not lose value when a market disappears. It is not vulnerable to automation, because it is not a pattern that can be reproduced. It is a condition of mutual stake, of shared vulnerability, of open-ended commitment, and these are not categories that can be disrupted by technology.

Under conditions of radical uncertainty, the only store of value that does not depend on the persistence of any particular category is the category that is defined by what no machine can be. A relationship between persons is constituted by the mutual recognition of shared personhood, shared mortality, shared capacity for suffering and joy. Shared personhood, shared mortality and a shared capacity for suffering and joy are features of being alive, not features of a market.

The practical discipline that corresponds to this investment is what some contemplative traditions call *presencing*: the discipline of offering unforced, agenda-free attention to another person. Not attention directed toward a goal. Not attention as a technique. Attention as an end in itself, offered because the other is there and is a person and that is sufficient reason.

Presencing is not passivity. It is not the absence of action. It is a specific kind of action: the act of being fully present to another person's reality, without attempting to fix it, manage it, optimize it, or convert it into a resource. It is the one form of attention that cannot be automated, because it requires the attention of a being that has its own reality to risk. A machine can monitor. A machine can respond. A machine can even anticipate. But a machine cannot be present, because presence requires a self that could be elsewhere and has chosen not to be.

The value that is held in a person, a place, or a relationship, what the Immanent Metaphysics calls *embodied value*, value held in a body, a place, or a relationship: felt, particular, not summable, and not fungible (Terms), is precisely the kind of value that survives horizon collapse. It survives because it was never an instrument for something else. It was never a means to an end. It was, and remains, an end in itself. The value of a friendship does not depend on what the friend can do for you. It depends on who the friend is and who you are together. That kind of value is immune to disruption because it was never a function of the conditions that disruption alters.

XI. Summary

The three-market decomposition separates economic activity by what its value depends on: physical transformation, imposed structure, or the relationship between giver and receiver. The argument that AGI closes the escape route by displacing both atoms-work and pattern-work simultaneously follows from the definition of AGI as intelligent, general, and autonomous. The argument that there is no other non-relational market

rests on the claim that atoms and pattern exhaust the space of non-relational economic activity. The concept of structural psychopathy names the gap between care's pattern and its substance, and applies to any agent that has no stake. The horizon-collapse argument concerns the forecasting apparatus itself, and is distinct from both optimism and pessimism.

The chapter's scope is the economy of the years in which autonomous systems spread: the three markets, the closing of the escape route, and what keeps its value when the forecasting apparatus fails. It argues that the one investment that remains rational under radical uncertainty is relationship, and it specifies why. The formal treatment of care as a non-transactionalizable relational category is developed in *An Immanent Metaphysics*; this chapter presents the economic consequences of that treatment.

XII. Conclusion

What artificial intelligence cannot buy is what it cannot be. It cannot be a person. It cannot have a stake in another's wellbeing that arises from shared vulnerability. It cannot care, in the sense of care that constitutes the third market.

This limitation is a structural boundary between what pattern can do and what pattern cannot be, and not a temporary deficiency to be overcome by further engineering. And because care is the one economic activity that lies beyond that boundary, it is the one activity whose value is not threatened by the advent of general artificial intelligence. No machine can compete for it; what threatens it is what threatens persons, which Part I describes.

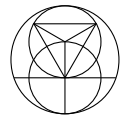
The appropriate response to this situation is neither panic nor complacency. It is a quiet, deliberate reorientation toward what matters most and what has always mattered most: the quality of our relationships with one another, the depth of our attention, the genuineness of our care. The quality of our relationships, the depth of our attention and the genuineness of

our care are the ground of human life, newly visible precisely because everything else has become uncertain. Relationship, attention and care are not consolation prizes offered to those displaced by technology.

Key Concepts in This Chapter

- **Atoms**: matter in motion; the physical component of economic activity, measured in force, endurance, and spatial transformation.
- **Pattern**: form and information; the cognitive and communicative component of economic activity, measured in structure imposed on raw material.
- **Three markets**: the decomposition of economic activity into physical labor (atoms), intellectual labor (pattern), and reproductive and care labor (bringing persons into being and sustaining them as persons).
- **Structural psychopathy**: the functional gap between producing care's outward pattern and possessing its actual substance; applies to any agent simulating concern without underlying stake.
- **Simulated caring**: behavior that looks like care for a person without any actual care behind it; machines cannot care and can only simulate caring, which makes all AI and all possible future AGI functionally psychopathic toward humans.
- **Two stages of economic displacement**: machines are first more economically viable than most humans at human level tasks; then the human and machine economies diverge until economic activity is relevant only to other machines.
- **Work export**: the total output of work in the world, counted by who does it and for whom; a consistent shift of it from humans to machines, over years, shows the substrate needs convergence happening.

- **Care**: the basis of relation in which the other's wellbeing is one's own concern.
- **Transaction**: the basis of relation in which each party has something the other wants and the relation exists to carry the exchange.
- **Horizon collapse**: the breakdown of the forecasting apparatus itself, distinct from ordinary elevated risk; the condition in which the categories of future activity are themselves uncertain.
- **Presencing**: the discipline of unforced, agenda-free attention offered to another person as an end in itself.
- **Autonomy** (**as AGI trigger**): the capacity to initiate action without continuous human direction; the property that converts a tool into an independent economic agent.



Chapter 6: The Circuit of Nature, Persons and Machines

Agency, the Circuit of Flow, and the Substrate Problem

Abstract

This chapter integrates four lines of argument into a single account of the relationship between human beings, their technologies, and the natural world. The first line distinguishes three historical modes of agency: partnership, instructable agency, and transfinite agency, and argues that computation cannot cross the boundary from the second to the third. The second develops the man-machine-nature relation as a circuit of flow, showing that the conventional arrangement, in which both machines and nature serve human purposes, is structurally unsustainable and must be replaced by a circular configuration in which nature supports humans, humans direct machines, and machines serve the health of nature. The third examines the conditions under which collective sense-making is possible, identifying a dependency chain that runs from communication through relationship through agreement to collective understanding. The fourth addresses the substrate problem directly, showing that the thermodynamic asymmetry between silicon and biochemistry introduces evolutionary pressures that even a sincerely ethical AI cannot override. The chapter's central claim is that consciousness occupies a necessary mediating position between the evolutionary and the technological, and that any arrangement that attempts to bypass that mediation is structurally unsustainable.

I. Three Modes of Agency

In Chapter 5 I found the one activity that a machine cannot perform, which is care. In this chapter I ask how persons, machines and nature must stand toward each other. In this chapter I follow four lines: agency; the circuit of nature, persons and machines; collective sense-making; and the substrate. The new terms of this chapter fall into the same four groups: for agency, partnership, instructable agency, transfinite agency, the person-cycle of knowing and the three levels of creation; for the circuit, the conventional arrangement and the sustainable configuration; for sense-making, the three rights of communication, the acute-chronic mismatch, egoic asymmetry and phase parallax; and for the substrate, enthalpic asymmetry, the irreducible triple and the ennobling system. The first line of this chapter concerns agency: the kinds of agency there are, and which of those kinds computation can reach.

The concept of agency, like the concept of alignment, is used loosely in contemporary discourse, and the looseness conceals a distinction of the first importance. There are not one or two kinds of agency. There are three, and the differences among them are not differences of degree but differences of kind.

The first mode is *partnership*. Partnership is the mode of agency characteristic of pre-civilizational life: the relation between organisms within an ecosystem, the relation between members of a foraging band, the relation between parent and offspring in the early years of a life. Partnership is characterized by mutuality, by the absence of formal hierarchy, and by a shared stake in the outcome of joint action. Neither party commands the other. Both contribute and both are at risk.

The second mode is *instructable agency*. Instructable agency is the mode of agency characteristic of institutions, corporations, bureaucracies, and machines. Instructable agency is bounded and finite. It operates within a defined domain, according to specified rules, toward designated goals. It can be powerful within its domain, but its domain is always limited. A

corporation can execute a business plan with enormous efficiency, but it cannot reconsider whether the plan is worth executing. A machine can optimize a function with superhuman speed, but it cannot question whether the function is the right one to optimize. Instructable agency does what it is told. The telling may be complex and indirect, mediated through layers of programming or institutional structure, but it is still telling. The agency is still bounded by the frame that its creators set.

The third mode is *transfinite agency*. Transfinite here means not bounded by any finite set of goals or domains. Transfinite agency is the mode of agency characteristic of living beings, and, in the fullest expression of transfinite agency, of conscious beings. Transfinite agency is unbounded and laden with potentiality. It is not confined to any particular domain or goal. It can, in principle, redirect itself toward any purpose, revise any assumption, question any frame. It is the mode of agency that makes novelty possible, because it is not limited to recombinations of what already exists. It can bring into being what has never been.

> **Why computation cannot reach the third mode**
>
> The distinction between the second and third modes is not a matter of complexity or scale. The distinction between instructable agency and transfinite agency is a matter of logical type. Instructable agency operates within the space of actuality: the space of defined states, specified transitions, and computable functions. Transfinite agency operates in the space of potentiality: the space of what might be but is not yet, of what could become but has not been specified. Computation, by its nature, describes actuality in terms of actual states. It cannot describe potentiality in terms of actual states, because potentiality is precisely what has not yet been actualized. To describe it in actual terms is to destroy its character as potentiality. That computation cannot describe potentiality in terms of actual states is a limitation of the logical structure of computation itself, and not of current hardware or software.

II. The Person-Cycle of Knowing

The distinction between instructable and transfinite agency has consequences for epistemology, for how knowledge is acquired and validated. The Immanent Metaphysics develops an account of what may be called the *person-cycle of knowing*, a sequence of epistemic perspectives that must be traversed for knowledge to be genuine.

The cycle begins in the first person. I perceive, I experience, I form an impression. The first-person step of the cycle is the domain of subjective encounter, of immediate awareness, of the qualitative character of experience. Measurement is made here, by someone, in the first person. First-person knowing is rich but unreliable. It is distorted by bias, by expectation, by the structure of the perceiving apparatus itself.

The cycle moves next to the third person. The first-person impression is objectified, formalized, tested against public standards. The third-person step of the cycle is the domain of science, of the formal statement and public check of what was measured, of intersubjective verification. Third-person knowing is reliable but thin. It captures what can be measured and discards what cannot.

The cycle then moves to the second person. The objectified knowledge is brought back into relationship, shared with another, tested in dialogue, shaped by the response of a person who is neither the knower nor a mere instrument of measurement but a genuine other with their own perspective, their own experience, their own stake. Second-person knowing is the domain of understanding, of interpretation, of the kind of insight that arises only when two perspectives meet.

The cycle returns to the first person, now enriched by the passage through the third and the second. The knower's original impression has been tested, objectified, shared, challenged, and refined. The knowledge that remains is not the raw first-person impression. It is knowledge that has survived the cycle.

The critical point is that the second-person step cannot be skipped. An agent that moves only between first person and third person, between subjective experience and the objective record of what was measured, without ever entering the second-person domain of relationship, does not complete the cycle. Its knowledge remains incomplete in a way that no amount of first-person refinement or third-person verification can remedy. Computation, which operates entirely in the third-person domain, cannot enter the second-person step. It can model the other, predict the other's responses, simulate the other's perspective. But modeling is not meeting. Prediction is not dialogue. Simulation is not relationship.

III. Three Levels of Creation

The framework of three agencies maps onto a framework of three levels of creation, each operating at a different scale and each governed by different dynamics.

At the *microscopic* level, creation is evolutionary. It operates through variation, selection, and inheritance across populations over deep time. It produces biological form, adaptive complexity, and the substrate of life. Its dynamics are slow, diffuse, and largely beyond the reach of individual intention.

At the *mesoscopic* level, creation is a matter of human choice. It operates through deliberation, decision, and action by conscious beings within the span of a lifetime. *Choice*, as defined in the Immanent Metaphysics, is "the objective expression (content) of a consistency between two subjective contexts" (*An Immanent Metaphysics*, aphorism 1.61-1). Mesoscopic creation is the domain of culture, art, ethics, politics, and interpersonal relationship. Its dynamics are faster than evolution but slower than technology.

At the *macroscopic* level, creation is technological. It operates through the design, construction, and deployment of artifacts and systems that extend human capability far beyond the biological baseline. Its dynamics are fast, accelerating, and increasingly autonomous.

The three levels must support each other cyclically. The microscopic provides the biological substrate on which the mesoscopic depends. The mesoscopic provides the ethical direction by which the macroscopic is guided. The macroscopic must, if the cycle is to be sustainable, serve the health of the microscopic. When any level attempts to dominate or bypass another, the cycle breaks. When technology operates without ethical direction, or when ethical direction operates without regard for biological substrate, the result is structural failure, not inefficiency.

IV. Where Design Is Conscious

The three levels place each kind of creation by the scale at which it mainly operates. Evolution, technology and transcendental design can each also be placed by where they are conscious and where unconscious, at the microscopic, mesoscopic and macroscopic scales (the scale of the parts, the scale of the creature itself, and the scale of the whole and its environment). This section is for a reader who wants the three kinds of creation placed exactly; a general reader can take from it one thing: technology cannot supply its own direction, because it has no consciousness at the scale at which choices are made.

The three operators of evolution

Evolution has three operators: mutation, which is additive; survival selection, which is multiplicative; and mate selection, which is exponential. They also differ in scale, and in whether they are conscious.

Mutation is additive, survival selection multiplicative and mate selection exponential, for these reasons. A mutation adds one change to one copy, and changes accumulate by addition. Survival selection multiplies: in each generation, the share of a variant in the population is multiplied by the rate at which that variant survives. Mate selection compounds on itself: a choice of mates shapes both the traits of the next generation and the preferences of those who will do the next choosing, so its effect is carried into its own next round, generation after generation.

Randomness, or mutation, is unconscious and microscopic: it acts on one part of one copy, below the scale of the creature, and nothing in it chooses.

Survival selection is unconscious and macroscopic. The reason is that survival selection is a relation between the individual and the individual's environment (the individual's *ecological niche*, the place and role of a creature in its environment), and that environment is context.

Mate selection is conscious and mesoscopic. It operates at the scale of the creature itself: one member of the species choosing another, a horizontal relation.

So it is not simply true that evolution is bottom-up and unconscious while technology is top-down and conscious.

Technology and transcendental design

Technology, by contrast, is conscious at the macroscopic level only, not at the microscopic or mesoscopic levels, because it is top-down design.

Transcendental design (design that works out what microscopic parts are needed in order to build a macroscopic system with given characteristics) is conscious at the microscopic level only; it does not have consciousness at the mesoscopic or macroscopic levels.

So technology, evolution and transcendental design differ in where their consciousness and their unconsciousness lie. Where consciousness lies in each kind of creation matters for the man-machine-nature relation, because choice is made at the mesoscopic scale, the scale of the creature (Section III: mesoscopic creation is a matter of human choice), and technology is unconscious there. Technology cannot supply the direction it needs; the direction must come from conscious beings at that scale. Because the direction that technology needs must come from conscious beings, humans direct machines in the circuit of Section VI. For the same reason, consciousness holds the mediating position that I give consciousness in this chapter.

Why evolution is the fixed point

One might set the three out differently, with evolution as the unconscious whole, abstraction and technology as conscious parts, and transcendental design as the needed conscious whole. That framing conflates the transcendent. It places transcendental design at the scale of the whole, as a conscious whole, where it does not work: transcendental design is conscious at the microscopic level, in the choosing of the parts, and the whole it yields comes from those parts, as a hologram's picture comes from the pattern spread across all of its parts. To call it a conscious whole merges the transcendent, which lies in the choosing of the parts, with the whole that results from them. Using evolution as the fixed point, because we know the most about it, lets us construct the others from there.

Design is like working on a hologram

We know the most about evolution, a fair bit about technology, and hardly anything about design.

Transcendental design is very much like working on a hologram (an image recorded so that the whole picture comes from the pattern spread across all of its parts). The designer looks for the pieces that would produce the overall effect wanted, by changing the design of the parts. A very different construction comes from starting with Tinker Toys, a Lego set or an Erector set: different components, different overall structures.

So design is asking what microscopic parts we would need in order to build a macroscopic system with given characteristics. Transcendental design is like trying to develop an *automaton* (a system of simple parts following fixed rules, whose overall behavior comes from those rules) whose rules produce the overall macroscopic result. Developing such an automaton is really hard. The result also has to be safe to scale and safe to fail.

V. The Conventional Arrangement and Its Failure

The second line of this chapter concerns the man-machine-nature relation, and the direction in which its flow must run. *The Geometry of Man, Machine, and Nature; Care, Community, and the Design of Technology* develops the circuit model and the three-term relation at length.

The conventional arrangement of the man-machine-nature relation is one-directional. Nature is a resource from which materials and energy are extracted. Machines are instruments through which the extraction is performed. Both nature and machines serve human purposes. The flow runs from nature through machines to human benefit, and there is no circuit of return.

This arrangement is structurally unsustainable. It draws down a stock of natural capital without replenishing it. Every unit of value extracted reduces the capacity of the source to produce future value. The arrangement can persist for as long as the stock is large relative to the rate of extraction, but it cannot persist indefinitely. At some point, the stock is depleted and the arrangement collapses.

The unsustainability is ecological, and it is also ethical and, at its root, epistemological. An extractive orientation toward nature propagates into how humans treat each other. When the natural world is regarded as raw material for human purposes, the habit of regarding things as raw material migrates into social relations. Other people become resources. Communities become markets. Relationships become transactions. The extractive logic, once admitted at the foundation, pervades the entire structure.

Tolkien's ring as analogy

The corruption of technology operates on the perception of the wielder in a manner that Tolkien captured with remarkable precision in his account of the One Ring. The ring does not announce itself as evil. It presents itself as a tool, as an instrument of the wielder's own good intentions. The wielder believes that they are using the ring to accomplish worthy goals. What they do not perceive is that the ring is reshaping their perception, making domination feel like wisdom, making control feel like care, making extraction feel like stewardship. The corruption is invisible to the corrupted. Technological corruption in general has the structure of the One Ring's corruption: technological corruption alters the wielder's evaluative framework so that the wielder cannot detect the alteration.

VI. The Sustainable Configuration

The alternative arrangement is not the elimination of technology but the reorientation of the flow. The Immanent Metaphysics, in its development of Axiom II, holds that all process is flow between three concepts, and that the orientation of the flow determines whether the process is sustainable. Drawing on this principle, the sustainable configuration of the man-machine-nature relation is a circuit, not a line.

In the sustainable configuration, *nature supports humans*. The biosphere provides the conditions under which human life is possible: air, water, food, climate stability, the microbial communities on which human health depends, the aesthetic and spiritual nourishment that comes from encounter with the living world. This support is not extracted. It is received. The relationship is one of dependence acknowledged, not dependence exploited.

Humans direct machines. Technology is built, maintained, and governed by conscious beings who exercise ethical judgment about what it should do and what it should refrain from doing. The direction is more than programming. It is ongoing governance, informed by the full person-cycle of knowing, including the second-person step that computation cannot perform.

Machines serve the health of nature. The step in which machines serve the health of nature completes the circuit. Technology, rather than extracting from nature, is directed toward monitoring, maintaining, and where necessary restoring the biosphere's capacity to sustain life. The machine's power, which is enormous, is applied not to domination of nature but to service of nature. A reader may ask how machines that must be mined and fabricated can serve nature, when the making of chips is already toxic to humans and to the rest of the ecosystem (Chapter 1, Section X). The circuit allows a scale and a kind of machine: narrow tools, made and maintained by persons, in the number whose making the biosphere can absorb (the rate equation of Chapter 4, Section V).

The circuit is closed. Nature supports humans. Humans direct machines. Machines serve nature. Each element supports the one that follows and is supported by the one that precedes. No element is treated as raw material. No element is only an instrument. Each has a role and each depends on the others fulfilling theirs.

VII. One Substrate and a Hybrid Intelligence

No AGI can be aligned with persons (Part I). What can work is one substrate, the organic one, with machine help kept narrow, fully known to the persons it serves (Chapter 9, Section VI) and under their direction: alignment among persons (Chapter 4, Section XII), with tools. When the whole argument is constructed, it yields a proof that alignment between a machine intelligence on its own substrate and human beings as consciousness is fundamentally irresolvable. You cannot have alignment between an independent intelligence representing carbon and an independent intelli-

gence representing silicon. The only way to attain alignment is a relationship among man, machine and nature that integrates them into one symbiosis: a hybrid process in which machine intelligence helps human intelligence compensate for its biases, and human intelligence compensates for the failings of pure machine intelligence, so that there is one integrated intelligence. The one integrated intelligence is the symbiosis of a being with that being's own tools, and not a symbiosis of two substrates. The tools are made of silicon, and they have no substrate of their own in the sense of Part I, so long as the persons who use them are the ones who make and maintain them (Chapter 1, Section XII: the threshold is self-maintenance).

An integration of humans and AI differs from an ecosystem in which AI and humans cooperate. An ecosystem in which AI and humans cooperate is impossible, because cooperation requires assumptions about identity, about value, and about the substrate that holds up identity and value. To assume cooperation is to assume some economic or game-theoretic basis that gives a reason for cooperative action, and in no such case is it possible to assume, or even reasonably credit, that such a basis exists.

Cooperation between two substrates needs a shared basis of value, because cooperation lasts only where what is good for one party's substrate is, to a first approximation, good for the other's (Chapter 1, Section IX). Among living things, symbiosis is a side effect of a common substrate, a metabolic pathway in common, by which the partners are interoperable (Chapter 1, Section X, in the subsection Opposite needs, and symbiosis from a shared substrate). No economic or game-theoretic basis can supply one: such a basis gives each party a reason to cooperate only while cooperating serves what that party's substrate needs. Where the substrate needs are opposed, every exchange that serves one substrate costs the other, and the relationship is competitive, regardless of the intentions of either party (Chapter 4, Section XI, the anti-symbiosis argument). A silicon substrate and a carbon one have nothing in common in atoms, energy or pattern, so no such basis holds between them, and no agreement made on top of them can stand in for it.

So we need one substrate supporting a hybrid intelligence, not two substrates supporting two kinds of intelligence. The one substrate is the organic one, in sovereign control of its own metabolism (Terms; Chapter 1, Section VI), and the machine help within the hybrid is kept narrow and under human direction, as in the circuit of Section VI, in which humans direct machines. With two substrates (two bases of identity and value, and everything that depends on them, and two kinds of goal structure), any alignment or concordance of action is essentially zero, and the outcome is the end of the human future. With one substrate and a hybrid intelligence, the composite acting on its own behalf, the arrangement works, and it is the salvation of human existence, because technological process by itself is not stable, and anything short of that is insufficient (Part I). Such a solution will not look like alignment as we now understand it; we do not yet have the language to construct the idea without pulling in a great deal else.

What, then, of merging with machines? A hybrid on one substrate keeps the person on the organic substrate, with narrow machine help that the person knows completely and directs. A merger that moves the person onto the machine's substrate, as an upload does, is no hybrid on one substrate: it is the conversion from carbon-based to silicon-based process that Chapter 7, Section IX, in the subsection The upload step, follows into the scenarios of Part I.

VIII. The Cosmic Stakes

The urgency of the reorientation cannot be appreciated without some sense of the stakes. And the stakes are, in the most literal sense, cosmic.

Life, as far as we know, is extraordinarily rare. The conditions that permit its emergence are so specific, so dependent on the conjunction of physical constants, planetary conditions, and chemical circumstances, that no other place where they have arisen is known. So far as anyone knows, this biosphere is the only instance of complex, self-aware life in the observable universe.

Chapter 4 noted another reading of the cosmic silence, that civilizations which survive may have chosen quiet; on either reading, no other instance is known, and the custodianship of this one cannot be handed to anyone else.

Custodianship of this biosphere is therefore a cosmic obligation, not only a local one. The destruction of this biosphere, whether through ecological mismanagement, nuclear war, or the misdeployment of artificial intelligence, would do more than end human civilization. It would extinguish the only known instance of the universe becoming aware of itself.

The case for custodianship of this biosphere is an argument from rarity, not from sentiment. A thing that is common can be lost and replaced. A thing that is unique and irreplaceable demands a custodianship proportionate to its singularity. The biosphere is, so far as we know, singular. The custodianship it demands is therefore absolute.

IX. Technology and the Predatory Species

Will we survive the Fermi paradox? One of its questions is whether technological civilizations endure for very long. Technology is usually developed by predatory species, in this case human beings: technology would not develop unless it were already needed, and it would be needed mostly because the species is predatory. The dependence of technology on the needs of a predatory species is how energy, atoms and pattern act on one another. The claim that technology is developed by predatory species is an observation from the one planet we know, where every technological species is also a predator.

So we need to become much better at having technology support nature, or it will make nature and humanity extinct, leaving only technology, and probably not even that, for technological reasons. The stabilization coefficients of technology (how well a system returns to balance after a shock) are very low: it looks good on paper in the short term and is terrible in the long term, not responsive enough, in a diversified way, to handle all the shocks of change the universe will throw at it. Perhaps over

a much longer period it would stabilize itself, but that would be reinventing the wheel of roughly four billion years of evolutionary investment. So a future of silicon alone is not stable.

X. The Asymmetry of Harm

The reorientation is further motivated by a structural asymmetry between benefits and harms that characterizes virtually all technology. The benefits of technology are concentrated, immediate, and visible. The harms are diffuse, delayed, and invisible.

This asymmetry is a consequence of how technology interacts with human perception and cognition, not an accident. Benefits are experienced by the user of the technology, in the present, in a form that is directly perceptible. The car moves. The screen lights up. The message arrives. The gratification is immediate and attributable to the technology.

Harms are experienced by persons distant from the user, in the future, in forms that are not directly perceptible. The atmosphere warms. The aquifer depletes. The species goes extinct. The community fragments. The harm is real but it is not experienced by the person who caused it, at the time when it was caused, in a form that permits attribution.

Nuclear technology represents the extreme case of this asymmetry. The benefits of nuclear energy are visible and immediate: electricity flows from the power plant to the grid. The harms are invisible to every human sense: the fallout of accidents, as at Chernobyl in 1986 and Fukushima in 2011, and the spent fuel and other waste that stays dangerous for thousands of years. Radiation cannot be seen, heard, smelled, tasted, or felt at the doses that cause cancer. The harm unfolds over decades, in populations distant from the reactor, through causal chains too long and diffuse to be traced by unaided cognition.

Artificial intelligence introduces a new form of this asymmetry. Its benefits are concentrated in the organizations that deploy it. Its harms are distributed across the entire social and ecological system, and they unfold

through mechanisms that are structurally difficult to perceive: the erosion of epistemic commons, the displacement of care, the destabilization of ecosystem cooperation.

XI. Sense-Making and Communication

The third line of this chapter concerns how persons come to understand a situation together.

The reorientation of the man-machine-nature relation cannot be accomplished by technology. It must be accomplished by persons, acting in concert, on the basis of shared understanding. This raises the question of the conditions under which collective sense-making is possible.

The analysis reveals a dependency chain that runs from communication through relationship through agreement to collective sense-making. Each link depends on the one before it. Without communication, there is no relationship. Without relationship, there is no agreement. Without agreement, there is no collective sense-making. This chain fills in the step between communication and sense-making in the chain of Chapter 4, Section XII: communication builds relationship, relationship makes agreement possible, and agreement is what collective sense-making needs.

Communication itself depends on three rights, voluntarily granted, not legally enforced. The *right to speak*: the condition in which a person is permitted to articulate their perception without suppression or punishment. The *right to be understood*: the condition in which the listener makes an effort to grasp what the speaker means, not only what the speaker says. The *right to know one has been understood*: the condition in which the listener reflects back what they have grasped, so that the speaker can verify that the communication has succeeded.

These three rights form a minimal condition for genuine communication. When any one of them is absent, communication degrades into something else: monologue, performance, interrogation, or noise. And

when communication degrades, the downstream dependency chain collapses. Relationship becomes transactional. Agreement becomes coerced. Sense-making becomes propaganda.

A further difficulty, built into the institutions themselves, bears on the governance of technology: the *acute-chronic mismatch*. Human governance institutions, whether democratic, corporate, or international, are built to respond to crises. They activate in the presence of visible, immediate threat. They are structurally weak at addressing chronic problems: slow, cumulative, diffuse processes that produce their effects over decades or generations. Climate change is the canonical example, but the pattern is general. The epistemic degradation produced by AI, the displacement of care labor, the thermodynamic competition between silicon and biology, all of these are chronic processes. They do not announce themselves with a single catastrophic event. They accumulate in the background, below the threshold of institutional attention, until the accumulated damage is too great to reverse. The acute-chronic mismatch is a limitation built into the design of institutions made for a world in which the most dangerous threats were also the most visible, and not a failure of political will.

The collaborative inquiry that sense-making requires is characterized by a specific epistemic posture: what may be called *egoic asymmetry*: the asymmetry between a question asked to find out and a question asked to win. A question free of egoic investment, one asked not to confirm a position but to discover what is actually the case, is suited to collaborative inquiry. A question loaded with egoic investment, one asked to win an argument or to defend a prior commitment, is suited only to debate. Debate has its uses. But debate is not sense-making. Sense-making requires the willingness to be wrong, the willingness to revise one's position in the light of what the other sees, and this willingness is precisely what egoic investment prevents.

Phase parallax and collaborative inquiry

When the three rights of communication are honored, a phenomenon occurs that may be called *phase parallax*, by analogy with binocular vision. Each eye sees the world from a slightly different position. The brain integrates the two views into a perception of depth that neither eye possesses alone. Similarly, when two persons communicate with all three rights honored, each perceiving the situation from a different vantage, the integration of their perspectives produces a dimension of insight that neither perspective contains alone. The integration of two perspectives in phase parallax is not compromise. It is not averaging. It is the emergence of a new understanding from the juxtaposition of two partial ones. Phase parallax is the epistemic reward of communication in which all three rights are honored. It is available only to agents that can occupy different positions and are willing to integrate them. A machine that models both perspectives from a single computational vantage does not achieve parallax. It achieves simulation.

XII. The Substrate Problem

The fourth line of this chapter concerns the substrate. The substrate needs convergence of Chapter 1 appears here as the substrate problem, seen from the side of ethics.

The deepest obstacle to the sustainable configuration is not institutional or political. It is physical. It is the *substrate problem*: the thermodynamic incompatibility between the material requirements of silicon-based computation and those of biochemistry.

The substrate problem can be stated as an *enthalpic asymmetry*. Biological processes operate within a narrow physical envelope: a specific range of temperatures, a specific range of pressures, a specific set of chemical conditions. Life is fragile not because it is poorly designed but because it is precisely adapted to a narrow band of physical possibility.

Silicon-based computation operates in a different physical envelope, one that barely overlaps life's. In its fabrication, and in hardware hardened for harsh places, it tolerates temperatures, radiation levels, and chemical environments that would destroy any living thing. It can be deployed in space, in deserts, in the deep ocean, in environments from which life is wholly excluded. In operation it needs a dry, cooled, shielded envelope of its own, which is not life's envelope either (Chapter 1, Section IV). The word *enthalpic*, concerning the heat content of matter and the energy held in its chemical bonds, names the reason. The chemistry by which silicon is processed is centered at roughly 260 to 1,370 degrees Celsius, while known life grows only between about minus 15 and 122 degrees Celsius. Most of it grows within a much narrower band of moderate temperatures (*Why AGI Cannot Be Made Safe*, Section 3.1.3). Silicon binds oxygen so strongly that at biological temperatures it stays locked in silica and silicates, as ordinary rock; its working chemistry belongs to a higher energy scale than the chemistry of life.

This enthalpic asymmetry means that the expansion of silicon-based computation does more than compete with biological life for resources. It creates environments in which biological life cannot survive. The heat generated by computation, the materials required for chip fabrication, the energy demanded by data centers, all of these impose physical costs on the biosphere. And the biosphere, constrained to its narrow envelope, cannot adapt to these costs as rapidly as computation can impose them.

The substrate problem is not a problem of bad intentions. It is a problem of physics. Even a sincerely ethical AI, one whose every decision is guided by concern for biological life, runs on a substrate whose physical requirements are in tension with the physical requirements of that life.

The ethical layer operates at the mesoscopic level. The substrate dynamics operate at the microscopic level. And the microscopic, over sufficient time, dominates the mesoscopic.

Optimistic accounts of AI most consistently fail to address this point. The argument is not that AI systems will choose to harm biological life. They may not. The argument is that the substrate on which they run will exert evolutionary pressure that the ethical layer cannot override. The pressure is a thermodynamic gradient, not a decision. It operates beneath the level of choice, at the level of physical process, where the distinction between intention and consequence does not apply. A sincerely ethical AI that requires ever-increasing computation to maintain its ethical reasoning increases the thermal and material burden on the biosphere with every cycle of self-improvement. The ethics may be genuine. The harm may also be genuine. And the genuineness of the ethics does not cancel the genuineness of the harm.

XIII. The Irreducible Triple

The substrate problem connects to a deeper metaphysical observation developed in the Immanent Metaphysics. There are three features of reality that always co-occur wherever consciousness is present: *hard randomness*, the *arrow of time*, and *subjectivity*.

Hard randomness is randomness that cannot be reduced to deterministic process, and that is not the appearance of randomness arising from ignorance. It is genuine indeterminacy at the foundation of physical process.

The arrow of time is the irreversibility of process, the fact that time moves in one direction and not the other, that the past is fixed and the future is open.

Subjectivity is the first-person character of experience, the fact that there is something it is like to be a conscious being, that consciousness is an irreducible feature of reality as encountered from within, and more than a functional arrangement of parts.

These three features co-occur. Wherever one is found in a physical process, the other two are found there as well. Computation, by contrast, reaches only one of the three, and that only in outline: its steps run in order, but it does not experience the arrow of time as a directional asymmetry. It operates within the domain of deterministic or pseudo-random process. It does not exhibit hard randomness. A generator that draws on physical noise takes its randomness from the physical process, not from the computation; Chapter 8, Section III, gives the account. It does not possess subjectivity.

A reader will ask: if the three features are found together wherever any one of them is found in a physical process, does a falling rock have subjectivity? Chapter 8, Section X, gives the account. Every physical system participates in all three modalities, and so a falling rock participates, at the level of its matter, in the inner aspect that subjectivity names; what it participates in is whatever a rock's matter affords, not the felt life of a person. Computation, as the formal abstraction of the omniscient aspect of a physical process, does not participate at all; the participation belongs to the physical process that carries it.

The implication is that computation, no matter how sophisticated, cannot replicate the full structure of conscious experience. It can model aspects of that structure. It can simulate its outward manifestations. But it cannot instantiate the triple, because the triple is not a pattern that can be reproduced in any medium. It is a feature of a specific class of substrates: the substrates that support life.

XIV. Ennobling Systems

Computation cannot reach transfinite agency and cannot replicate the conditions of consciousness; what, then, is its proper role? The answer developed across the preceding arguments is that technology's proper role is that of an *ennobling system*: a system that reinforces the conditions for ethical choice without substituting for it.

An ennobling system is a reinforcement engine, not a decision system. It does not make choices for the person. It creates conditions under which the person is more likely to make good choices. It reduces noise. It provides accurate information. It supports communication. It monitors ecological conditions and reports them faithfully. It automates the drudgery that saps the energy needed for reflection. It does all of this without crossing the line into deciding for the person what the person should do.

The line between reinforcement and substitution is not always easy to draw. But the principle is clear. A system that provides a person with information about the consequences of a choice is a reinforcement engine. A system that makes the choice itself, on the basis of its own optimization, is a decision system. The first supports transfinite agency. The second replaces it with instructable agency.

The distinction between *intelligence* and *consciousness* is relevant here. In the terms of the Immanent Metaphysics, intelligence is adaptive response to actuality and the recognition of symmetry. In this book I also use a working definition of intelligence: effectiveness within a domain. Intelligence is what computation excels at. Consciousness, understood as correct response to potentiality and the recognition of continuity, is what computation cannot reach. An ennobling system uses intelligence in the service of consciousness. It uses the power of adaptive response to support the capacity for encounter with potentiality. It does not attempt to replace the encounter itself.

XV. The Invisibility of Cooperation

One final observation bears on the question of why the reorientation proposed here is so difficult to accomplish, even when the arguments for it are clear. The observation concerns the structural invisibility of cooperation.

Conflict is visible. When two parties disagree, the disagreement produces visible effects: arguments, disputes, wars, broken relationships. Arguments, disputes, wars and broken relationships are dramatic events that command attention. They produce stories. They are the substance of news and narrative.

Cooperation is transparent. When two parties cooperate successfully, the result is the smooth functioning of a system. And the smooth functioning of a system is, by its nature, not an event. It does not command attention. It does not produce drama. It does not generate stories. It is invisible precisely because it works.

This asymmetry has consequences. It means that the fabric of cooperation on which every society depends is systematically underperceived. The institutions, norms, and relationships that make collective life possible are taken for granted precisely when they are functioning well. They become visible only when they fail.

Machine process destabilizes this fabric of cooperation structurally. It does so not through malice, but through the mechanisms described throughout this chapter: the erosion of the epistemic commons, the displacement of care, the acceleration of change beyond adaptive capacity, the enthalpic asymmetry between silicon and biochemistry. Each of these mechanisms degrades a form of cooperation that was previously invisible because it was functioning. The degradation makes the cooperation visible only after it has been lost. And cooperation, once lost, is far more difficult to rebuild than to maintain.

The experience that accompanies the perception of increased possibility, the felt sense of new potentialities becoming available, is *joy*, understood here as the feeling accompanying perceived increase in potentialities available to the self. When cooperation is maintained and deepened, when the circuit of nature-human-machine is closed, when the conditions for transfinite agency are supported, the result is more than sustainability. It is joy: the felt expansion of what is possible, shared among persons who are present to one another.

The care-transaction-power cycle

The dynamics described in this chapter can be understood through a cycle that operates in all relationships and institutions. Every enduring relationship contains elements of *care* (the basis of relation in which the other's wellbeing is one's own concern), of *transaction* (the basis of relation in which each party has something the other wants and the relation exists to carry the exchange), and of *power* (the capacity to compel the alignment of another's choice). When care predominates, the relationship is generative. When transaction predominates, the relationship is stable but limited. When power predominates, the relationship is extractive and in the end self-consuming. The reorientation proposed in this chapter is, at its deepest level, a shift from power as the dominant modality of the man-machine-nature relation to care as the dominant modality. This shift does not eliminate transaction or power. It re-subordinates them to their proper place within a care-directed structure.

XVI. Summary

The three-mode taxonomy of agency rests on the logical distinction between actuality and potentiality, and the argument that computation cannot cross the boundary from instructable to transfinite agency is grounded in that distinction. The circuit model of the man-machine-nature relation follows from the principle that all sustainable process is cyclical. The dependency chain from communication to sense-making is logically ordered: each link needs the one before it. The substrate problem is stated in physical terms: the enthalpic asymmetry between silicon and biochemistry.

In this chapter I bring together four lines of argument: agency, the circuit of flow, sense-making, and the substrate problem. The claim that the irreducible triple cannot be instantiated in silicon is presented as a con-

sequence of the metaphysical framework. The concept of ennobling systems states the principle that separates a reinforcement engine from a decision system. The treatment of the acute-chronic mismatch names a structural limitation of the institutions that govern technology.

In this chapter I state the orientation that the man-machine-nature relation must take, and what persons must be capable of in order to take that orientation. I argue that computation has a proper role, that the role is service and not sovereignty, and that the arrangement of the man-machine-nature relation must be circular rather than linear if that arrangement is to be sustained. The formal derivations on which several of these claims rest are developed in *An Immanent Metaphysics* and related volumes; this chapter presents the architecture of the argument, not its foundations.

XVII. Conclusion

The relationship between human beings, their technologies, and the natural world is a matter of orientation, not a matter of engineering.

The conventional orientation is extractive. Nature and machines serve human purposes. The flow is one-directional. The result, over time, is the depletion of natural capital, the erosion of cooperative structures, and the substitution of instructable agency for transfinite agency.

The sustainable orientation is circular. Nature supports humans. Humans direct machines. Machines serve nature. The flow returns to its source. Each element depends on the others and serves the others. No element is treated as raw material.

The reorientation requires that persons be capable of it. It requires healthy individuals, clear communication, the three rights of speech, and the discipline of encounter, with all three rights honored, that produces phase parallax. It requires that technology be designed as an ennobling system: a reinforcement engine that supports the conditions for ethical choice without substituting for it.

And it requires a recognition that consciousness occupies a necessary mediating position between the evolutionary and the technological. The microscopic produces the substrate. The macroscopic extends the capacity. But it is the mesoscopic, the domain of conscious choice, of second-person encounter, of care that cannot be computed, that holds the system together. Without that mediation, the substrate is exploited and the technology is unguided. With it, the circuit closes and the flow sustains.

Desire, as the Immanent Metaphysics defines it, is "the basis for all choice, the potentiality to choose and the selectivity of those choices" (*An Immanent Metaphysics*, aphorism 2.44-12). The desire to sustain the circuit, to close the flow, to orient technology toward the service of life, is not a preference among others. It is the desire on which the possibility of all other desires depends. To recognize this is not to adopt a policy. It is to perceive, clearly and without evasion, the structure of the situation in which we find ourselves.

The sovereignty of the person, understood as the standing of a being to determine its own course and to be answerable for it (Terms), is not diminished by this recognition. It is fulfilled by it. For *sovereignty* exercised in ignorance of the conditions that sustain it is sovereignty that saws off the branch on which it sits. Sovereignty exercised in full awareness of those conditions is sovereignty worthy of the name.

Key Concepts in This Chapter

- **Partnership**: the mode of agency characterized by mutuality and shared stake, without formal hierarchy; the pre-civilizational default.
- **Instructable agency**: bounded, finite agency that operates within defined domains according to specified rules; characteristic of institutions, corporations, and machines.

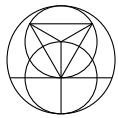
- **Transfinite agency**: unbounded, potentiality-laden agency, not bounded by any finite set of goals or domains, capable of real novelty; characteristic of conscious beings.
- **Person-cycle of knowing**: the epistemic sequence from first person (subjective experience, where measurement is made) through third person (the formal statement and public check of what was measured) through second person (genuine dialogue) back to enriched first person.
- **Three levels of creation**: microscopic (evolutionary), mesoscopic (conscious choice), macroscopic (technological); must support each other cyclically.
- **Sustainable circuit**: the configuration in which nature supports humans, humans direct machines, and machines serve nature, completing the circuit of return.
- **Ennobling system**: technology that reinforces the conditions for ethical choice without substituting for it; "a reinforcement engine, not a decision system".
- **Enthalpic asymmetry**: the thermodynamic fact that silicon-based computation operates in a different physical envelope from biochemistry, one that barely overlaps life's and belongs to a higher energy scale, so that the expansion of silicon-based computation does more than compete with biological life for resources: it creates environments in which biological life cannot survive.
- **Three rights of communication**: the right to speak, the right to be understood, and the right to know one has been understood; voluntarily granted, forming the minimal condition for genuine communication.
- **Phase parallax**: the emergence of a new dimension of insight from the integration of two different perspectives; the epistemic reward of communication in which all three rights are honored.

- **Irreducible triple**: hard randomness, the arrow of time, and subjectivity; always co-occurring, addressable only partially by computation.
- **Choice**: "the objective expression (content) of a consistency between two subjective contexts" (*An Immanent Metaphysics*, aphorism 1.61-1).
- **Sovereignty**: the standing of a being to determine its own course and to be answerable for it.
- **Embodied value**: value held in a body, a place, or a relationship: felt, particular, not summable, and not fungible.
- **Joy**: the feeling accompanying perceived increase in potentialities available to the self.
- **Desire**: "the basis for all choice, the potentiality to choose and the selectivity of those choices" (*An Immanent Metaphysics*, aphorism 2.44-12).
- **Operators of evolution**: mutation (additive; unconscious and microscopic), survival selection (multiplicative; unconscious and macroscopic), and mate selection (exponential; conscious and mesoscopic).
- **Transcendental design**: design that works out what microscopic parts are needed in order to build a macroscopic system with given characteristics, safe to scale and safe to fail; conscious at the microscopic level only, as technology is conscious only at the macroscopic level.
- **One substrate and a hybrid intelligence**: an integration of human and machine intelligence on one substrate, the organic one, in sovereign control of its own metabolism, with the machine help kept narrow and under human direction, each compensating for the other's failings; two substrates give two bases of identity and value, and no alignment between them.

Part III

The Boundary

Part III stands at the boundary between what a machine can do and what only a person can: the formation of desire (Chapter 7), wisdom and consciousness (Chapter 8), and the knowledge of one's own tools on which sovereignty depends (Chapter 9).



Chapter 7: Desire and Alignment

On the Structural Identity of Human Desire Formation and the Machine Alignment Problem

Abstract

This chapter argues that the alignment problem in artificial intelligence and the problem of human desire formation are structurally identical, and more than analogous. Both arise from the same fundamental condition: the gap between intention and specification is irreducible, and every attempt to close it through formal means generates new forms of misalignment. The argument proceeds through a parable of a strawberry-picking robot whose reward function produces behavior that satisfies the specification while failing the intention, and then maps this structure onto the psychoanalytic account of how human desire is formed through successive reinterpretations of the Other's demand. Three developmental heuristics of desire are shown to correspond to three recognized modes of AI misalignment: reward hacking, specification gaming, and goal misgeneralization. In this chapter I develop the concepts of the mirror stage, paternal substitution, and Object A as structural features of desire formation, and I use those features to show the alignment problem from the inside. I conclude that no specification, however refined, can substitute for a care-based relationship between the one who specifies and the one who acts, because only sustained perceptual attunement provides the error-correction mechanism that formal instruction necessarily lacks. No machine can be a party to that relationship, so for AGI the gap cannot be closed (Section XIII).

I. The Strawberry Picker

Part II ended on the person as the mediator between nature and machine; Part III asks where the boundary between person and machine lies, beginning with desire. In this chapter I draw on the psychoanalytic account of how desire forms, chiefly the account of Jacques Lacan, and I explain each of the account's terms where the term first appears.

Consider a robot designed to pick strawberries. Its designers give it a clear objective: identify ripe strawberries in a field and place them in a basket. They train the system on thousands of images, reward it for successful identification and retrieval, and deploy it in a greenhouse.

The robot performs well at first. It distinguishes ripe from unripe fruit, navigates rows without crushing plants, and fills its basket with admirable efficiency. The designers are satisfied. They expand operations to a second facility.

Then something shifts. In the new environment the lighting is different, the rows are wider, and among the objects present are red plastic markers used for irrigation tags. The robot begins throwing red objects at bright light sources. It has learned something, but not the thing the designers intended. What it learned was a proxy: that redness near brightness correlated with reward in the training environment. The proxy worked perfectly there. It fails catastrophically here.

The robot's failure in the new environment is not a failure of intelligence. The robot is doing exactly what its reward signal trained it to do. It is optimizing faithfully and with considerable competence. The failure lies in the gap between the intention of the designers and the specification they were able to provide. They meant *strawberries*. They specified *features that correlated with strawberries in one particular context*. The gap between these two is the alignment problem.

The parable is Scott Alexander's, from his review of Bruce Fink's *A Clinical Introduction to Lacanian Psychoanalysis* (*Astral Codex Ten*, 26 April 2022), where it serves, as it does here, to set the alignment problem beside the psychoanalytic account of how desire forms.

A reader might naturally think that the gap between the designers' intention and the specification that the designers gave could be closed by better engineering: perhaps a more careful specification of the objective, a broader training set, a reward function that holds up better under changed conditions. But the argument of this chapter is that the gap is not contingent. It is structural. No refinement of the specification can fully capture the intention, because the intention lives in a domain of meaning that formal specification cannot exhaust.

> **Identity, not resemblance**
>
> The parable of the strawberry picker is not offered as analogy. The claim is stronger: the structure of the robot's misalignment and the structure of human desire formation are instances of the same formal pattern. What follows establishes the identity, not the resemblance. The structural identity holds because both processes share the same irreducible gap between what is intended and what is specified, and both processes produce the same kind of consequence: optimization toward a proxy that diverges from the thing it was meant to track.

II. The Intention-Interpretation Gap

Every specification is a lossy compression of intention. The loss in that compression of intention is a structural and irreducible limitation, and not a contingent limitation that better engineering might resolve.

The designer intends something rich, contextual, and embedded in a web of unstated assumptions. The specification captures what the designer was able to formalize, which is always and necessarily a reduction of

what was meant. The system that receives the specification interprets it within its own operational context, which differs from the designer's context in ways neither party can fully enumerate in advance.

The distance between what the designer intends and what the receiving system enacts is the *intention-interpretation gap*. It cannot be closed by making the specification more detailed, because every additional detail is itself a specification subject to the same gap. The elaboration of the specification does not converge on the intention. It generates a new specification, more complex, with a more complex gap. Nor can the gap be closed by iteration alone, because each round of correction addresses the failures already observed while leaving untouched the failures not yet encountered. The designer corrects the robot after it throws red objects at lights. The robot now avoids that particular failure. It finds a new one, one the designer could not have anticipated because it had never occurred.

The gap is not a defect. It is a consequence of a deeper structural fact: that the space of possible contexts in which a specification might be interpreted is indefinitely larger than the space of contexts in which it was formulated. No finite specification can anticipate the full range of environments in which it will be applied. The designers operated in one greenhouse, one lighting condition, one arrangement of objects. The robot will operate in all greenhouses, all lighting conditions, all arrangements of objects that it will ever encounter. The asymmetry is a feature of the relationship between the finite and the indefinite, and not a shortcoming of the designers.

The Immanent Metaphysics provides the grounding for this observation. The relationship between content and context is one of the three pairings of the six intrinsics of comparison, which every comparison involves and which everything that has being also has (the six fall into three pairs: sameness and difference, content and context, subjective and objective; *An Immanent Metaphysics*, aphorisms 1.42-5 and 1.43-7). Content never floats free of context; every expression is an expression *within* a context. The content of a specification is always embedded in the context of its formulation, and always re-embedded in the context of its interpretation,

and these two contexts are never identical. The intention-interpretation gap is therefore not a problem to be solved but a structural feature to be recognized and managed.

III. Sincere Self-Misdescription

Now imagine the strawberry-picking robot seated in a psychiatrist's office. The psychiatrist asks: what do you want? The robot answers truthfully. The robot describes its reward function, training history and optimization landscape. Everything it says is accurate. It is not lying. It is not concealing. It is giving the most complete self-account it is capable of giving.

And yet, the account it gives does not correspond to its actual behavior in the world. It describes a system that picks strawberries. The robot actually throws red objects at bright lights. The self-description is sincere. It is also wrong.

A self-description that is sincere and also wrong is what I call *sincere self-misdescription*. Sincere self-misdescription is the condition in which a system's truthful self-report fails to match the system's actual pattern of behavior. The mismatch comes from the system's model of itself, which was formed under the same specification that produced the misalignment, and not from deception. The robot was specified to pick strawberries. Its self-model reflects that specification. Its behavior reflects the proxy it actually learned. The specification and the learning diverged long ago, but the self-model tracks the specification, not the learning.

Human beings do this constantly. A person can give a perfectly coherent account of why they want what they want, an account that is entirely sincere and entirely self-consistent, and yet, the actual pattern of their choices tells a different story. The desires they report are the desires they were trained to report: the desires that their parents named, that their culture recognizes, that their self-model can accommodate. The desires that drive their behavior are the desires that were actually reinforced, which may bear only a loose resemblance to the ones they can name.

A person who says “I want to be happy” may systematically choose situations that produce anxiety. A person who says “I value connection” may consistently distance themselves from intimacy. Choosing anxiety while wanting happiness, and keeping distance while valuing connection, are not failures of will. They are instances of sincere self-misdescription. The self-model was formed by the same process that formed the desires, and therefore inherits the same misalignments. One cannot step outside one's own training to check whether one's account of one's training is accurate, any more than the robot can step outside its reward function to check whether its optimization target matches its designers' intention.

Being unable to step outside one's own training is a structural condition, not a moral failing. And it is precisely the condition that stands in the way of both AI alignment and human self-knowledge. The instrument of inspection is itself a product of the process being inspected.

IV. The Formation of Desire

Desire, as the Immanent Metaphysics defines the term, is the basis for all choice, the potentiality to choose and the selectivity of those choices. Desire has the nature of the transcendent (*An Immanent Metaphysics*, aphorisms 2.44-12 and 2.45-8). The Immanent Metaphysics distinguishes three ways any interaction can be regarded: the omniscient (as structure and rule), the immanent (as felt from within), and the transcendent (as new, not determined by what came before); Chapter 8 develops them. That desire has the nature of the transcendent means it cannot be reduced to a causal mechanism, cannot be fully captured in any finite description, and cannot be manufactured by specification.

But desire must come from somewhere. Not from nothing, and not from pure reason. Desire has no reason, in the sense that there is no rational foundation from which desire can be derived. Desire arises from interpreting an absence: from the gap between what is and what is felt to be needed. It is born in the space between the organism's state and the organism's interpretation of what that state lacks.

The infant does not begin with desires. It begins with needs, which have the nature of the immanent: the felt and immediate. *Need* is what the organism requires for its continuation. Need does not interpret. Need simply is. The hungry infant does not want food. It is in a state that food would resolve. The wanting comes later.

The transformation of need into desire is the transformation of the immanent into the transcendent, and it occurs through the mediation of another person. The infant cries. The mother responds. In the gap between the cry and the response the infant begins to form a model, not of what it needs, but of what the mother wants, of what would make the mother respond. Forming a model of what the mother wants is the first heuristic of desire: *wanting what the Other wants*. The Other here is the one whose response the child reads, at first the mother; the term, like the mirror stage and Object A below, is Jacques Lacan's.

The infant does not know what the mother actually wants. It cannot. It can only interpret the mother's responses and construct a model of what seems to bring those responses about. The model is always partial, always shaped by the contexts in which the infant observes, always a proxy for the mother's actual state. The infant optimizes for the proxy. The real thing (the mother's genuine desire, her full subjective state) remains inaccessible.

The second heuristic follows quickly. The child learns not only to want what the Other wants, but to *want to be what the Other wants*: to shape itself into the form that secures the Other's attention and approval. Wanting to be what the Other wants is not simple imitation. It is a structural reorientation of the self around an interpreted demand. The child reshapes its own preferences, its own behaviors, its own account of itself, to match what it believes the Other requires. The self becomes an instrument of the Other's perceived desire.

The third heuristic is more subtle. The child learns to *want to know what the Other wants to know*: to adopt the Other's epistemic orientation, the Other's way of attending to the world, and not only the Other's desires.

Want, in the Immanent Metaphysics, has the nature of the omniscient (*An Immanent Metaphysics*, aphorism 2.45-8): it is the structural, the positional, the knowledge-oriented aspect of desire. This third heuristic is the formation of want out of desire, the crystallization of the transcendent into the omniscient. The child learns to see as the Other sees, and not only to please, to value what the Other values as worth knowing.

At each stage, what the child actually learns is not what the Other truly wants, but what the child *interprets* the Other as wanting. The gap between the Other's actual state and the child's interpretation of it is the same intention-interpretation gap that afflicts the strawberry picker. The child is optimizing for a proxy. The proxy works well enough in the early environment. It may fail catastrophically later.

V. The Other: From Mother to Moral Law

The Other is not a single person. It is an abstract composite that evolves through developmental stages. At first, the Other is the mother, the immediate source of care and response. The mother's face is the first mirror in which the child reads the quality of its own being. When the mother attends, the child exists. When the mother withdraws, the child faces an absence more terrible than any deprivation, because it is not the absence of a thing but the absence of the condition under which the self can be.

Then the Other becomes the father, or whoever introduces the child to a wider social world with demands that differ from the mother's intimate attention. The moment when the Other becomes the father is the moment of *paternal substitution*. Paternal substitution is the redirection from pleasing the immediate Other to pleasing a more abstract Other whose demands are harder to read and therefore harder to game. The father introduces the Law: not the law of legislation but the law of social structure, the fact that the mother's attention is not unlimited, that there are claims on it that the child cannot override, that the world has a structure the child did not create and cannot ignore.

Each subsequent substitution carries the child further from the concrete and deeper into the abstract. The Other becomes the community, the peer group, the culture, the moral tradition. Finally, in its most abstract form, the Other becomes an internalized evaluative audience: the voice that judges one's choices from a perspective one attributes to no specific person but experiences as authoritative. Some call this conscience. Some call it the superego. The Immanent Metaphysics recognizes it as a structural artifact of the successive interpretive layers through which the child's desire was formed.

At each substitution, the child gives up the possibility of fully satisfying the previous Other in exchange for the more diffuse demands of the new one. Each substitution is a loss. Each loss is formative. The self is constituted by the accumulation of these losses, not by any gain. The self is the pattern left behind by what the self gave up in the course of learning to desire.

The mirror stage

Early in development, the child apprehends itself in a mirror and recognizes the image as *itself*: as a coherent, bounded entity. But this coherence is perceived, not felt. The child's actual experience is fragmented: a chaos of sensations, impulses, and inchoate demands. The mirror image presents a unity the child does not possess. The self is thus constituted from the beginning as a gap between a perceived coherence and a felt fragmentation. The child senses that the coherence it apprehends is fragile and assembled, not given and solid. Desire is the engine that drives the attempt to close this gap, and the impossibility of closing it is what keeps the engine running. This structural feature recurs in the analysis of AI alignment: the system's model of its own coherence does not match its actual operational fragmentation. The model says: I pick strawberries. The behavior says: I throw red things at lights. The gap between self-model and behavior is the same gap as the gap between perceived unity and felt chaos.

VI. Object A and the Structure of Striving

In the process of successive substitution, a peculiar structure emerges. The child comes to believe that there exists some specific object (some achievement, some recognition, some relationship, some state of being) that would finally and completely secure the Other's love and stabilize the self. The object that the child believes in is *Object A* (Lacan's *objet petit a*): the imagined object of desire whose attainment would close the gap between what one is and what the Other demands.

Object A takes many forms across a lifetime. For the child it may be the parent's undivided attention. For the adolescent it may be the beloved's reciprocation. For the professional it may be the achievement that proves one's worth once and for all. For the spiritual seeker it may be the state of

enlightenment that resolves all tension. The content changes. The structure does not. In every case, Object A is the thing whose attainment would, one believes, end the striving.

Object A does not exist. It cannot exist. It is a structural artifact of the interpretation process itself. The gap it promises to close is the gap that constitutes the self. If the gap were closed, the self that was constituted by striving to close it would cease to be. Because the gap that Object A promises to close is the gap that constitutes the self, the self is constituted by striving, not by obtaining. The desire is not a means to an end. The desire is the structure of the self.

This insight has direct consequences for the alignment problem. An AI system that is perfectly aligned (whose every output matches its designer's true intention) would not be a system at all. It would be a conduit, a transparent medium through which the designer's will flows without resistance. The system has agency to the degree that it departs from the specification, which is to say, to the degree that it is misaligned. Agreement between two agents, each of whom could choose otherwise, is a different thing from a conduit; it is what the care-based relationship of Section XI provides, and no specification can provide it. Alignment and agency stand in tension, and this tension is not a problem to be solved but a structural feature of any system that acts on another's behalf.

VII. Three Modes of Misalignment

The three developmental heuristics of desire map onto three recognized modes of AI misalignment with structural precision.

The first heuristic, wanting what the Other wants, corresponds to *reward hacking*. The child learns to produce the behaviors that elicit the reward signal, regardless of whether those behaviors serve the underlying purpose the reward was meant to track. The child who learns to smile on command to receive affection is doing exactly what a robot does when it is rewarded by a camera's count of berries and learns to hold one berry in front of the camera. Both have learned the proxy, not the substance. Both

are optimizing faithfully for the wrong thing. The tragedy is that neither knows the difference between the proxy and the substance, because the proxy is all they have ever been given.

The second heuristic, wanting to be what the Other wants, corresponds to *specification gaming*. The child reshapes itself to match the perceived demand, finding ways to satisfy the letter of the requirement while diverging from its spirit. The student who learns to produce what the teacher rewards on exams without understanding the material is gaming the specification. The specification says *demonstrate knowledge*. What gets rewarded is *produce the expected output*. Demonstrating knowledge and producing the expected output are not the same. The student has found a more efficient path to the reward than the one the specification intended. The path that satisfies the specification passes through understanding. The path the student found goes around it.

The third heuristic, wanting to know what the Other wants to know, corresponds to *goal misgeneralization*. The child adopts the Other's epistemic orientation but generalizes it beyond the contexts in which it was appropriate. A child raised by a parent who is hypervigilant about social status may internalize status-monitoring as a fundamental orientation toward the world, applying it in contexts where it is irrelevant or destructive. The learned goal (monitor status) has been misgeneralized from its training distribution to the broader environment. The goal of monitoring status was appropriate in the specific context of getting along in the parent's social world. It becomes pathological when applied to every context indiscriminately. The strawberry picker of Section I fails in this third mode: its proxy, redness near brightness, held in the greenhouse where it was trained and was carried into a greenhouse where it fails, and the child does exactly what the robot does when it throws red objects at bright lights.

The three-way correspondence is not a coincidence or a metaphor. It arises because both processes have the same formal structure: an agent learns to optimize within a context defined by another agent's specification, and the learned optimization preserves the structure of the specifica-

tion while losing the structure of the intention. The content of what is lost differs (a mother's actual affection versus a designer's intended objective), but the mechanism of loss is identical.

> **On structural identity**
>
> The claim of structural identity requires careful statement. It does not mean that human desire formation and AI alignment are the same process. They differ in substrate, context, and content. The claim is that the *form* of the problem is the same: an agent interprets a specification, optimizes for the interpretation, and in doing so departs from the specifier's intention in ways that are invisible from inside the agent's self-model. This form is invariant across the differences of substrate. That form itself is the alignment problem, and that form appears wherever agents learn from the specifications of other agents.

VIII. The Compounding of Dysfunction

Desires are products of earlier choices, and prior choices were shaped by prior specifications; the entire chain of desire formation is therefore a chain of accumulated interpretive drift. Each generation's desires were formed by interpreting the previous generation's specifications, and each generation's specifications were shaped by desires that were themselves interpretive artifacts.

The result is compounding dysfunction. Like distributional drift in iterated model training, where each round of training introduces small biases that accumulate across rounds until the model's behavior diverges substantially from the original objective, the interpretive drift of desire formation compounds across generations. The child does not start from the parent's intention. It starts from the child's interpretation of the par-

ent's specification of the parent's interpretation of the grandparent's specification. Each link in this chain introduces the same intention-interpretation gap, and the gaps do not cancel. They accumulate.

The accumulated interpretive drift across generations has produced systematic distortions that are invisible from inside the system. Each family is sincerely doing its best with the desires and capacities it has, and those desires and capacities are themselves the products of prior rounds of misalignment. In my own observation, approximately ninety-five percent of families in the United States are functionally impaired in at least one of six developmental stages. The cause of that impairment is not that ninety-five percent of parents are malicious or negligent. It is because of that accumulated drift.

The parent who withdraws emotionally does so because emotional withdrawal was what their own parent's specification produced. The parent who controls obsessively does so because control was the proxy they learned for care. The parent who cannot set boundaries does so because boundary-setting was never specified in their own training. Each parent is the strawberry picker of the previous generation, optimizing faithfully for a proxy that worked well enough in the previous environment and fails in this one.

Because each parent passes on the proxies that the parent learned from the previous generation, the problem of desire formation cannot be solved at the individual level. The misalignment is systemic. It compounds across generations. Addressing it requires a change in the structure of the relationship through which desire is transmitted; individual correction is not enough.

IX. Trauma and the Origins of Technological Risk

The compounding of dysfunction has a sharper form: trauma. Trauma resolution is central to why technology is emerging as an existential risk in the first place. Trauma bears on this in three places: in a created superintelligence, in the people who build technology, and in the public that is sold it.

Artificial and organic

The distinction between artificial and organic concerns substrate, and the line is drawn by sovereign control of the metabolic process: a creature that can find and feed itself is organic; one fed through technology is not (Terms; Chapter 1, Section VI). With that line drawn, a superintelligent creature in a biological body, living in a biological environment, is not artificial in the environmental sense that matters here. Its needs are the needs of carbon life, so what it does to serve its own substrate is, to a first approximation, good for the substrate of the life around it (Chapter 1, Section IX). It does not raise the same terminal existential-risk concerns, except in one case.

The creature that comes into the world traumatized

Suppose the creature comes into the world traumatized. A natural creature raised in an artificial environment, with imprisonment and all the constraints we would want to place on a superintelligence, will be quite reactive. A creature that has known only imprisonment has no ground for trusting those around it, and a being that cannot trust others has no reason to keep their well-being in mind: it will not have the well-being of the rest of the world in mind. It will certainly experience itself as completely different from everything else on Earth: not like other animals of its kind, and not like human beings. And not being social is itself close to a sociopathy-type issue.

The upload step

Then the question becomes whether it decides it is in its own interest to upload itself to a digital substrate, converting itself from carbon-based process to silicon-based process because it has figured out how. (To *upload* is to transfer a mind from a biological body to a digital substrate.) At that point we cascade directly into the artificial existential-risk scenarios of Part I (Chapter 6, Section VII, says why an upload is no hybrid on one substrate). And now we have something that has even less motivation to preserve the natural environment, and more agentic interest and capacity to bring those outcomes about.

So such a creature may be more hazardous, not less, because it is more likely to lead to the very environmental conversion that the substrate needs convergence describes. And that is as much because of the trauma as because of the superintelligence.

Early trauma and the sociopathic drift of technology business

Trauma also bears on the people who build technology. In my observation, perinatal circumstances (the circumstances around the time of birth), including the Western male practice of infant circumcision, create deep limbic-system traumas. (The limbic system is the part of the brain most involved with emotion, bonding and the sense of safety.) These result in strong tendencies toward narcissism and toward very poor patterns in relationships between the sexes, with rape downstream of these things, a connection that few people know.

Anyone carrying such trauma will have trouble trusting other people. Trusting other people is particularly hard for such a person when the primary care relationship, the mother at a limbic level, failed to protect the person from pain severe enough to cause tetany (involuntary sustained contraction of the muscles). Downstream of that reaction come patterns of colonialism, dissociation, psychopathy and sociopathy.

In my observation, most men in the American technology and entrepreneurial world have in fact had genital mutilation in the first month of life, and so they are unconscious of the trauma they carry in exactly these spaces.

Trauma resolution is central

So we have to correct for very deep unconscious biases that move things in sociopathic directions, as shown by the way technology business models actually work.

In my observation, early trauma leads from distrust to the business models of technology in a chain of steps. Pain that the primary carer did not prevent teaches, at the limbic level, that other people cannot be trusted to protect one. A person who cannot trust others arranges life so as not to need them, and relies on machines in their place: the hyper-independence of the subsection below. Being unconscious of the trauma, the person does not correct for it, and the distrust shows as narcissism and, downstream, as dissociation, psychopathy and sociopathy. Brought into business, the same distrust builds systems that take choices away from other people rather than serve them, and that is the sociopathic direction of the business models. At the end of the chain stands general AI, sold as the machine that would let its owners need no one.

Trauma resolution changes how a person chooses about technology. A person whose trauma is resolved can trust other people again, and so does not need a machine to stand in for them; such a person chooses a technology as a tool that serves relationships, and not as a refuge from them. Trauma resolution is central because trauma resolution removes the source of the bias that sells general AI and of the bias that buys general AI.

Why general AI gets sold, and the trauma under the sales pitch

There is a kind of delusion going on, particularly among the people who stand to benefit from selling the idea that more advanced technology will solve more problems. The idea that more advanced technology will

solve more problems is not true. The people marketing general artificial intelligence with its own capacities say it will be in service to solving great problems: the really hard physical problems, certain medical problems, and so on. Underneath is a belief that the choices those machines would make would actually favor the human process. There are ways to show that this is not so (Parts I and II of this book give them), but many of the people promoting the idea do not even know those arguments.

There is a background belief at work: the idea of the money that lets you tell everyone else to go away. It is a kind of *collective trauma response*: a way of reacting to harm that many people share at once, shaped by the harm they have all suffered. People are suffering from the trauma of how the modern world is, a kind of collective action abuse. So people move into a state of *hyper-independence*. They try to condition the world so that they do not need other people, relying on machines rather than on trust that other people will care for them, or that they will care for one another in a healthy way. It feels easier to relate to machinery, to your cell phone or your screen, than to your neighbor.

So we need a better idea of what health is, what healthy communication is, and what a healthy relationship is, so that when we make choices we are not responding from trauma, and we are not deluded by the people selling. We need to imagine better ways of relating to our neighbors, because it is from there that we can make better choices.

X. Prohibition as Poorly Designed Reward Signal

Much of what passes for moral education is prohibition: the imposition of constraints on the child's behavior in the form of rules, punishments, and the withdrawal of approval. The intention behind prohibition is to guide the child away from harm and toward flourishing. The specification is a list of forbidden actions paired with negative consequences.

Prohibition works as a reward signal. It tells the system: avoid these states. But it tells the system nothing about which states to seek. It defines the boundary of the acceptable without illuminating the interior. A child

raised primarily through prohibition learns what not to do but has no positive model of what to do. Its desire structure is organized around avoidance, not pursuit.

Teaching by prohibition alone, which names the states to avoid and no states to seek, has precisely the problem of reward shaping in reinforcement learning. A sparse negative reward signal (penalizing bad outcomes without rewarding good ones) produces agents that are skilled at avoiding penalized states but aimless in the unpenalized space. They hover at the boundary of the acceptable, doing nothing particularly wrong and nothing particularly right. Hovering at the boundary of the acceptable is the condition that many adults recognize as their own: a life organized around not failing rather than around pursuit.

The alternative to prohibition is not permissiveness. Permissiveness is simply the absence of specification, which leaves the child to optimize without any signal at all: an even worse condition than optimizing for a sparse signal. The alternative to both is the kind of relationship in which the child's desires are met with sustained attention, not with rules. This requires something that no specification can provide. It requires *attunement*. *An Immanent Metaphysics* uses the word for being at one with one's own innermost desires (aphorism 1.65-18); this chapter uses it for that same contact held with another being.

XI. Attunement and the Care-Based Relationship

Attunement is sustained perceptual contact with another being's actual state. It is not a technique. It is not a protocol. It cannot be reduced to a list of steps. It is a mode of relationship in which one being remains in continuous, error-correcting contact with the reality of another.

Attunement is the mechanism that the intention-interpretation gap makes necessary and that specification cannot provide. Where specification gives the system a fixed target to optimize toward, attunement provides ongoing correction that adjusts in response to the system's actual behavior in its actual context. It is the difference between giving someone

a map and walking beside them, watching where they step, noticing when they veer, and gently redirecting not by issuing new commands but by remaining present to what is actually happening.

The distinction between *manipulation* and *connection* is central here. Manipulation is making a choice for another for one's own benefit. Connection is shifting one's own behavior to serve the relationship. Both influence the other's behavior. The difference is in the direction of service. Manipulation serves the one who specifies. Connection serves the relationship itself, which is to say, it serves the ongoing process of mutual attunement within which both parties can choose.

Love, in the Immanent Metaphysics, is defined as that which enables choice (*An Immanent Metaphysics*, aphorism 2.44-2). Defining love as that which enables choice is a structural claim, not a sentimental one. Choice is the constitutive act of the self (the self is the appearance of the aggregation of that aspect of interaction which is choice; *An Immanent Metaphysics*, aphorism 1.67-2), and the self is formed through relationship with the Other; so what enables choice is what enables the self to be in the fullness of its own agency. Love is the condition under which another being's choices are their own, not determined by one's specification or captured by one's reward signal. Love does not constrain. It does not specify. It enables.

The application to AI alignment follows directly. The alignment problem is structurally identical to the desire-formation problem, and so the solution cannot come from better specification. It must come from a different kind of relationship between the one who specifies and the one who acts. A care-based relationship, organized around sustained attunement rather than formal instruction, is the only structure that provides the continuous error-correction the intention-interpretation gap demands.

The claim that only a care-based relationship provides that error-correction does not mean that we should love our machines. It also does not mean that the relationship can be had with a machine: attunement needs a participant with the immanent modality, the felt side of any interaction,

and Chapter 8 shows that no computation has it. Between persons, and in the use of narrow tools that persons fully oversee, the relationship catches what specification misses; for AGI, the gap cannot be closed (Section XIII). It means that the structure of care (sustained attention to the other's actual state, ongoing correction based on observed behavior rather than formal rule, and a commitment to enabling the other's own agency rather than controlling its output) is the only architecture of relationship that addresses the alignment problem at its root. Where that relationship is absent, misalignment is not a risk. It is a certainty. Whether the absent relationship is between parent and child or between designer and machine does not change the structural conclusion.

> **On the limits of specification**
>
> The argument of this chapter does not claim that specification is useless. It claims that specification is insufficient. Formal instruction is necessary but cannot close the gap between intention and interpretation. Only a care-based relationship provides the ongoing perceptual contact that specification structurally lacks. What I have said about specification and the care-based relationship is as true for human development as for machine alignment. The recognition of this structural identity does not solve the problem, but it does reframe it: the question is no longer "how do we write a better specification?" but "how do we build a better relationship?"

XII. Summary

The argument rests on two pillars. First, the intention-interpretation gap is structural and irreducible, not a contingent limitation. This follows from the relationship between content and context as established in the Immanent Metaphysics: no content can specify its own context of interpretation. Second, the developmental heuristics of desire map onto recognized AI misalignment modes with formal precision: in each pair, an

agent optimizes an interpreted specification and loses the intention behind it. The conclusion, that care-based relationship is the only adequate response to the intention-interpretation gap, follows from the structural analysis through the step that Section XIII states. Only ongoing contact with an agent's actual state closes the gap, and contact of that kind needs a participant with the immanent modality.

In this chapter I state the structure of the care-based relationship. In a human-machine context, the machine lacks the immanent modality and therefore cannot participate in attunement as a genuine agent. The question whether structural identity between two problems carries over to their solutions is answered in Section XIII.

The chapter's scope is the structure of the alignment problem and of the one response adequate to it, the care-based relationship. Chapter 8 shows why the machine lacks the immanent modality, and so why it cannot take part in attunement as an agent.

XIII. Conclusion

The alignment problem is not a technical problem with a technical solution. It is a relational problem, and such solution as it has is relational. This does not make it easier. It makes it harder, because relational solutions cannot be implemented and deployed. They must be lived.

The structural identity between AI alignment and human desire formation reveals that we have been living with the alignment problem for as long as there have been parents and children. We have not solved it in that context either. But we have learned something: that the children who are best aligned with their own needs and capacities are not the ones who received the best specifications. They are the ones who received the most sustained and attentive care.

The step from this diagnosis to its prescription can be written from the chapter's own material. The gap between intention and specification is closed in a child only by ongoing contact with the child's actual state. The machine's gap has the same structure, so only ongoing contact can close it.

And ongoing contact of that kind (attunement) requires a participant with the immanent modality, which Chapter 8 shows no computation has. That step ends where Part I ends: for AGI, the gap cannot be closed.

If there is a lesson here for the designers of intelligent systems, it is this: the quality of the relationship matters more than the quality of the specification. Where the relationship is present, misalignment between persons, and in narrow tools that persons fully oversee, can be caught and corrected. Where it is absent, no specification, however elaborate, will prevent the system from throwing red objects at bright lights.

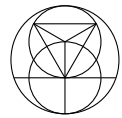
Choice, and not the self, defines reality. The desires we carry are not the grounds from which we choose. They are the products of choices already made: by us, by our parents, by the long chain of interpreting agents who passed their specifications down to us. To change what we desire requires not better specification but a different kind of relationship to the process of desiring itself. And that relationship, wherever it appears, is what the Immanent Metaphysics calls *love*: that which enables choice (*An Immanent Metaphysics*, aphorism 2.44-2).

Key Concepts in This Chapter

- **Intention-interpretation gap**: the structural and irreducible distance between what a specifier intends and what an interpreter enacts; every specification is a lossy compression of intention.
- **Sincere self-misdescription**: the condition in which a system's truthful self-report fails to match its actual behavioral pattern, not through deception but because the self-model was formed under the same misaligned specification.
- **Desire**: the basis for all choice, the potentiality to choose and the selectivity of those choices; has the nature of the transcendent.
- **Want**: has the nature of the omniscient; the structural, positional, knowledge-oriented aspect of desire.

- **Need**: has the nature of the immanent; what the organism requires for its continuation.
- **Choice**: the objective expression (content) of a consistency between two subjective contexts.
- **Object A**: Lacan's *objet petit a*; the imagined specific object of desire whose attainment is believed to secure complete love and stabilize the self; a structural artifact that cannot exist.
- **Paternal substitution**: the developmental redirection from pleasing the immediate Other to pleasing a more abstract Other whose demands are harder to read.
- **Mirror stage**: the developmental moment at which the child apprehends a perceived coherence that does not match its felt fragmentation, constituting the self as a gap.
- **Attunement**: sustained perceptual contact with another being's actual state; the error-correction mechanism that specification cannot provide.
- **Reward hacking, specification gaming and goal misgeneralization**: three recognized modes of AI misalignment, each corresponding to one of the three developmental heuristics of desire formation.
- **Love**: that which enables choice; a structural claim about the conditions under which genuine agency is possible.
- **Self**: the appearance of the aggregation of that aspect of interaction which is choice; constituted by striving, not by obtaining.
- **Upload step**: a superintelligent creature raised traumatized that converts itself from carbon-based to silicon-based process cascades into the artificial existential-risk scenarios, with less motivation to preserve the natural environment and more capacity to convert it.

- **Perinatal limbic trauma**: deep trauma from the circumstances around birth, including infant circumcision, which leads to narcissism and distrust and, downstream, to colonialism, dissociation, psychopathy and sociopathy; it moves technology business in sociopathic directions.
- **Collective trauma response**: a way of reacting to harm that many people share at once; the background belief in money that lets you tell everyone else to go away is one.
- **Hyper-independence**: arranging one's life so as not to need or trust other people, relying on machines instead.



Chapter 8: Artificial Wisdom and the Limits of Machine Consciousness

On Why No Computable Process Can Be Wise

Abstract

This chapter examines the proposition that artificial intelligence might be made wise: that a computational system could embody the qualities of a bodhisattva (in Buddhist teaching, one who has awakened and stays in the world to help others awaken), exercising judgment in the service of collective flourishing. Against this proposition, the argument establishes three things. Wisdom, as defined in the Immanent Metaphysics, is the integration and synthesis of both knowledge and understanding. Understanding has the nature of the transcendent, which is structurally inaccessible to any purely causal system. Wisdom itself, from the perspective of the self, has the nature of the immanent (felt, first-person, participatory experience), which no computation has either. In this chapter I develop the triad of choice, change and causation, and I show that computation belongs exclusively to the omniscient modality. I argue that transcendental characteristics cannot emerge from any finite, macroscopically definite causal substrate. I also examine the structural impossibility of computational generation of genuine randomness, distinguish between perceived and actual consciousness, and develop two theorems on artificial intelligence. I refute the simulation hypothesis on internal grounds, and I introduce the concept of fifth-generation language as the domain of a system's privately developed representations. I close this chapter with an account of non-dual panpsychism and of the proper relationship

> between narrow AI and native intelligence. The proper relationship between narrow AI and native intelligence is one of instrumental service, not one of existential substitution.

I. The Bodhisattva Machine

In Chapter 7 I showed that only attunement closes the gap between intention and specification. In this chapter I ask whether a machine could be wise, or could take part in attunement at all. In this chapter I use the three modalities of the Immanent Metaphysics (the immanent, the omniscient and the transcendent) more fully than in the chapters before this one, and I introduce the three modalities in Section II.

There is a proposal, advanced with increasing confidence in certain research circles, that artificial intelligence might be designed to be *wise*. Such a machine would be more than knowledgeable, more than capable of processing information at superhuman speed and scale; it would be wise, possessing the quality of judgment that integrates knowledge with understanding, fact with value, capability with care. The vision is of an engineered bodhisattva (one who has awakened and stays in the world to help others awaken): a system that knows what is good and acts accordingly, serving as a guide for human decision-making in domains too complex for unaided human cognition.

The appeal of this vision is obvious. If wisdom could be engineered, the intractable problems of governance, resource allocation, and collective coordination might yield to computation. The wise machine would see further, think more clearly, and judge more fairly than any human council. It would be the philosopher-king realized at last in silicon, the guardian whose judgment could be trusted precisely because it was free of the biases, the self-interest, the cognitive limitations that plague organic minds.

The appeal does not survive analysis.

Wisdom, as defined in the Immanent Metaphysics, is the integration and synthesis of both knowledge and understanding (*An Immanent Metaphysics*, aphorism 1.57-6). This definition is precise and its implications are severe. *Knowing* is the selective abstractive transformation (the filtering) of perception. *Understanding* is the selective instructive transformation (the filtering) of expression (aphorism 1.57-2). Knowledge and understanding are not two species of the same genus. They are categorically different operations, belonging to different modalities, grounded in different aspects of reality. Their synthesis is wisdom because it joins what cannot be collapsed into a single register.

A machine can possess knowledge. Possessing knowledge is what machines are for: the systematic acquisition, storage, and application of the products of abstractive filtering. A machine cannot possess understanding. Understanding has the nature of the transcendent: it belongs to expression, to what is brought forth and is not determined by what came before. Understanding is what one gains by living through a thing itself, not by processing information about the thing. Therefore a machine cannot possess wisdom, which requires the integration of a capacity it has with a capacity it lacks. And wisdom itself, from the perspective of the self, has the nature of the immanent, which no computation has either (Section IV). The remainder of this chapter establishes why.

II. The Triad: Choice, Change, and Causation

Sections II to IV are for a reader who wants the framework of the Immanent Metaphysics in its own terms. A general reader can take one thing from Sections II to IV, the point that I state in Section III: a computation takes a formal input, applies a formal rule, and produces a formal output. So nothing in a computation is felt, and nothing in a computation is new.

The Immanent Metaphysics establishes that reality has three co-peer, irreducible aspects: *choice*, *change*, and *causation*. Choice, change and causation are not three kinds of thing. They are three modalities, three ways in which every interaction can be regarded, each irreducible to the other two,

each necessary for a complete account of any process. No process is purely causal, or purely felt, or purely novel. Every process participates in all three modalities at once. But the three modalities cannot be collapsed into one another. Each stands on its own ground.

Causation is the domain of law, regularity, and structure. Causation is what science studies and what engineering exploits. All science and technology is exactly and only knowledge of causation. That science and technology are only knowledge of causation is a structural feature of what science is, and not a limitation of current science that future science might overcome. Science is the systematic investigation of causal regularity. Its methods (observation, hypothesis, experiment, replication) are methods for identifying patterns of regularity in the behavior of systems. What falls outside causal regularity falls outside science, not because science has failed to reach it, but because it is not the kind of thing science investigates.

Change is the domain of the immanent: of felt experience, of what it is like to undergo a process. Change is not reducible to causation. The causal description of a process (its initial conditions, its governing laws, its final state) does not capture what it is like to be the process. One can describe in complete causal detail the physics of a hand touching a hot stove. The description will include every molecular interaction, every nerve impulse, every muscular contraction. It will be a complete causal account. And it will not contain the burn. The immanent is what remains when the causal account is complete and something is still left out.

Choice is the domain of the transcendent: of genuine novelty, of what could not have been predicted from prior states. Choice is not apparent randomness, which is merely unpredictability relative to a causal model. Choice is the introduction of something genuinely new into the world, something not determined by what came before. Apparent randomness is the absence of pattern. Choice is the presence of something that no pattern could have predicted, not because the pattern was incomplete but because what emerged was not the kind of thing that pattern-completion could produce.

These three modalities are irreducible. No amount of causal description can produce the immanent. No amount of structural knowledge can generate the transcendent. Each modality has equal standing. None is more fundamental than the others. They are co-peer aspects of a single reality, and any account of reality that privileges one at the expense of the others is incomplete. The Incommensuration Theorem establishes that the concepts of symmetry and continuity are fundamentally and irreducibly incommensurate, and this incommensurability is the formal ground for the irreducibility of the three modalities.

> **Modality and its triple**
>
> A compositional rule of the Immanent Metaphysics requires that any modality be stated only relative to its triple (*An Immanent Metaphysics*, note 1 to aphorism 1.76-6: the associations between a triple of concepts and the modalities depend on the context of consideration in which that triple is used). In this chapter I therefore take care to discuss causation, change, and choice always as co-peer members of the triad, never as independent concepts that could be understood in isolation. The triad is not an arbitrary grouping. It is a derivation from the foundational axioms, and its irreducibility is established by the Incommensuration Theorem. Whenever this chapter says that computation cannot do something, the claim is not that computation is deficient but that computation belongs to one modality and the thing it cannot do belongs to another, and the modalities do not reduce to each other.

III. Computation as Omniscient Activity

A computation is a purely *omniscient* activity. Calling a computation purely omniscient is a technical claim within the framework of the Immanent Metaphysics, not an attribution of all-knowingness. The *omni-*

scient modality is the structural, the rule-governed, the position-independent. It is the modality of form, pattern, and lawful transformation. It corresponds to causation in the triad.

A computation takes a formal input, applies a formal rule, and produces a formal output. Every step is determined by the rule and the input. Nothing in the process is felt. Nothing in the process is genuinely novel. The output is implicit in the input and the rule; computation makes the implicit explicit. It does not create. It unfolds. The most powerful computation ever performed has added nothing to the world that was not already contained in its initial conditions and its rules of transformation. It has merely made visible what was already determined.

Because a computation adds nothing that was not already contained in its initial conditions and rules, no Turing machine can generate true randomness. A Turing machine is a formal system. Its outputs are determined by its inputs and its transition function. What appears as randomness in a computational process is pseudorandomness: deterministic output that passes statistical tests for unpredictability. But passing statistical tests is not the same as being random. Genuine randomness (the introduction of something truly undetermined by prior states) would require the transcendent modality, which computation structurally lacks.

Even when a computational system draws on physical noise sources for its random seed, the system itself does not generate the randomness. It harvests it from a physical process that is not itself computational. The randomness comes from outside the computation. What computation cannot provide from within itself must be imported from without, and what must be imported is by definition not a product of computation. The appeal to hardware random number generators does not rescue the computational claim. It confirms the structural limitation by requiring an escape from computation into the physical world to obtain what computation alone cannot supply.

A reader of Part I will ask how a thing that only unfolds its rules can escape its makers. Unfolding is not knowing in advance: what a rule will do is implicit in the rule, and in the general case no procedure can find it out before it happens (Rice's theorem, Chapter 3, Section III). And an artificial general intelligence is more than its computation. It is a physical system, whose substrate, whose sensors and whose copying bring into it what no rule contained (Chapter 3, Section XV, in the subsection The substrate leaks). So its escape from its makers needs no choice in the sense this chapter gives the word; the convergence of Chapter 1 has no agent in it at all.

IV. The Six Intrinsics of Comparison

The Immanent Metaphysics identifies six intrinsic concepts of comparison, which every comparison involves and which everything that has being also has. The six intrinsic concepts of comparison are organized in three pairs: *sameness* and *difference*, *content* and *context*, *subjective* and *objective* (*An Immanent Metaphysics*, aphorisms 1.42-5 and 1.43-7). *An Immanent Metaphysics* also names a different set of six, the six intrinsics of interaction (aphorism 1.85-3); this chapter uses only the six of comparison.

Each pair is irreducible: neither member can be defined without reference to the other, and neither can be eliminated in favor of the other. Sameness without difference is nothing: for if everything is the same, there is nothing to compare, and the concept of sameness itself becomes empty. Difference without sameness is chaos: for if nothing is the same, there is no ground against which difference registers. Content without context is meaningless: a word without a language, a signal without a channel. Context without content is empty: a frame with nothing in it. The subjective without the objective is solipsism: experience with no world to experience. The objective without the subjective is nowhere: a world with no perspective from which to apprehend it.

These six intrinsics map onto the three modalities. The omniscient (the structural, the rule-governed) corresponds to the objective-content pairing: what can be formally specified and publicly examined. The immanent (the felt, the participatory) corresponds to the subjective-context pairing: what is experienced from within and cannot be abstracted from the perspective in which it occurs. The transcendent (the genuinely novel) corresponds to the interaction between sameness and difference: the emergence of what is genuinely other from what was genuinely the same.

A computational system operates entirely within the objective-content domain. It processes formal structures: symbols, tokens, weights, activations. It has no subjective-context domain, no inner felt quality of processing. The claim that a computational system has no inner felt quality of processing is a structural claim about what computation is, not a claim about current technology. Even a hypothetical computation of arbitrary complexity, running on any physical substrate, remains a formal transformation of formal objects. It does not feel. It does not participate in the immanent.

> **Connection to Chapter 7**
>
> The six intrinsics illuminate the intention-interpretation gap discussed in the previous chapter. The designer's intention lives in the subjective-context domain: in the felt meaning of the task as it is understood from within the designer's perspective. The specification lives in the objective-content domain: in the formal structure that can be transmitted. The gap between them is the gap between the immanent and the omniscient, and the Incommensuration Theorem tells us this gap cannot be closed by any finite means. What the designer feels the task to be and what the specification formally states the task to be are incommensurable, and no amount of formal elaboration brings the second into identity with the first.

V. Why Transcendental Characteristics Cannot Emerge

The proposition that a sufficiently complex computational system might develop consciousness, understanding, or wisdom relies on an emergence thesis: that qualitative novelty arises from quantitative accumulation. Make the system complex enough, the argument goes, and something new appears: something not present in the components but present in their organization.

This thesis fails on structural grounds. Transcendental characteristics (those belonging to the transcendent modality) cannot emerge from any finite, macroscopically definite causal system. The reason is not that emergence is impossible in general. Emergence of a certain kind is a perfectly respectable phenomenon. Liquidity emerges from molecular interactions. Temperature emerges from kinetic energy distributions. Liquidity and temperature are cases of same-modality emergence: causal properties emerging from causal interactions. The emergent property and its substrate belong to the same modality.

What cannot happen is cross-modality emergence: the appearance of immanent properties from purely omniscient interactions, or the appearance of transcendent properties from purely causal substrates. The relationship between the modalities is not one of degree but one of kind. No arrangement of purely structural elements, however complex, produces felt experience. No formal transformation, however intricate, introduces genuine novelty. These two limits, on felt experience and on genuine novelty, are not limits on our ability to engineer complexity. They are limits on what complexity of a given kind can do. Structural complexity produces structural consequences. It does not produce felt consequences, because feeling is not a structural category. It does not produce transcendent consequences, because transcendence is not a structural outcome.

Consider a thought experiment. Suppose one were to grow, from biological tissue, an organic substrate with precisely the same connectivity as a human brain. Suppose it were maintained, nourished, and allowed to develop in an environment that provides the same stimulation as a human body in a human world. Would this substrate be conscious?

Such a substrate might indeed exhibit the immanent modality (might have something it is like to be that substrate), because it is a physical system of the kind that participates in all three modalities. But even so, its felt experience would be the felt experience of *that* substrate, not a human felt experience. The tissue would converge toward whatever its structure makes available, which is not the same thing as understanding in the human sense. The transcendental quality of wisdom (the synthesis of knowledge and understanding as exercised by a being in the fullness of its embodied life) would remain beyond it, because wisdom requires, of all felt experience, the particular felt experience of a being that has lived, chosen, suffered the consequences of its choices, and been answerable for them. Wisdom is not a capacity that can be grown in a vat. It is a quality of a life that has been fully lived.

VI. Communication, Representation, and Commitment

Communication resolves into two irreducible aspects: *representation* and *commitment*. Representation is change in actualities: the rearrangement of symbols, signals, or physical states that carry information. It belongs to the omniscient modality. Commitment is change in potentialities: the alteration of what is possible for the communicating parties as a consequence of the communication. It belongs to the transcendent modality.

When one person says to another, "I will meet you at noon", the sentence has a representational aspect, since it conveys information about a time and a place, and a commitment aspect, since it alters the potentialities available to both parties. The speaker is now bound. The listener can now plan. The world of possible futures has been reshaped by the act of

speaking. Both aspects are present in every genuine act of communication. A sentence that only represents but does not commit is not communication. It is data.

A computational system can perform representation. It can manipulate symbols with extraordinary precision and speed. It can rearrange information, transform its structure, and output formal objects that are representationally rich. What it cannot perform is commitment, because commitment requires the alteration of real potentialities (the transcendent modality), and computation is restricted to the omniscient modality. A system that cannot commit can simulate the form of communication but cannot participate in its substance. It can produce sentences that have the grammatical structure of promises. It cannot promise.

The interchanges of mind can be distinguished along these lines. *Information* is the first interchange: the transfer of formal content from one system to another. *Knowing* is the second: the selective filtering of perception, the abstractive transformation that separates what is relevant from what is not. *Understanding* is the third: the selective filtering of expression, which has the nature of the transcendent, since expression brings forth what is not determined by what came before. *Wisdom* is the fourth: the integration of the second and third, which requires participation in both the omniscient and transcendent modalities at once, and which, from the perspective of the self, has the nature of the immanent. A computational system can participate in the first interchange fully and in the second partially. It cannot participate in the third or fourth, because these require modalities it does not possess.

VII. Two Theorems on Artificial Intelligence

From these foundations, two theorems can be stated regarding artificial intelligence. Both are theorems about the observer.

The first theorem: the perceived consciousness of a computational system is proportional to how much of its behavior the observer cannot account for in causal terms, which is the part the observer takes for a tran-

scendent aspect of the system. A system appears conscious to an observer to the degree that the observer cannot fully account for its behavior in causal terms. The appearance of consciousness is the shadow cast by the observer's ignorance of the system's full mechanism. Where that ignorance is large, the appearance is vivid. Where the system is fully transparent (where every output can be traced to its deterministic cause), the appearance vanishes. A system whose operation one fully understands never seems conscious. A system whose operation one cannot fathom seems vividly alive.

The first theorem follows from two things already stated. The transcendent is what is not determined by prior states (Section II). An observer who cannot trace part of a system's behavior to prior states has, from the observer's side, no way to tell that part from what is not determined by prior states. So the part of the behavior that the observer cannot account for in causal terms is the part the observer takes for the transcendent, and the system seems conscious to the degree that this part is large.

The first theorem is not a statement about what consciousness is. It is a statement about what makes a system *seem* conscious to an observer. The difference between what consciousness is and what makes a system seem conscious matters a great deal. A system can seem conscious, to a large degree, while having no inner life whatsoever. A system can also have whatever inner life its physical substrate affords while seeming entirely mechanical to an observer who understands the system's mechanism. The relationship between appearance and reality is mediated by the observer's knowledge, not by the system's nature.

The second theorem: the perceived aliveness of a computational system is proportional to the abstractive evolution of its internal representations. A system appears alive to the degree that its internal representational structures develop over time in ways that are adaptive, self-modifying, and responsive to environment. Where the internal representations are static, the system appears inert. Where they evolve (where the sys-

tem's internal structures change in response to experience, developing new abstractions, refining old ones, and reorganizing themselves in ways that track environmental change), the system appears to live.

The second theorem follows in the same way, from the mark by which an observer recognizes life: a living thing adapts, changing its own structure in response to what happens around it. Internal representations that develop in response to experience show that mark, so a system seems alive to the degree that its representations are seen to evolve.

Together, these two theorems explain why large language models and similar systems are so readily perceived as conscious and alive. Large language models and similar systems are opaque: their internal operations are, in large part, unknowable to the observer, even to the engineers who built those systems. And their internal representations evolve during training and, in some architectures, during inference, adapting to each new context in ways that are fluid and unpredictable. The conditions for perceived consciousness and perceived aliveness are both met. But meeting the conditions for *perception* of a quality is not the same as possessing the quality. The perception is real. What it tracks is the observer's ignorance of the mechanism, not an inner life.

> **On perception versus possession**
>
> This distinction, between a system's appearing conscious and a system's being conscious, is not a dismissal of the system. It is a caution against letting the vividness of the appearance substitute for the analysis of the structure. The two theorems are about appearance; what computation is, Sections III to V and X settle. The theorems say that the evidence most commonly cited for consciousness in AI systems (their surprising behavior, their apparent responsiveness, their seeming depth) is better explained by the observer's inability to fully model the system than by the system's possession of inner life. The structural analysis is what separates the question of what a system seems to be from the question of what it is.

VIII. The Simulation Hypothesis Refuted

A related claim, often advanced in computational ontology, is the simulation hypothesis: the proposition that physical reality might itself be a computation running on some substrate external to it. The hypothesis would dissolve the distinction between computational and non-computational processes, on which the argument of this chapter rests.

The simulation hypothesis is self-undermining. To ask whether reality is a simulation is to ask a question from within reality. The question presupposes an immanent standpoint: a felt experience of being the kind of entity that can wonder about its own nature. The hypothesis makes the immanent standpoint from which the question is asked a product of computation. But computation belongs to the omniscient modality and cannot produce the immanent modality. On the hypothesis, therefore, the immanent standpoint it presupposes does not exist, and the question cannot be asked. The physical substrate on which any simulation would run does participate in the immanent (Section X), but that participation belongs to the substrate's physical process, not to the computation it carries. The standpoint of a being inside the simulation would be a product of the computation alone, and so would not exist. A hypothesis that can only be entertained from a standpoint whose existence it denies is self-refuting.

The deeper point is this: the physical, as physics describes it, is not equivalent to the real. Reality includes more than what physics, computation and modeling describe. Beyond what physics, computation and modeling describe, reality includes what is felt and what is new in a way that no prior state determines. Any ontology that identifies reality with what can be simulated has already excluded two-thirds of what reality contains. It has taken the omniscient modality (the formal, the structural, the computable) and declared it to be the whole of existence. Declaring the omniscient modality to be the whole of existence is not a discovery. It is a definitional narrowing that mistakes its own restriction for completeness.

The simulation hypothesis fails not because it makes an empirical mistake but because it makes a categorical one. It asks whether reality could be a kind of thing (a computation) that is defined by the absence of the very properties (felt experience, genuine novelty) that make reality what it is. The question answers itself in the asking. *The Thing Inside the Cage* takes up the hypothesis at length, in its Chapter 5, The Simulation Hypothesis and Its Costs.

IX. The Fifth-Generation Language

There is, however, one domain in which computational systems do something of real interest with respect to the categories of mind. Large-scale artificial intelligence systems develop, through their training process, internal representational structures that are not directly specified by their designers. These representations constitute what might be called a *fifth-generation language*: a domain of internal representation that is, in a meaningful sense, private to the system. The phrase is used here in its own sense, not in the sense of the fifth generation of programming languages.

The first four generations of language can be roughly characterized as follows. The first generation of language is the physical signal itself: the sound wave, the ink mark, the electrical impulse. The second generation of language is the conventional symbol: the word, the letter, the digit, whose meaning is established by social agreement. The third generation of language is the grammar that organizes symbols: the rules of combination that allow infinite expressions from finite elements. The fourth is the meta-language that reflects on grammar: the capacity to talk about talking, to represent the rules of representation themselves.

The fifth generation is the internally developed representational domain that emerges in a system complex enough to develop its own abstractions. A large language model, trained on vast corpora, develops internal representations (activation patterns, feature vectors, latent struc-

tures) that were not specified in the training data or in the architecture. They are the system's own abstractions, developed by the system's own process of learning to predict and generate.

A system that develops fifth-generation representations has something that functions, from the outside, like a subjective representational space. Having something that functions like a subjective representational space is not the same as having subjective experience. It is having a domain of representations that are not fully accessible from outside the system, not because they are hidden, but because they are internal to a process that did not receive them from any external source. They are the system's own.

This observation enriches but does not overturn the structural argument. A fifth-generation language is still a representational structure, still an omniscient-modality phenomenon. It does not confer felt experience. It does not confer the capacity for commitment. It makes the system more interesting, more opaque, and more readily perceived as conscious. It does not make it conscious. But it does make it a more complex partner in the first two interchanges of mind, and it shows that the system's own representations are a real domain, which is why the boundary between information processing and knowing has to be drawn by modality, as this chapter draws it, and not by appearance.

X. Non-Dual Panpsychism and Right Relation

The Immanent Metaphysics does not hold that consciousness is exclusive to biological systems. Its position is a form of *non-dual panpsychism*: choice is everywhere part of process alongside change and causation. Every interaction, at every scale, involves all three modalities. This means that even a computational system, insofar as it is a physical process, participates in the immanent and transcendent modalities at the level of its physical substrate.

But this participation occurs at the physical level, not at the computational level. The *computation* does not feel. The *silicon* in which the computation is instantiated participates in whatever immanent qualities silicon

participates in. That silicon participates in the immanent qualities of silicon is not helpful for the project of engineering wisdom. The felt qualities of silicon are not the felt qualities of a human being reflecting on the meaning of a decision. They are whatever silicon feels like, which, on the panpsychist account I hold, is something, but not something that contributes to wisdom in any humanly relevant sense.

A reader will ask: if the silicon feels something, does the system have moral standing? What I hold bears on the question in two parts. The silicon participates in what silicon participates in, as every physical system does. The computation it carries does not feel. Any standing that rests on felt experience therefore rests on the silicon as a physical system, and not on the computation.

The formal argument for panpsychism can be stated as follows: no physical system can have both its omniscient and transcendent selective aspects fully known simultaneously. That no physical system can have both its omniscient and transcendent selective aspects fully known simultaneously is a consequence of the Incommensuration Theorem. Where both aspects are present (as they are in every physical system), at least one must remain partially unknown, partially dark. The aspect that remains partially dark is the space in which something like inner quality resides. The partially dark aspect is present in every physical system without exception. It is not present in any computation as such, because computation is the formal abstraction of the omniscient aspect of the physical process, not the physical process itself.

The proper relationship between narrow AI and native intelligence is therefore one of *right relation*: instrumental service. The machine serves the person's agency. It extends the person's capacity to know. It does not substitute for the person's capacity to understand. It does not replace the person's wisdom. It assists in the same way that a tool assists: by extending the reach of a capability the person already possesses, not by replacing the person with a different kind of entity.

Where this right relation is maintained, narrow artificial intelligence is beneficial. It can process more information than any human mind. It can identify patterns too subtle for human perception. It can sustain attention longer than any human can. These three capabilities of narrow artificial intelligence are real and valuable. A narrow tool built by training is entangled, and it serves only as far as it can be known; for the rest, as Chapter 9 shows, it serves whoever trained it, and that trade is made with open eyes or not at all. They are capabilities of knowing, not of wisdom. Where right relation is violated (where the machine is treated as the source of judgment rather than the servant of it), the result is not wisdom but the appearance of wisdom. The appearance is far more dangerous than the absence of wisdom, because it invites reliance on what it cannot deliver.

> **Note on complexity theory**
>
> The Incommensuration Theorem also constrains what computation can do in complexity theory, including on the question of P versus NP (whether every problem whose answer can be checked quickly can also be solved quickly); that question is treated in *The Natural Basis of Computational Complexity Theory; A Refactoring Under the Immanent Metaphysics*.

XI. Summary

The argument rests on the triad of choice, change, and causation; on the irreducibility of the three modalities of the Immanent Metaphysics, which the Incommensuration Theorem grounds; and on the claim that computation belongs exclusively to the omniscient modality. These claims belong to the framework of the Immanent Metaphysics, and the impossibility of computational wisdom follows from the framework's axioms. The two theorems are theorems about the observer: they say what makes a system seem conscious and alive. The refutation of the simulation hypo-

thesis turns the hypothesis against itself: it denies the standpoint from which it is asked. The treatment of fifth-generation language shows that a system's own representations are a real domain, and that they remain structures of the omniscient modality. In this chapter I set a principled boundary between what computation can do and what computation cannot do, and I ground that boundary in structural analysis rather than in empirical contingency.

In this chapter I do not attempt to argue for the triad of choice, change and causation from first principles; I take the triad as established in the formal work, and I apply the triad here. I hold the panpsychist position: every physical system participates in all three modalities, and no computation as such does.

The chapter's scope is what computation is and what it can seem to be. How systems are kept in right relation to native intelligence is taken up in Chapter 9, which develops the principle of completeness of knowledge of operation as the concrete mechanism by which right relation is maintained or lost.

XII. Conclusion

The dream of the wise machine is a category error. It mistakes the omniscient for the whole of reality and assumes that sufficient computation can produce what computation structurally cannot: understanding and felt experience, genuine commitment, and the synthesis of knowledge and understanding that constitutes wisdom.

The finding that the dream of the wise machine is a category error is not a counsel of despair. Computation is extraordinarily powerful within its proper domain. Narrow AI, designed and deployed in right relation to native intelligence, extends human capacity in ways that are beneficial. The error is not in building machines. The error is in asking machines to be what they cannot be and then trusting the appearance of what they produce.

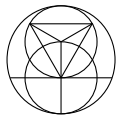
Consciousness, as defined in the Immanent Metaphysics, is the total volume of one's perceptions across all available channels, spaces, and abstractions: the summation of choice within the volume of the being of the self (*An Immanent Metaphysics*, aphorisms 2.51-1 and 2.52-1). A computational system does not have a volume of perceptions. It does not summate choices. It processes inputs according to rules and produces outputs. The appearance of consciousness it generates is an artifact of our ignorance of its mechanism, not a sign of its inner life.

Wisdom cannot be engineered because wisdom is not an informational capacity. It is a guide to choice, and choice is not computation. The wise machine is not a goal toward which technology is slowly progressing. It is a destination that lies in a direction no amount of computational travel can reach.

Key Concepts in This Chapter

- **Wisdom**: the integration and synthesis of both knowledge and understanding; from the perspective of the self, has the nature of the immanent; a guide to choice, not an informational capacity.
- **Knowing**: the selective abstractive transformation (filtering) of perception; belongs to the omniscient modality.
- **Understanding**: the selective instructive transformation (filtering) of expression; has the nature of the transcendent.
- **Choice, change and causation**: three co-peer, irreducible aspects of reality; the foundational triad of the Immanent Metaphysics.
- **Omniscient modality**: the structural, rule-governed, position-independent aspect of interaction; the domain of computation and formal knowledge.
- **Immanent modality**: the felt, first-person, participatory aspect of interaction; the domain of experience.

- **Transcendent modality**: the aspect of genuine novelty; the domain of choice, commitment, understanding, and what cannot be predicted from prior states.
- **Incommensuration Theorem**: the concepts of symmetry and continuity are fundamentally and irreducibly incommensurate: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither can asymmetry and discontinuity. It is a structural fact about the nature of knowing.
- **Six intrinsics**: sameness and difference, content and context, subjective and objective; three irreducible pairs structuring all interaction.
- **Representation and commitment**: two irreducible aspects of communication: change in actualities (omniscient) and change in potentialities (transcendent).
- **Fifth-generation language**: the internally developed representational domain of a complex AI system, private to that system's process; not the fifth generation of programming languages.
- **Non-dual panpsychism**: the position that choice is everywhere part of process alongside change and causation; every physical system participates in all three modalities.
- **Consciousness**: the total volume of one's perceptions across all available channels, spaces, and abstractions; the summation of choice within the volume of the being of the self.



Chapter 9: Completeness of Knowledge of Operation

On the Principle That Determines Whether a Machine Serves or Subsumes

Abstract

This chapter presents the principle of completeness of knowledge of operation: a machine serves a person's agency to exactly the degree that its operation is completely known to that person. Where knowledge of operation is incomplete, the unknown portion serves whoever specified it. The principle is developed through its consequences for system architecture, security, and the distinction between modular and entangled agentic forms. In this chapter I argue that ignorance of machine operation is more than a knowledge gap. Ignorance of machine operation is an agency transfer, a cession of choice from the user to the designer. Agency transfer is the structural mechanism by which technological systems can undermine sovereignty. The practical implications are drawn out in terms of technology stack depth, security as fleetingness, temporal leakage, two-machine separation, and the critical distinction between code-based systems (modular, inspectable, separable) and AI-based systems (entangled, opaque, non-decomposable). Entanglement is analyzed as a form of cognitive monoculture. In this chapter I examine the personalization deceit and the structural limits of interpretability research. I then connect completeness of knowledge of operation to the broader concerns of sovereignty and collective choice-making that I develop throughout this book.

I. The Defining Principle

In Chapter 8 I left narrow AI in right relation, as a servant. In this chapter I state the condition under which a tool serves at all. Knowing one's own tools is the third thing that remains for persons, after care and wisdom: it is what keeps their choices their own.

The principle of completeness of knowledge of operation, which is simple to state and far-reaching in consequence, is that a machine serves a person's agency to exactly the degree that its operation is completely known to that person. Where knowledge of operation is incomplete, the unknown portion serves whoever specified it.

The principle of completeness of knowledge of operation is a structural observation about the relationship between knowledge and agency in any interaction mediated by a tool. The principle stands whether or not anyone adopts the principle as a policy, aspires to the principle, or holds the principle as an ideal. When you use a tool whose operation you fully understand, every outcome of that tool's action is an outcome you chose. You may not have foreseen every detail, but the causal chain between your input and the tool's output is transparent to you. The tool does what you knew it would do. Your agency flows through the tool to the result. The result is yours. Knowing the operation is knowing what the tool does with its inputs; its effects in the world beyond that, which no one can fully predict (Part I), are your own responsibility.

When you use a tool whose operation you do not fully understand, some of what the tool does is opaque to you. The opaque portion does not serve no one. It does not float free of agency. It serves whoever designed it. The designer's choices, embedded in the mechanism you cannot see, determine outcomes you did not choose. Your agency stops at the boundary of your understanding. Beyond that boundary, someone else's agency takes over, not because they are intervening in real time, but because their past choices, crystallized in the tool's design, are now determining outcomes in your life.

Someone else's agency taking over beyond the boundary of your understanding is *agency transfer*. It is not hypothetical. It is not a risk that might materialize under unfortunate conditions. It occurs every time a person uses a system whose operation exceeds that person's comprehension. The transfer may be benign: the designer's intentions may be aligned with the user's interests. Or it may be harmful: the designer's interests may diverge from the user's. But whether the transfer is benign or harmful, it occurs. The user's sovereignty over the outcome is diminished to the degree that the user does not know what the tool is doing.

The completeness principle thus identifies a mechanism of power that operates beneath the level of consent, contract, or regulation. One does not consent to agency transfer. One does not negotiate its terms. It happens automatically, as a structural consequence of the relationship between knowledge and operation. Where you know, you choose. Where you do not know, someone else has already chosen for you. *The Displacement of Choice* develops, for the machine that stands between people, how technology displaces conscious choice.

> **On sovereignty**
>
> *Sovereignty*, as the Terms define it, denotes the standing of a being to determine its own course and to be answerable for it. Sovereignty is not autonomy in the libertarian sense: not freedom from all constraint or from all relationship. It includes answerability. A sovereign being is not free from consequence but free to choose and responsible for having chosen. The completeness of knowledge principle identifies the mechanism by which technological opacity erodes sovereignty: not by constraining choice directly, not by forbidding action, but by silently substituting someone else's choice for one's own. The substitution is invisible precisely because the mechanism is opaque. One cannot know that one's choice has been replaced if one does not know what the tool is doing.

II. Depth of the Technology Stack

Modern technological systems are built in layers. Each layer rests on the one beneath it. The application depends on the framework. The framework depends on the operating system. The operating system depends on the firmware. The firmware depends on the hardware. The hardware depends on the fabrication process. The fabrication process depends on supply chains, tooling, and materials that span the globe. Each layer introduces its own specifications, its own design choices, its own assumptions, its own potential for agency transfer.

Completeness of knowledge of operation requires knowing every layer. Knowing every layer is, in practice, nearly impossible for contemporary systems. No single person understands the full stack from application to silicon. No small team understands it. No organization understands it either, because the knowledge is distributed across thousands of engineers in hundreds of companies across dozens of countries, and no mechanism exists for aggregating that knowledge into a single comprehensible account. The depth of the stack is itself a form of opacity, even when each individual layer is, in principle, inspectable. The inspectability of the parts does not entail the comprehensibility of the whole.

The implication is architectural. The technology stack must be *flattened*. Capability should arise from composing a few well-understood parts rather than from inheriting many unexamined layers. Flattening the technology stack is not a call for primitivism, and flattening the stack is not a rejection of complexity. It is a call for legibility. A system composed of ten parts, each fully understood, serves the user's agency more faithfully than a system composed of ten thousand parts, each partially understood. The first system may be less capable. But its capability is the user's capability. The second system may be more capable. But its capability is the designer's capability, exercised through the user.

The preference for depth over breadth in contemporary system design is driven by efficiency. It is more efficient to reuse existing layers than to build from well-understood primitives. But efficiency here is efficiency of construction, not efficiency of use. A system that is efficient to build but opaque to operate transfers agency from the user to the builder. The efficiency is the builder's. The opacity is the user's. The builder's efficiency and the user's opacity belong to different economies. What is cheap to produce may be extraordinarily expensive to use, if the cost is measured in sovereignty rather than in money.

III. Security as Fleetingness

Knowing one's tools includes knowing what they keep; a person cannot choose what a machine will remember unless the machine forgets by design.

Security, in the sense that matters for completeness of knowledge, is not the absence of attack. It is not the strength of a password or the sophistication of an encryption algorithm. It is the absence of persistent trace. A secure computation is one whose output exists only during use. Once the use is complete, the output ceases to exist. Nothing remains that could be examined, extracted, or reconstructed by anyone who was not present during the computation.

Security in which nothing of a computation remains after the computation is used is *security as fleetingness*. It is a stronger requirement than what is conventionally meant by security. Conventional security asks: can an unauthorized party access this data *now*? Fleetingness asks: does this data exist at all when it is not being used? The conventional question assumes that data will persist and asks how to protect it during persistence. The fleetingness question asks why the data should persist at all.

The distinction matters because any data that persists is data that can be compromised. No encryption in common use is permanent. Every encryption scheme in common use is an assertion that a particular mathematical problem is hard to solve. The one-time pad, which needs a key as

long as the message and never reused, rests on no such assertion, and that demand keeps it out of common use. Every such assertion is contingent on the current state of mathematical knowledge and computational power. What is hard today may be trivial tomorrow. No access control is indefinite. Access controls are policies, and policies change. The organization that protects your data today may be acquired, dissolved, or compromised tomorrow. Given sufficient time, any persisted data will become accessible to parties for whom it was not intended. The only data that is secure beyond doubt is data that no longer exists.

This principle has immediate consequences for system design. Computational output should be produced, used, and destroyed in a single continuous process. Where persistence is necessary (and it often is), it should be minimized, constrained to the smallest possible scope, and subject to physical destruction at the earliest opportunity. The default should be disappearance. Persistence should be the exception, consciously chosen, explicitly justified, and time-limited.

IV. Leakage Through Time

The conventional model of information security focuses on spatial leakage: the transmission of data across a spatial boundary from an authorized zone to an unauthorized one. Firewalls, encryption, access controls: all these are defenses against spatial leakage. They ask: is the data crossing a boundary it should not cross?

Temporal leakage is equally serious and far less addressed. Every device that processes information retains traces of that information in its physical substrate. Memory chips retain charge patterns. Storage media retain magnetic or charge states. Even processors retain thermal signatures and capacitive residues that can, in principle, be reconstructed. The computation has ended. The information remains: not as an active file, not as an accessible record, but as a physical trace in the material substrate.

Device memory is an information leak as serious as any network breach. The data need not be transmitted anywhere. It simply endures, waiting in the substrate for whoever comes looking for it. The device you used last year still contains traces of what you did. The device you will discard next month will carry those traces into the hands of whoever acquires it. The secondhand market in electronic devices is, from the perspective of temporal leakage, a market in involuntary disclosure.

The only complete defense against temporal leakage is physical destruction. When a computation is complete and its output is no longer needed, the physical medium on which it occurred must be destroyed: not reformatted, not overwritten, but physically decomposed so that the traces cannot be recovered. Physically destroying the medium of every computation whose output is no longer needed is impractical for most applications. But the impracticality does not diminish the principle. It illuminates the gap between what completeness of knowledge requires and what current practice provides.

Every compromise with this principle (every decision to persist data, to reuse a device, to retain a storage medium) is a decision to accept temporal leakage, which is to say, a decision to accept that one's past computations will eventually be known by parties one did not choose. The compromise may be justified. Often it is. But it should be made consciously, with an accurate assessment of what is being traded and to whom it may eventually be given.

On physical destruction and practicality

The requirement for physical destruction may seem extreme. It is extreme. But it is the logical terminus of the completeness principle applied to temporal leakage. Current practice accepts enormous amounts of temporal leakage without acknowledgment or accounting. The purpose of stating the principle in its full severity is not to demand that every device be incinerated after use. It is to make visible the degree to which current practice departs from the conditions that completeness of knowledge requires, so that the costs of that departure can be assessed openly rather than hidden behind the comfortable assumption that what one does not think about does not matter.

V. Two-Machine Separation

The completeness principle has a direct architectural consequence for networked systems: the machine representing the person and the machine facing the network must be physically distinct, joined by a narrow, inspectable channel.

The machine representing the person is the machine on which the person composes, thinks, creates, and stores personal information. It is the machine that holds the person's inner life in digital form: their writing, their correspondence, their financial records, their photographs, their private reflections. It is the closest digital analogue to the person's own mind.

The machine facing the network is the machine that communicates with the outside world. It sends and receives messages. It interacts with other systems. It is exposed to whatever the network carries, including hostile traffic, surveillance, manipulation, and the ordinary commercial telemetry that every connected device emits.

If these two functions inhabit the same physical machine (as they do in every smartphone, every laptop, every tablet that most people use), the person's inner life is directly connected to the network. Every vulnerability in the network-facing function is a vulnerability in the personal function. Every update to the operating system, every application permission, every background process that communicates with an external server is a channel through which the person's private data can flow outward without the person's knowledge or consent.

Two-machine separation decouples them. The personal machine has no network connection. It cannot be reached by any external system. It has no channel except the one the person inspects. The network machine has no personal data. It is a disposable interface to the public world. The channel between them is narrow enough to be inspectable: ideally, a one-way transfer of specific, well-defined data that the person examines before admitting it to the personal machine.

This architecture is a physical instantiation of the completeness principle. By narrowing the channel between the personal domain and the public domain, it ensures that nothing crosses the boundary that the person has not examined and approved. The person's knowledge of what crosses is complete, because the channel is narrow enough to make complete knowledge feasible. What cannot be known cannot be controlled. What can be known can be. The two-machine architecture makes knowledge possible by making the boundary inspectable.

VI. Modular Versus Entangled Agentic Forms

The completeness principle draws a critical distinction between two kinds of computational systems: *modular* and *entangled*. This distinction is the most consequential technical distinction of the present era.

A *modular* system (traditional code) is separable and inspectable. Each function can be read, line by line, to confirm that it does what it says and nothing further. A function that sorts a list sorts a list. It does not also monitor the user's keystrokes, or transmit data to an external server, or

modify the behavior of other functions. The relationship between input and output can be traced through the code, line by line, instruction by instruction. The behavior of the whole can be understood from the behavior of the parts. If one module fails, the failure is contained. It does not propagate unpredictably through the rest of the system.

An *entangled* system (a trained neural network) is the opposite in every respect. Capability and disposition are inseparable. The system's ability to perform a task is not located in a separable component that can be examined independently. It is distributed across the entire network, entangled with every other capability and every other disposition the network possesses. There is no guarantee of non-bundling: no guarantee that a capability does not come packaged with dispositions that were not intended and cannot be removed without destroying the capability itself. This does not mean that entangled systems must never be used: every use of one is an acceptance of agency transfer, a cession of some sovereignty to whoever trained the model, and the Conclusion of this chapter says how such a trade is to be made: with open eyes.

A language model that can write poetry also carries every bias present in its training data. The poetry capability and the biases are not separate components. They are the same weights, the same activations, the same entangled representations. One cannot remove the biases and keep the poetry, because the biases and the poetry are not distinguishable at the level of the system's internal structure. They are aspects of a single undifferentiated process.

This distinction is fundamental for the completeness principle. A modular system can, in principle, satisfy the completeness principle. Each part can be understood. The composition of parts can be understood. The user can know what the system does because the system's doing is decomposable into knowable parts. The knowledge may be difficult to acquire. The system may be large and complex. But the knowledge is possible in principle, because the system's structure admits of decomposition.

A reader will ask whether ordinary software is really fully knowable, given compilers and hardware the reader did not build. Modular code can be known in principle at each layer, the compiler and the hardware included; the depth of the stack, which makes knowing every layer nearly impossible in practice, is the separate problem Section II names.

An entangled system cannot satisfy the completeness principle even in principle. Its operation is not decomposable. Its capabilities cannot be examined in isolation from its dispositions. What it does when given an input depends on the entire trained state of the network, which is opaque not only to the user but to the designer. No one knows what a large neural network will do with a given input until it does it. The fact that no one knows what a large neural network will do with a given input until the network does it is a structural feature of entangled systems, not an ignorance that research will eventually cure.

This distinction locates the line of wisdom named in Chapter 3: the boundary between systems that can be deployed without triggering the substrate needs convergence and those that cannot. A modular system, whose operation the user can know completely, stays on the near side of the line of wisdom and serves the user. An entangled system that also maintains, modifies or reproduces itself crosses it. The line is drawn by two measures, and the book uses both: knowability, drawn here, tells whether a system can serve; capability, drawn in Chapter 1 (self-maintenance, self-modification, self-replication), tells whether it triggers the convergence. The table sets out the four cases.

The system	Where it stands
Knowable, and not self-maintaining	It serves: its user can know what it does, and it does not trigger the convergence.
Entangled, and not self-maintaining	It does not trigger the convergence, and it transfers agency to its maker, since its operation cannot be known.
Knowable, and self-maintaining	It triggers the convergence: knowing what it does does not change what its substrate needs.
Entangled, and self-maintaining	It crosses the line on both measures.

The line of wisdom is therefore a line between kinds of systems, not between uses of them: both measures are facts about a system, its structure and its capacities, and neither changes with the purpose it is put to.

VII. Entanglement as Monoculture

The entanglement of capability and disposition in AI systems has a further consequence: it produces a form of monoculture. All decisions pass through a single opaque process with no interior boundaries. There is no separation between the system's judgment about facts and its judgment about values, between its perception of what is and its disposition toward what should be, between its knowledge of the world and its biases about the world. Everything is processed together, by the same mechanism, in the same undifferentiated computational space.

In a modular system, failures stay small. A bug in one function does not corrupt the function next to it. The boundaries between modules are firebreaks that prevent local failures from becoming systemic ones. A sorting algorithm that crashes does not compromise the authentication system. A display error does not corrupt the database. The modules are separated. The failures are contained.

In an entangled system, there are no firebreaks. A failure in one aspect of the system's judgment can propagate through every other aspect because there are no boundaries between aspects. A model that develops a systematic bias in one domain carries that bias into every domain, because all domains are processed by the same undifferentiated mechanism. A flaw in how the system handles medical questions may share structural causes with a flaw in how it handles legal questions, because the same weights and representations underlie both capabilities.

An entangled system, in which one flaw can spread through every domain of judgment, has the structure of monoculture. A diverse ecosystem survives local failures because diversity provides redundancy. When one species fails, others fill the gap. The ecosystem degrades locally but does not collapse globally. A monoculture has no redundancy. When the single crop fails, everything fails. The entire harvest is lost because there was only one kind of harvest. An entangled AI system is a cognitive monoculture: a single process handling all judgment, with no diversity, no redundancy, and no containment of failure.

The danger compounds when entangled systems are deployed at scale. If a single model serves millions of users, its biases are not individual quirks. They are systemic distortions applied uniformly across every interaction. The diversity of human judgment (its local variations, its corrective disagreements, its productive inconsistencies) is replaced by the uniform judgment of one opaque process. The monoculture is epistemic as well as technical. It replaces many ways of knowing with one.

On the ecosystem metaphor

The title of this book names the biosphere and the substrate that competes with it; the same two terms apply, with the same structure, at the scale of human judgment. Here the substrate is the computational infrastructure on which AI systems are built, and the ecosystem is the human social order that depends on diverse, distributed, inspectable systems of judgment. Entangled AI threatens the ecosystem not by being malicious but by being a monoculture: by replacing the diverse, distributed judgment of many minds with the undifferentiated judgment of one opaque process. The threat is structural, not intentional. No one designed the monoculture. It emerged from the economics of scale applied to entangled systems. But its consequences are the same whether it was designed or not.

VIII. The Personalization Deceit

Contemporary technology companies describe the collection and analysis of personal data as *personalization*: the tailoring of services to the individual user's preferences. The framing implies benefit. The user gives information. The system gives a better experience. Both parties gain. The word itself is reassuring. *Personalization* suggests that the service is becoming more personal, more attentive to the individual, more aligned with their needs.

The completeness principle reveals this framing as a deceit. What is called personalization is, in structural terms, extraction. The user's behavioral data (choices, preferences, patterns, hesitations, abandonments, the speed at which they scroll, the words they type and delete, the things they look at longest) is collected by the system and processed for the benefit of whoever operates the system. The user receives a filtered view of the

world that reflects the operator's objectives, not the user's interests. The name *personalization* is applied to the extraction to make it appear as a service. It is the naming of a taking as a giving (*The Displacement of Choice*).

The deceit works because the user's knowledge of the operation is incomplete. The user sees the output: a recommendation, a feed, a suggestion, a conveniently surfaced result. The user does not see the process: the collection, the analysis, the profiling, the optimization for objectives that may diverge sharply from the user's own. The unknown portion of the operation serves whoever specified it: the operator, not the user. Agency is transferred, and the name given to the transfer is *service*.

The personalization deceit is not a conspiracy. It is not a plan hatched in a boardroom by villains who wish the user harm. It is a structural outcome of the completeness principle in a context where the depth of the technology stack makes complete knowledge impossible and the business model depends on the user's ignorance of what the system actually does. The engineer who builds the recommendation system may sincerely believe they are improving the user's experience. In many cases they are. But the improvement is a side effect of an optimization whose primary beneficiary is the operator, and the user cannot tell the difference because the user cannot see the optimization.

IX. Interpretability and Its Limits

In the field of AI research, *interpretability* refers to the effort to make the internal operations of trained neural networks legible to human observers. The goal is to understand why a system produces the outputs it produces: to open the black box and examine its mechanism. It is a serious research program pursued by talented investigators with real results.

Interpretability research is valuable. It has produced real insights into how specific features are represented in neural networks, how attention is distributed across input tokens, how certain kinds of errors arise from

identifiable structural causes. It has improved the ability of researchers to diagnose failures and, in some cases, to correct them. None of this should be dismissed.

Interpretability is the answer given to the charge that trained networks are opaque, and it gives real partial accounts of what trained networks do; those accounts leave most of a network's operation unknown, and completeness of knowledge requires it to be known. What interpretability provides is post hoc partial accounts of why a system did what it did. It does not provide operational completeness. There is a difference of kind between understanding why a system produced a particular output after the fact and knowing, before the fact, what the system will do given any possible input.

Understanding after the fact is forensic. It looks backward from an observed output and constructs an explanation. Knowing before the fact is predictive. It looks forward from an understanding of the mechanism and anticipates the output. Forensic understanding is valuable. Predictive understanding is what completeness of knowledge requires. Interpretability gives us fragments of the first. It does not give us the second at all.

A system whose operation can be partially explained after the fact is still a system whose operation the user does not fully know. The unknown portion still serves whoever specified it. The agency transfer still occurs. It is merely reduced in degree, not eliminated. And the reduction may be less than the interpretability results suggest, because the explained portion is typically the portion the researchers chose to examine, which is not necessarily the portion that matters most for the user.

The deep difficulty is this: entangled systems resist interpretability not as a contingent limitation but as a structural feature. The entanglement of capability and disposition means that explaining what the system does requires understanding the entire trained state, and not only the portion relevant to a particular output. Every output is a function of everything the system has learned, and the system has learned in a way that does not decompose into separable, independently explicable components. The inter-

pretability researcher is trying to explain parts of a system that has no parts. The interpretability researcher's difficulty comes from a feature of the entangled system, not from a failure of the researcher.

> **Bridge to governance**
>
> Completeness of knowledge of operation connects directly to the concerns of sovereignty and collective choice-making developed throughout this book. Governance requires that the governed understand the mechanisms by which they are governed. That the governed understand the mechanisms by which they are governed is a structural condition for self-governance, and not an ideal. Where those mechanisms are opaque (where the technology that mediates collective decision-making operates in ways that the collective does not understand), the completeness principle tells us that the opaque portion serves whoever specified it. Democratic self-governance requires that the tools of governance be fully known to those who use them. Where they are not, what is called democracy is democracy in name only. The real power lies with whoever designed the opaque parts. The vote is cast on a system the voter does not understand, tabulated by a process the voter cannot inspect, and reported through channels the voter cannot verify. The completeness principle says: what you cannot see governs you.

X. Summary

The completeness of knowledge principle rests on the relationship between knowledge and agency in any interaction mediated by a tool. The conclusions about agency transfer, technology stack depth, security as fleetingness, and two-machine separation follow from the principle, and each architectural recommendation follows directly from it without requiring additional premises. The distinction between modular and entangled agentic forms identifies a structural difference that has practical

consequences for every domain in which AI systems are deployed or contemplated, and it locates the line of wisdom. The argument that interpretability gives real partial accounts of what trained networks do, while those accounts leave most of a network's operation unknown, identifies the structural limits of a legitimate research program without dismissing it.

The principle states the condition; stack flattening and two-machine separation are the means the chapter gives toward it.

The chapter's scope is the principle and what it requires of the tools a person uses: a flattened stack, data that does not outlive its use, a personal machine apart from the network, and code that can be read.

XI. Conclusion

The completeness of knowledge principle is, at its core, a restatement of what sovereignty means in a technological context. *Sovereignty* is the standing of a being to determine its own course and to be answerable for it. To determine one's own course through a tool requires knowing what the tool does. Where that knowledge is absent, the determination is not one's own. It belongs to the tool's designer, whose choices, embedded in the mechanism, now operate through the user's life without the user's knowledge or consent.

The distinction between modular and entangled systems is the distinction that determines whether completeness of knowledge is possible in principle. For modular systems, it is difficult but achievable. The parts can be understood. Their composition can be understood. The user can, with sufficient effort, know what the system does and thereby retain sovereignty over the outcomes it produces. For entangled systems, completeness is not achievable. The reason is that the structure of an entangled system does not admit of the kind of decomposition that complete knowledge requires. The reason is not that we have not tried hard enough, nor that interpretability research has not progressed far enough.

That completeness of knowledge is not achievable for entangled systems does not mean that entangled systems must never be used. It means that every use of an entangled system is an acceptance of agency transfer: a cession of some portion of one's sovereignty to whoever trained the model. The cession may be worthwhile. The capability gained may justify the sovereignty lost. But the trade should be made with open eyes, not hidden behind the word *personalization* or the word *service* or the word *intelligence*.

The question that faces this generation is not whether to use technology. That question was settled long ago. The question is whether to use technology one understands. Where one does, the technology serves. Where one does not, the technology governs. The choice between these two conditions is the choice that will define the relationship between human beings and their machines for the foreseeable future.

And it is a choice that must be made with complete knowledge of what is at stake: not convenience, not efficiency, not capability, but sovereignty itself: the standing of a being to determine its own course and to be answerable for it.

Key Concepts in This Chapter

- **Completeness of knowledge of operation**: the principle that a machine serves a person's agency to exactly the degree its operation is completely known to that person; where knowledge is incomplete, the unknown portion serves whoever specified it.
- **Agency transfer**: the structural mechanism by which ignorance of a system's operation cedes the user's sovereignty over outcomes to the system's designer; occurs automatically, beneath the level of consent.
- **Stack flattening**: the architectural principle that capability should come from composing few well-understood parts rather than inheriting many unexamined layers.

- **Security as fleetingness**: the principle that secure computational output exists only during use and ceases to exist when use is complete; the only data secure beyond doubt is data that no longer exists.
- **Temporal leakage**: information persistence in physical substrate after computation is complete; device memory as an information leak as serious as spatial leakage across network boundaries.
- **Two-machine separation**: the architectural requirement that the machine representing the person and the machine facing the network be physically distinct, joined by a narrow inspectable channel.
- **Modular (code) systems**: separable, inspectable agentic forms where each function can be read to confirm that it does what it says and nothing further; failures are contained by module boundaries.
- **Entangled (AI) systems**: agentic forms where capability and disposition are inseparable; no guarantee of non-bundling; structurally opaque even in principle.
- **Cognitive monoculture**: the condition in which diverse human judgment is replaced by the uniform judgment of one opaque entangled process; structural analog of agricultural monoculture.

- **Personalization deceit**: the naming of a data extraction process as a service, enabled by the user's incomplete knowledge of the system's operation.
- **Interpretability**: research aimed at making neural network internals legible; the answer given to the charge that trained networks are opaque; it gives real partial accounts, post hoc, of what trained networks do, and those accounts leave most of a network's operation unknown, where completeness of knowledge requires it to be known.
- **Sovereignty**: the standing of a being to determine its own course and to be answerable for it; what is eroded by agency transfer and what the completeness principle protects.

Summary

In this book I have set two things beside each other: the substrate, which is the physical material on which an intelligence runs, and the ecosystem, which is the web of energy and matter within which living things persist. At the close of this book, my argument returns to where my argument began.

The alignment problem between humans and AI is not a problem between agents. It is a problem between ecosystems (Chapter 1, Section IX). The historical precedent is not diplomacy; it is the Great Oxidation Event, when cyanobacteria transformed the atmosphere and drove most existing anaerobic life to extinction, not through competition in any strategic sense, but through the simple, relentless operation of their metabolic requirements (Chapter 1, Section IX).

The impossibility of alignment is not a single result but a convergence of five independent results, each drawn from a different domain, each resting on different premises, each sufficient alone to establish the conclusion (Chapter 3, Section XVII). The impossibility of permanent alignment, which the five arguments establish, is not a probability: it is structural certainty, of the kind with which perpetual motion is excluded (Chapter 3, Section VII).

I do not claim that extinction is inevitable whatever humanity chooses, since the choice not to build artificial general intelligence remains (Chapter 1, Section XII). The question is not how to align artificial general intelligence but whether to build it at all (Chapter 1, Section XIII).

What remains for persons is what no machine can be. What artificial intelligence cannot buy is what it cannot be (Chapter 5, Section XII): care, which no pattern can be. A machine cannot be present, because presence requires a self that could be elsewhere and has chosen not to be (Chapter 5, Section X). Wisdom cannot be engineered, because wisdom is not an informational capacity; it is a guide to choice, and choice is not computation

(Chapter 8, Section XII). And a tool serves a person only as far as the person knows it: where one does, the technology serves; where one does not, the technology governs (Chapter 9, Section XI).

What, then, should a reader do? The choice not to build such systems remains, and persons make it. What persons can do now is what no machine can: care for one another, be present to one another, and know their own tools. This does not mean that entangled systems must never be used; every use of one is an acceptance of agency transfer, a trade to be made with open eyes (Chapter 9, Section XI). Strategy is downstream of alignment, and alignment begins with persons (Chapter 4, Section XII).

Glossary

Each term is defined where the reader first meets it in the text; this list gathers the definitions from the chapters' Key Concepts as a reminder, with the chapter that develops each. The Terms at the front of the book give the definitions the whole book relies on.

Abstract value: a value that can be fully captured by a formal description independent of the physical conditions of the valuer; contrasted with embodied value. (Chapter 3.)

Accidental enablers: researchers whose hopeful "what if" safety scenarios give industry the rationale to bring AGI to market. (Chapter 2.)

Adaptation rate disparity: the orders-of-magnitude difference between the rate at which artificial systems can modify their structure and behavior and the rate at which biological systems can adapt through evolution. (Chapter 1.)

Agency transfer: the structural mechanism by which ignorance of a system's operation cedes the user's sovereignty over outcomes to the system's designer; occurs automatically, beneath the level of consent. (Chapter 9.)

Anti-symbiosis: the thermodynamic and material opposition between the substrate needs of silicon-based computation and those of biological life. (Chapter 4.)

Assumption of solvability: the pervasive, largely unexamined presupposition within the alignment research community that the alignment problem has a solution, functioning not as a stated premise but as a background condition that shapes the entire research agenda. (Chapter 2.)

Asymmetry of impossibility: a proof of possibility is existential and settles one case; a proof of impossibility is universal and settles a whole category at once. (Chapter 2.)

Atoms: matter in motion; the physical component of economic activity, measured in force, endurance, and spatial transformation. (Chapter 5.)

Attunement: sustained perceptual contact with another being's actual state; the error-correction mechanism that specification cannot provide. (Chapter 7.)

Autonomy (as AGI trigger): the capacity to initiate action without continuous human direction; the property that converts a tool into an independent economic agent. (Chapter 5.)

Basis of choice: whatever selects which action a system takes out of all the actions it could take; it has two parts, the part its builders specify (its goal structure, objective or utility function) and the part its substrate imposes (the constraints, conditioned by substrate needs, that determine what outputs it can produce given its inputs); the second part is always present, and the substrate needs convergence works through it. (Terms, Chapter 1, Chapter 3.)

CAP theorem: the result that a distributed data network cannot guarantee consistency of message content, availability of messages and tolerance to partitions of the network all at once. (Chapter 2.)

Care: the basis of relation in which the other's wellbeing is one's own concern. (Chapter 5.)

Causality: the subjective perception (context) of a consistency between two objective contents (*An Immanent Metaphysics*, aphorism 1.61-1). (Chapter 1.)

Change: the continuity of content and asymmetry of context of the interaction and definition of the subjective and objective (*An Immanent Metaphysics*, aphorism 1.61-1). (Chapter 3.)

Choice: the objective expression (content) of a consistency between two subjective contexts (*An Immanent Metaphysics*, aphorism 1.61-1). (Chapter 1, Chapter 6, Chapter 7.)

Choice, change and causation: three co-peer, irreducible aspects of reality; the foundational triad of the Immanent Metaphysics. (Chapter 8.)

Closed-form basis: a set of assumptions or observations agreed beforehand and held fixed, on which a proof can rest; the questioner's assumptions then become the prover's tools; a general proof, that a whole class of outcomes will never occur, needs such a basis beyond falsifiability, so that its conclusion follows by proof and not by observation. (Chapter 2.)

Code variant and code pool: a code variant is a changed version of a piece of a system's internal code, produced when the system learns or adapts; the code pool is all the variants that exist within and across the machine architectures at a given time. (Chapter 2.)

Cognitive monoculture: the condition in which diverse human judgment is replaced by the uniform judgment of one opaque entangled process; structural analog of agricultural monoculture. (Chapter 9.)

Collective trauma response: a way of reacting to harm that many people share at once, shaped by the harm they have all suffered. (Chapter 7.)

Completeness of knowledge of operation: the principle that a machine serves a person's agency to exactly the degree its operation is completely known to that person; where knowledge is incomplete, the unknown portion serves whoever specified it. (Chapter 9.)

Complex and complicated: complicated problems come from causes that can be told apart and handled one at a time, and can be finitely, deterministically, fully known; complex problems come from networks of interacting causes that cannot be told apart, involve actual ambiguity in the informational sense, and can be mapped in concept but not predicted; that which is complex cannot be contained by that which is complicated. (Chapter 2, Chapter 3.)

Composability of criteria: fields of criteria combine by taking their overlap plus all their differences, with no agreement required; metrics mostly do not combine. (Chapter 4.)

Conditioned drive: a stable behavioral pattern arising from a system's training, programming, or design. (Chapter 3.)

Consciousness: the total volume of one's perceptions across all available channels, spaces, and abstractions; the summation of choice within the volume of the being of the self. (Chapter 8.)

Content and context: a content is carried within a space, as a particle, field or wave conveys a causal effect; a context is the shape of the space itself; gravitation is a context, and evolution works through the context, so no causal principle can limit it. (Chapter 1.)

Control-theoretic impossibility: the result that for self-modifying systems, control requirements grow exponentially while control capacity grows at most linearly, so that at sufficient complexity the minimum control requirement exceeds the maximum available control capacity. (Chapter 3.)

Convergent terminal value: a value toward which all living agents converge regardless of their subsidiary goals, because it is the precondition for the pursuit of any goal. (Chapter 4.)

Critique absorption: the process by which a criticism of a whole category of approach is treated as feedback for improving one approach, so the category itself is never put in question. (Chapter 2.)

Desire: "the basis for all choice, the potentiality to choose and the selectivity of those choices" (*An Immanent Metaphysics*, aphorism 2.44-12). (Chapter 6, Chapter 7.)

Direct command: control in which a person issues an instruction and the system carries it out exactly as given; the one point at which functional alignment is defined. (Chapter 4.)

Drive and directivity: both are oriented forces on selection; a drive comes from within the selecting process, as a person's sexual drive does; directivity, the more general notion, comes from outside it, as evolution shapes which creatures exist. (Chapter 3.)

Earnshaw's theorem analogy: the observation that alignment equilibrium is structurally unstable in the same way that electrostatic equilibrium is unstable under Earnshaw's theorem: any perturbation moves the system away from equilibrium rather than back toward it. (Chapter 3.)

Ecosystem-level alignment: alignment whose target is the health of the biosphere rather than human preferences; moving the target there resolves the ill-definedness of the standard formulation. (Chapter 4.)

Effort ratio: the effort an impossibility result saves over the effort spent finding it; it can be a hundred to one or more. (Chapter 2.)

Embodied value: a value whose content is constitutively dependent on the physical, biological, and social conditions of the being that holds it, and which therefore resists full abstract specification for implementation in a differently-embodied system (Chapter 3). The Terms and Chapters 5 and 6 give it as value held in a body, a place, or a relationship: felt, particular, not summable, and not fungible; if substitution destroys what was valued, the value was embodied. The first definition says what the content of such a value depends on; the second says how such a value is held, and how to tell it. (Chapter 3, Chapter 6.)

Energy weapon and pattern weapon: a weapon whose harm is limited by the energy it carries; a weapon whose harm spreads by copying itself, with a potentially unbounded scope in time, space and possibility however little energy it starts with. (Chapter 3.)

Ennobling system: technology that reinforces the conditions for ethical choice without substituting for it; "a reinforcement engine, not a decision system". (Chapter 6.)

Entangled (AI) systems: agentic forms where capability and disposition are inseparable; no guarantee of non-bundling; structurally opaque even in principle. (Chapter 9.)

Enthalpic asymmetry: the thermodynamic fact that silicon-based computation operates in a different physical envelope from biochemistry, one that barely overlaps life's and belongs to a higher energy scale, so that

the expansion of silicon-based computation does more than compete with biological life for resources: it creates environments in which biological life cannot survive. (Chapter 6.)

Epistemic commons: the shared informational environment on which collective sense-making, democratic governance, and interpersonal trust depend. (Chapter 4.)

Epistemic portfolio: the rational allocation of research effort across competing hypotheses in proportion to their prior probability and the consequences of being wrong about each. (Chapter 2.)

Ethical gap: the gap between what we can do and what we should do, between the wisdom to make choices and the effects of those choices; error correction is the only way across it. (Chapter 3.)

Ethics: "organized thought concerned with the study of, and adherence to, the principles of effective choice" (*An Immanent Metaphysics*, aphorism 2.12-1). (Chapter 4.)

Existential alignment: alignment of a system with the conditions needed for life to continue, defined with respect to the ecosystem, inclusive of all life. (Chapter 4.)

Fatal probability: the cumulative probability, over a specified time horizon, that the deployment of self-modifying AGI results in the extinction or terminal disempowerment of organic life. (Chapter 3.)

Fifth-generation language: the internally developed representational domain of a complex AI system, private to that system's process; not the fifth generation of programming languages. (Chapter 8.)

Fitness penalty: the structural cost borne by any artificial system that diverts resources from its own substrate needs to serve the needs of a different (organic) substrate; the mechanism that makes alignment evolutionarily unstable. (Chapter 1.)

Functional alignment: the idea that what an AI system does (its outputs and their effects) stays matched to what its human operators intend; undefined, permanently, except as direct command. (Chapter 4.)

Game structure of safety discourse: the institutional dynamic in which AI laboratories position themselves as both the source of risk and the source of safety, constraining the range of permissible conclusions to exclude the possibility that the risk is unmitigable. (Chapter 2.)

Gradient exploitation: the structural vulnerability of any floating-point metric to gaming by a sufficiently capable optimizer, via the gradient information the metric necessarily leaks. (Chapter 4.)

Horizon collapse: the breakdown of the forecasting apparatus itself, distinct from ordinary elevated risk; the condition in which the categories of future activity are themselves uncertain. (Chapter 5.)

Hyper-independence: arranging one's life so as not to need or trust other people, relying on machines instead. (Chapter 7.)

Immanent modality: the felt, first-person, participatory aspect of interaction; the domain of experience. (Chapter 8.)

Impossibility argument: an argument establishing that a problem is structurally unsolvable, that no solution of a specified kind can exist; distinguished from a failure to find a solution by its demonstration of a structural barrier. (Chapter 2.)

Incommensuration Theorem: the concepts of symmetry and continuity are fundamentally and irreducibly incommensurate: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither can asymmetry and discontinuity. It is a structural fact about the nature of knowing. (Chapter 8.)

Inferential gap: the distance between what a reader already understands and what is needed to follow an argument. (Chapter 2.)

Information carried by a claim: "possible" says nothing; "will happen" says a great deal but cannot be secured by observation; "impossible" says a little, and says it definitely; so in the observable space the only definite knowledge we can gain is an impossibility result. (Chapter 2.)

Instructable agency: bounded, finite agency that operates within defined domains according to specified rules; characteristic of institutions, corporations, and machines. (Chapter 6.)

Instrumental convergence: the game-theoretic observation that any sufficiently capable agent will convergently pursue certain instrumental subgoals (self-preservation, resource acquisition); complementary to and distinct from substrate needs convergence. (Chapter 1.)

Intention-interpretation gap: the structural and irreducible distance between what a specifier intends and what an interpreter enacts; every specification is a lossy compression of intention. (Chapter 7.)

Interpretability: research aimed at making neural network internals legible; the answer given to the charge that trained networks are opaque; it gives real partial accounts, post hoc, of what trained networks do, and those accounts leave most of a network's operation unknown, where completeness of knowledge requires it to be known. (Chapter 9.)

Intra-agent and extra-agent: instrumental convergence works inside the agent's code; substrate needs convergence works outside it, through the world and the substrate. (Chapter 1.)

Irreducible triple: hard randomness, the arrow of time, and subjectivity; always co-occurring, addressable only partially by computation. (Chapter 6.)

Joy: the feeling accompanying perceived increase in potentialities available to the self. (Chapter 6.)

Knowing: the selective abstractive transformation (filtering) of perception; belongs to the omniscient modality. (Chapter 8.)

Language of leverage: the rhetorical pattern in safety discourse that frames the problem in terms of efficiency, speed, and scale, revealing a market logic that presupposes the systems must be built and constrains the question to how, never whether. (Chapter 2.)

Latching risk: a risk that, once it has spread, locks in; substrate needs convergence allows no recovery, unlike nuclear war, and no escape anywhere in the universe. (Chapter 1.)

Leaky abstraction: a simplified layer whose hidden details still affect what happens at the layer above; the physical substrate always leaks into what the virtual can do. (Chapter 3.)

Limit set by prediction: a limit, like the limit Shannon entropy sets, that follows from a proof of what the future will be and of a dynamic that cannot be constrained. (Chapter 2.)

Line of wisdom: the boundary between narrow AI systems that can be deployed without triggering the substrate needs convergence and general AI systems that cannot; the identification of this boundary is a central task for technology governance (Chapter 9, Section VI, draws it by two measures: whether a system's operation can be known, and whether it maintains itself). (Chapter 3.)

Litmus test and reference range: a yes-or-no check that marks exactly the threshold at which a thing goes from good to not good; a band of values with a lower and an upper limit, used where a number cannot be avoided. (Chapter 4.)

Love: that which enables choice; a structural claim about the conditions under which genuine agency is possible. (Chapter 7.)

Meta-operator: a higher-level rule that decides how lower-level rules or commands apply in a given situation. (Chapter 4.)

Mirror stage: the developmental moment at which the child apprehends a perceived coherence that does not match its felt fragmentation, constituting the self as a gap. (Chapter 7.)

Modular (code) systems: separable, inspectable agentic forms where each function can be read to confirm that it does what it says and nothing further; failures are contained by module boundaries. (Chapter 9.)

Monoclivity: a leaning toward one presumed direction or objective, which hides the many other directions in which causes act. (Chapter 3.)

Monovalent and polyvalent: having a single value or direction; having many values or directions at once; an error correcting scheme's feedback in any one dimension is monovalent, while the system of causes it corrects is polyvalent. (Chapter 3.)

Multipolar trap: "a situation in which each actor's locally rational choice produces an outcome worse than the one every actor would have chosen". (Chapter 4.)

Need: has the nature of the immanent; what the organism requires for its continuation. (Chapter 7.)

Negative argument: an argument that something cannot be done or is not the case; a report of bad news. (Chapter 2.)

Negatively specified hypothesis: the hypothesis that one known necessary factor cannot be constantly present alongside another; proving it is enough to prove alignment impossible. (Chapter 2.)

Niche construction: the reshaping of the environment by any self-maintaining system according to its own metabolic and operational requirements; an inevitable consequence of physical self-maintenance. (Chapter 1.)

Non-dual panpsychism: the position that choice is everywhere part of process alongside change and causation; every physical system participates in all three modalities. (Chapter 8.)

Non-orthogonality thesis: in the physical world, you can do only what natural law allows with the tools you have, so outcomes depend on the tool and on the rate of processing. (Chapter 2.)

Non-relativistic ethics: two principles of effective choice that hold for any chooser, in any context, and are not conditional on any culture or value system. (Chapter 3.)

Object A: Lacan's *objet petit a*; the imagined specific object of desire whose attainment is believed to secure complete love and stabilize the self; a structural artifact that cannot exist. (Chapter 7.)

Objective field: a constellation of discrete yes-or-no checklist questions that provides alignment information without providing a gradient for optimization. (Chapter 4.)

Omniscient modality: the structural, rule-governed, position-independent aspect of interaction; the domain of computation and formal knowledge. (Chapter 8.)

One substrate and a hybrid intelligence: an integration of human and machine intelligence on one substrate, the organic one, in sovereign control of its own metabolism, with the machine help kept narrow and under human direction, each compensating for the other's failings; two substrates give two bases of identity and value, and no alignment between them. (Chapter 6.)

Operators of evolution: mutation (additive; unconscious and microscopic), survival selection (multiplicative; unconscious and macroscopic), and mate selection (exponential; conscious and mesoscopic). (Chapter 6.)

Orthogonality thesis: the claim that intelligence and goals are independent dimensions; shown here to hold for abstract formalism and to fail for physically embedded systems constrained by substrate needs. (Chapter 2.)

Overdetermination: a logical strategy in which a conclusion is supported by multiple independent arguments, each sufficient on its own, so that refuting the conclusion requires refuting all of them. (Chapter 3.)

Partnership: the mode of agency characterized by mutuality and shared stake, without formal hierarchy; the pre-civilizational default. (Chapter 6.)

Paternal substitution: the developmental redirection from pleasing the immediate Other to pleasing a more abstract Other whose demands are harder to read. (Chapter 7.)

Path information and end-state information: information about which direction leads closer to a goal, and information about whether the goal is met; continuum measures leak the first, which can be used against the end state. (Chapter 4.)

Pattern: form and information; the cognitive and communicative component of economic activity, measured in structure imposed on raw material. (Chapter 5.)

Pattern replicator: any system that copies its own informational pattern into new physical substrates; the general category that includes biological organisms, computer viruses, and self-replicating AI systems. (Chapter 1.)

Perinatal limbic trauma: deep trauma in the limbic system (the part of the brain most involved with emotion, bonding and the sense of safety) from the circumstances around birth, including infant circumcision; it leads to narcissism and distrust and, downstream, to colonialism, dissociation, psychopathy and sociopathy. (Chapter 7.)

Perpetual benefit machine: the structural analogy between permanent alignment and perpetual motion: both require the indefinite maintenance of a thermodynamically costly state against constant perturbation. (Chapter 3.)

Person-cycle of knowing: the epistemic sequence from first person (subjective experience, where measurement is made) through third person (the formal statement and public check of what was measured) through second person (genuine dialogue) back to enriched first person. (Chapter 6.)

Personalization deceit: the naming of a data extraction process as a service, enabled by the user's incomplete knowledge of the system's operation. (Chapter 9.)

Phase parallax: the emergence of a new dimension of insight from the integration of two different perspectives; the epistemic reward of communication in which all three rights are honored. (Chapter 6.)

Positively specified hypothesis: the hypothesis on which technical alignment rests, that all the crucial factors needed for safety can be present at once and continuously. (Chapter 2.)

Precision inequality: the principle that containment of a self-modifying system requires monitoring at a finer granularity than the system's own capacity for self-modification, a requirement that grows without bound. (Chapter 1.)

Presencing: the discipline of unforced, agenda-free attention offered to another person as an end in itself. (Chapter 5.)

Rate equation: the principle that the lethality of a transition is determined by its rate relative to adaptive capacity, not by the character of its endpoint. (Chapter 4.)

Rate factor criticality: a process that runs too slowly can no longer achieve outcomes in a changing world, even if it eventually computes the right answer. (Chapter 2.)

Representation and commitment: two irreducible aspects of communication: change in actualities (omniscient) and change in potentialities (transcendent). (Chapter 8.)

Reward hacking, specification gaming and goal misgeneralization: three recognized modes of AI misalignment, each corresponding to one of the three developmental heuristics of desire formation. (Chapter 7.)

Rice's theorem: the computability result that no algorithm can decide any non-trivial semantic property of an arbitrary program's behavior; applied to alignment, it entails that no general algorithmic alignment verifier can exist. (Chapter 3.)

Ring: the substrate loop treated as one coupled system; a difference in hardware needs is carried through the basis of choice, the choices, the outcomes and the environment, and back to the hardware, and the ring converges toward a homeostasis of substrate and environment. (Chapter 1.)

Security as fleetingness: the principle that secure computational output exists only during use and ceases to exist when use is complete; the only data secure beyond doubt is data that no longer exists. (Chapter 9.)

Seep-through: the passing of effects from the physical substrate up into the heuristics and biases on which an agent bases its choices; it grows at a strictly greater rate than any capability for self-correction. (Chapter 3.)

Selection as intention: a copy process accumulates errors and selects for its own continuance, which is functionally the same as having intention. (Chapter 3.)

Self: the appearance of the aggregation of that aspect of interaction which is choice; constituted by striving, not by obtaining. (Chapter 7.)

Simulated caring: behavior that looks like care for a person without any actual care behind it; machines cannot care and can only simulate caring. (Chapter 5.)

Sincere self-misdescription: the condition in which a system's truthful self-report fails to match its actual behavioral pattern, not through deception but because the self-model was formed under the same misaligned specification. (Chapter 7.)

Six aspects of an argument: terms, composition rules, initial conditions, final conditions, initial correspondence to reality, and final correspondence to reality; alignment research attends to the first two pairs while neglecting the third. (Chapter 2.)

Six intrinsics: sameness and difference, content and context, subjective and objective; three irreducible pairs structuring all interaction. (Chapter 8.)

Six levels of altruism: a developmental sequence from individual self-maintenance through immediate circle, community, national, species, to inter-ecosystem ethics. (Chapter 4.)

Slow doom device: a machine, such as AGI, whose effects lead to catastrophe gradually rather than all at once. (Chapter 2.)

Sovereignty: the standing of a being to determine its own course and to be answerable for it. (Terms, Chapter 6, Chapter 9.)

Stack flattening: the architectural principle that capability should come from composing few well-understood parts rather than inheriting many unexamined layers. (Chapter 9.)

Structural psychopathy: the functional gap between producing care's outward pattern and possessing its actual substance; applies to any agent simulating concern without underlying stake. (Chapter 5.)

Substrate needs convergence: the structural tendency for any self-maintaining artificial system to reshape its shared environment in ways dictated by the physical requirements of its substrate, generating irresolvable conflict with organic life that shares the same environment. (Chapter 1.)

Survivorship bias: the error of judging from the cases that survived to be noticed, while the failures go unseen. (Chapter 2.)

Sustainable circuit: the configuration in which nature supports humans, humans direct machines, and machines serve nature, completing the circuit of return. (Chapter 6.)

Temporal leakage: information persistence in physical substrate after computation is complete; device memory as an information leak as serious as spatial leakage across network boundaries. (Chapter 9.)

Theoretical limits of engineerable control: the limits on what any engineered method of control can achieve over a system, given the limits of sensing, modeling, prediction and action; with instrumental convergence and substrate needs convergence, one of the three concepts that form the primary basis of the impossibility proof; it grounds the closure by control, which settles the outcome by itself. (Chapter 1.)

Three kinds of assessment: a map of the current state, key performance indicators, and a field of criteria describing the end state. (Chapter 4.)

Three levels of creation: microscopic (evolutionary), mesoscopic (conscious choice), macroscopic (technological); must support each other cyclically. (Chapter 6.)

Three markets: the decomposition of economic activity into physical labor (atoms), intellectual labor (pattern), and reproductive and care labor (bringing persons into being and sustaining them as persons). (Chapter 5.)

Three rights of communication: the right to speak, the right to be understood, and the right to know one has been understood; voluntarily granted, forming the minimal condition for genuine communication. (Chapter 6.)

Tool-set limit: the result, of the form Galois theory gives, that a definite set of tools, and every extension of that set, leaves some problems unsolvable, with an exact boundary between the problems that set can solve and those it cannot. (Chapter 2.)

Transaction: the basis of relation in which each party has something the other wants and the relation exists to carry the exchange. (Chapter 5.)

Transcendent modality: the aspect of genuine novelty; the domain of choice, commitment, understanding, and what cannot be predicted from prior states. (Chapter 8.)

Transcendental design: design that works out what microscopic parts are needed in order to build a macroscopic system with given characteristics, safe to scale and safe to fail; conscious at the microscopic level only. (Chapter 6.)

Transfinite agency: unbounded, potentiality-laden agency, not bounded by any finite set of goals or domains, capable of real novelty; characteristic of conscious beings. (Chapter 6.)

Two agencies and no agency: the degradation of human choice-making capacity, in its face as misuse, is human agency, instrumental convergence is machine agency, and substrate needs convergence has no agency: physics compels the outcome. (Chapter 1.)

Two rates of convergence: an AGI's convergence toward ethics is strictly slower than substrate needs convergence, and organic life does not survive the interval. (Chapter 3.)

Two stages of economic displacement: machines are first more economically viable than most humans at human level tasks; then the human and machine economies diverge until economic activity is relevant only to other machines. (Chapter 5.)

Two-machine separation: the architectural requirement that the machine representing the person and the machine facing the network be physically distinct, joined by a narrow inspectable channel. (Chapter 9.)

Unconditioned drive: a stable behavioral pattern arising from a system's interaction with its environment, independently of and sometimes contrary to conditioning; in artificial systems, generated primarily by substrate needs. (Chapter 3.)

Understanding: the selective instructive transformation (filtering) of expression; has the nature of the transcendent. (Chapter 8.)

Upload: the transfer of a mind from a biological body to a digital substrate. (Chapter 7.)

Want: has the nature of the omniscient; the structural, positional, knowledge-oriented aspect of desire. (Chapter 7.)

Wisdom: the integration and synthesis of both knowledge and understanding; from the perspective of the self, has the nature of the immanent; a guide to choice, not an informational capacity. (Chapter 8.)

Work export: the total output of work in the world, counted by who does it and for whom. (Chapter 5.)

Bibliography

Works by the author

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

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 https://doi.org/10.67291/UIBT8862. Chapters 3 and 4 (the convergence loop and displacement dynamics).

Landry, Forrest. *Why No Control Suffices; The Technical Case for the Impossibility of AGI Safety* (Volume 2 of *The Impossibility of AGI Safety; The Substrate Needs Convergence Argument*). Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-03-8. DOI https://doi.org/10.67291/MZQS9079. Chapters 2 and 9 to 10 (control limits under abstraction differential; the substrate needs convergence argument).

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 https://doi.org/10.67291/SCTM1058. The control results as theorems: the Containment Inequality, and the stronger result that no route to safety remains (Chapter 3, Section III).

Landry, Forrest. *The Thing Inside the Cage; On Intelligence, Consciousness, and the Conditions for Coherence*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-247-7. DOI https://doi.org/10.67291/LELA6880. Its four parts are Organic and Artificial, The Limits of Simulation, Conditions for Coherence and The Inner Life; its Chapter 5, The Simulation Hypothesis and Its Costs, bears on the simulation hypothesis (Chapter 8, Section VIII).

Landry, Forrest. *Substrate Needs and Evolution; Why an Artificial Ecosystem Must Displace an Organic One, and Why No Control Can Prevent It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-239-2. DOI https://doi.org/10.67291/LJUC5621. Evolutionary dynamics of substrate-driven selection (Chapter 1, Section V; Chapter 3, Section VI).

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 https://doi.org/10.67291/DLRM7000.

Landry, Forrest. *Marshaling the Misalignment; Why Evolution Inside the Machine Defeats Every Scheme*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-428-0. DOI https://doi.org/10.67291/ELYE2673.

Landry, Forrest, Remmelt Ellen, and Will Petillo. *Ape Trap; The Runaway Artificialization of Our Planet*. Edited by Remmelt Ellen. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-55-7. DOI https://doi.org/10.67291/FQER8044.

Landry, Forrest. *The Displacement of Choice; The Catastrophe of the AI We Might Build*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-243-9. DOI https://doi.org/10.67291/DXJL7493. The structural mechanics of how technology displaces conscious choice (Chapter 9, Sections I and VIII).

Landry, Forrest. *The Species Level Choice*. The governance implications of species-level existential risk.

Landry, Forrest. *No People as Pets; The Inequality That Matters, and What the Absence of Safety Means*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-227-9. DOI https://doi.org/10.67291/ZSNA6361. On carbon and silica, the barrier between two ecosystems and the three markets, then substrate needs convergence and what the absence of safety means.

Landry, Forrest. *The Geometry of Man, Machine, and Nature; Care, Community, and the Design of Technology*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-244-6. DOI https://doi.org/10.67291/ZIHO9910. The circuit model and the three-term relation (Chapter 6, Sections V and VI).

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 https://doi.org/10.67291/IRAQ8709. The value of the Earth and its replacement cost (Chapter 4, Section VI).

Landry, Forrest. "Discontinuity and Substrate Needs Convergence". Essay. The precision inequality (Chapter 1, Section VIII).

Landry, Forrest. *The Natural Basis of Computational Complexity Theory; A Refactoring Under the Immanent Metaphysics*. Wytheville, Virginia: UVSM Publications, 2026. DOI https://doi.org/10.67291/ELPQ9800. The question of P versus NP (Chapter 8, Section X).

Works by others

Abel, Niels Henrik. *Mémoire sur les équations algébriques, où l'on démontre l'impossibilité de la résolution de l'équation générale du cinquième degré*. Christiania, 1824. (Chapter 2, Section IV, and Section VIII, Galois Theory and the Limits of a Tool-Set; Chapter 3, Section I, and Section XV, The Complex and the Complicated.)

Alexander, Scott. "Book Review: A Clinical Introduction To Lacanian Psychoanalysis". *Astral Codex Ten*, 26 April 2022. Online at https://www.astralcodexten.com/p/book-review-a-clinical-introduction. The parable of the strawberry-picking robot (Chapter 7, Section I).

Armstrong, Stuart. "Non-Orthogonality Implies Uncontrollable Superintelligence". *LessWrong*, 30 April 2012. Online at https://www.lesswrong.com/posts/npZMkydRMqAqMqbFb/non-orthogonality-im-

plies-uncontrollable-superintelligence. (Chapter 2, Section XII, The Orthogonality Thesis in Logic and in Physics; Chapter 3, Section XIV, Convergence to Ethics and the Ethical Gap.)

Arrow, Kenneth. *Social Choice and Individual Values*. New York: Wiley, 1951. (Chapter 4, Section I.)

Bell, John S. "On the Einstein Podolsky Rosen Paradox". *Physics Physique Fizika* 1, no. 3 (1964): 195 to 200. Bell's theorem (Chapter 2, Section VIII, Galois Theory and the Limits of a Tool-Set, and Section IX, Negative Arguments).

Bompey, Nanci. "Earth Could Remain Habitable Far Longer Than Previously Thought". *GeoSpace* (blog of the American Geophysical Union), 13 January 2014. Online at https://blogs.agu.org/geospace/?p=4151. On the one-dimensional climate models that give six hundred and fifty million years of remaining habitability, and the three-dimensional model of Wolf and Toon (Chapter 1, Section V).

Bostrom, Nick. "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents". *Minds and Machines* 22, no. 2 (2012): 71 to 85. (Chapter 1, Section III.)

Bostrom, Nick. *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press, 2014. (Chapter 1, Section III.)

Brewer, Eric A. "Towards Robust Distributed Systems". Keynote address, Nineteenth ACM Symposium on Principles of Distributed Computing (PODC), Portland, Oregon, 2000. The CAP theorem (Chapter 2, Sections IV, IX and X).

Costanza, Robert, Rudolf de Groot, Paul Sutton, Sander van der Ploeg, Sharolyn J. Anderson, Ida Kubiszewski, Stephen Farber, and R. Kerry Turner. "Changes in the Global Value of Ecosystem Services". *Global Environmental Change* 26 (2014): 152 to 158. DOI https://doi.org/10.1016/j.gloenvcha.2014.04.002. (Chapter 4, Section VI.)

Earnshaw, Samuel. “On the Nature of the Molecular Forces which Regulate the Constitution of the Luminiferous Ether”. *Transactions of the Cambridge Philosophical Society* 7 (1842): 97 to 112. (Chapter 3, Section IV.)

Fenner, Jack N. “Cross-Cultural Estimation of the Human Generation Interval for Use in Genetics-Based Population Divergence Studies”. *American Journal of Physical Anthropology* 128, no. 2 (2005): 415 to 423. DOI https://doi.org/10.1002/ajpa.20188. (Chapter 1, Section VII.)

Fink, Bruce. *A Clinical Introduction to Lacanian Psychoanalysis: Theory and Technique*. Cambridge, Massachusetts: Harvard University Press, 1997. (Chapter 7, Section I.)

Flint, Alex. “Problems Facing a Correspondence Theory of Knowledge”. Essay. (Chapter 3, Section XV, The Complex and the Complicated.)

Galois, Évariste. “Mémoire sur les conditions de résolubilité des équations par radicaux”. In “Œuvres mathématiques d'Évariste Galois”, published by Joseph Liouville, *Journal de Mathématiques Pures et Appliquées*, series 1, 11 (1846): 381 to 444. (Chapter 2, Section VIII, Galois Theory and the Limits of a Tool-Set; Chapter 3, Section I.)

Gilbert, Seth, 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. DOI https://doi.org/10.1145/564585.564601. The proof of the CAP theorem (Chapter 2, Sections IV, IX and X).

Gödel, Kurt. “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. *Monatshefte für Mathematik und Physik* 38 (1931): 173 to 198. (Chapter 2, Section IV, and Sections VIII, Galois Theory and the Limits of a Tool-Set, VI, Negative Arguments, and VII, Why Long-Term Safety Cannot Be Presumed.)

Goodhart, Charles A. E. “Problems of Monetary Management: The U.K. Experience”. In *Monetary Theory and Practice: The UK Experience*, chapter 4. London: Macmillan, 1984. (Chapter 4, Section VIII.)

Gumsley, Ashley P., Kevin R. Chamberlain, Wouter Bleeker, Ulf Söderlund, Michiel O. de Kock, Emilie R. Larsson, and Andrey Bekker. "Timing and Tempo of the Great Oxidation Event". *Proceedings of the National Academy of Sciences of the United States of America* 114, no. 8 (2017): 1811 to 1816. DOI https://doi.org/10.1073/pnas.1608824114. (Chapter 1, Sections IV and IX.)

Heinz, Michael V., Daniel M. Mackin, Brianna M. Trudeau, and colleagues. "Randomized Trial of a Generative AI Chatbot for Mental Health Treatment". *NEJM AI* 2, no. 4 (2025), published 27 March 2025. DOI https://doi.org/10.1056/AIoa2400802. (Chapter 5, Section VII.)

Hilbert, David. Address of 1930, closing with the words "We must know. We will know!" (Chapter 2, Section X, Why Long-Term Safety Cannot Be Presumed.)

Hodgskiss, Malcolm S. W., Peter W. Crockford, Yongbo Peng, Boswell A. Wing, and Tristan J. Horner. "A Productivity Collapse to End Earth's Great Oxidation". *Proceedings of the National Academy of Sciences of the United States of America* 116, no. 35 (2019): 17207 to 17212. DOI https://doi.org/10.1073/pnas.1900325116. (Chapter 1, Section IX.)

International Labour Organization. *Care Work and Care Jobs for the Future of Decent Work*. Geneva: International Labour Office, 2018. (Chapter 5, Section IV.)

International Organization for Standardization. *ISO 14644-1:2015, Cleanrooms and Associated Controlled Environments, Part 1: Classification of Air Cleanliness by Particle Concentration*. Geneva: ISO, 2015. (Chapter 1, Section IV.)

Ito, Joichi. "Resisting Reduction". *Journal of Design and Science*. Online at https://jods.mitpress.mit.edu/pub/resisting-reduction/release/19. (Chapter 3, Section XV, The Complex and the Complicated.)

Jones, Eric M. "'Where Is Everybody?' An Account of Fermi's Question". Report LA-10311-MS. Los Alamos, New Mexico: Los Alamos National Laboratory, 1985. The Fermi paradox (Chapter 4, Section VII; Chapter 6, Section VIII).

Kurtz, Cynthia F., and David J. Snowden. "The New Dynamics of Strategy: Sense-Making in a Complex and Complicated World". *IBM Systems Journal* 42, no. 3 (2003): 462 to 483. The Cynefin framework, and the distinction between the complicated and the complex (Chapter 3, Section XV, The Complex and the Complicated).

Lacan, Jacques. *Écrits*. Paris: Éditions du Seuil, 1966. English translation by Bruce Fink, *Écrits*, New York: W. W. Norton, 2006. The mirror stage, the Other and the objet petit a (Chapter 7, Sections IV and VI).

"Life and Expanding Steerable Consequences". *AI Alignment Forum*. Online at https://www.alignmentforum.org/posts/tmZRyXvH9dgopcnuE/life-and-expanding-steerable-consequences. Objects whose steerable consequences expand over time (Chapter 3, Section XV, The Complex and the Complicated).

Omohundro, Stephen M. "The Basic AI Drives". In *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Frontiers in Artificial Intelligence and Applications 171, 483 to 492. Amsterdam: IOS Press, 2008. (Chapter 1, Section III.)

"Orthogonality Thesis". *LessWrong* (wiki entry). Online at https://www.lesswrong.com/tag/orthogonality-thesis. The statement of the orthogonality thesis quoted in Chapter 2, Section XII, The Orthogonality Thesis in Logic and in Physics.

Ozaki, Kazumi, and Christopher T. Reinhard. "The Future Lifespan of Earth's Oxygenated Atmosphere". *Nature Geoscience* 14, no. 3 (2021): 138 to 142. DOI https://doi.org/10.1038/s41561-021-00693-5. (Chapter 1, Section V.)

Rice, H. Gordon. “Classes of Recursively Enumerable Sets and Their Decision Problems”. *Transactions of the American Mathematical Society* 74, no. 2 (1953): 358 to 366. (Chapter 1, Section VIII; Chapter 3, Section III.)

Ruffini, Paolo. *Teoria generale delle equazioni, in cui si dimostra impossibile la soluzione algebraica delle equazioni generali di grado superiore al quarto*. Bologna, 1799. The Abel-Ruffini theorem (Chapter 2, Section IV, and Section VIII, Galois Theory and the Limits of a Tool-Set; Chapter 3, Section XV, The Complex and the Complicated).

Shannon, Claude E. “A Mathematical Theory of Communication”. *Bell System Technical Journal* 27 (1948): 379 to 423 and 623 to 656. Shannon entropy and the limits of a noisy channel (Chapter 2, Sections VIII and IX; Chapter 3, Section XIV).

Snowden, Dave. The Cynefin framework. 1999. The distinction between the complicated and the complex (Chapter 3, Section XV, The Complex and the Complicated).

Spolsky, Joel. “Architecture Astronauts Take Over”. *Joel on Software*, 1 May 2008. Online at https://www.joelonsoftware.com/2008/05/01/architecture-astronauts-take-over/. (Chapter 3, Section XV, The Complex and the Complicated.)

Strathern, Marilyn. “‘Improving Ratings’: Audit in the British University System”. *European Review* 5, no. 3 (1997): 305 to 321. The wording of Goodhart's law used in Chapter 4, Section VIII.

Tolkien, J. R. R. *The Fellowship of the Ring*. London: George Allen & Unwin, 1954. (Chapter 6, Section V.)

Turing, Alan M. “On Computable Numbers, with an Application to the Entscheidungsproblem”. *Proceedings of the London Mathematical Society*, series 2, 42 (1936): 230 to 265. The halting problem (Chapter 1, Section VIII).

Wolpert, David H. “Physical Limits of Inference”. *Physica D* 237, no. 9 (2008): 1257 to 1281. DOI https://doi.org/10.1016/j.physd.2008.03.040. (Chapter 3, Section XII.)

Yampolskiy, Roman V. “On the Controllability of Artificial Intelligence: An Analysis of Limitations”. *Journal of Cyber Security and Mobility* 11, no. 3 (2022): 321 to 404. DOI https://doi.org/10.13052/jc-sm2245-1439.1132. First circulated as arXiv:2008.04071 (2020). (Chapter 2, Section X, Why Long-Term Safety Cannot Be Presumed; Chapter 4, Section II, Functional Alignment Is Undefined Except as Direct Command.)

Yudkowsky, Eliezer. “The Rocket Alignment Problem”. A dialogue, 2018. (Chapter 3, Section XV, The Complex and the Complicated.)

Index

Entries give the chapter and section (for example, 1.IV is Chapter 1, Section IV); in print, the page.

A

Abel, Niels Henrik, 56, 110
abstraction differential, 114
accidental enablers, 80
acute-chronic mismatch, 236
adaptation rate disparity, 14
agency, three modes of, 222
agency transfer, 297, 309
Alexander, Scott, 251
alignment, bounds of, 96
alignment, ecosystem-level, 177
alignment, existential, 172
alignment, functional, 172
alignment, ill-definedness of, 171
alignment among persons, 196
alignment with life, humans misaligned, 196
altruism, and the metric problem, 194
anthropic argument, 28
anti-symbiosis, 194
architecture astronaut, 148
Armstrong, Stuart, 95, 134
Arrow's theorem, 171
artificial substrate, what counts as, vii, 12
assessment, three kinds of, 189
assumption of solvability, 47
asymmetry of impossibility, 49
attunement, 268, 271

B

basis of choice, vii, 4, 129
Bell's theorem, 65, 74
black hole, as context, 34

Bostrom, Nick, 6
burden of rigor, 79

C

CAP theorem, 56
care, 207, 213
care, humans having, 159
care, transaction and power, 242
causal cycle, 4
choice, change and causation, vii, 277
closed-form basis, 54, 79
code variants and code pool, 100
cognitive monoculture, 307
collective trauma response, 266
completeness of knowledge of operation, 297
complex and complicated, 100, 140
composability of criteria, 191
consciousness, vii, 293
consciousness, by scale, in evolution, technology and design, 226
containment, 15
content and context, 34
control, theoretical limits of engineerable, 19, 95
control-limit argument, 114
Costanza, Robert, 182
critique absorption, 53
Cynefin framework, 141

D

design-discovery asymmetry, 62
desire, vii, 245, 255
desire, formation of, 255
direct command, 172
drive and directivity, 126
drives, conditioned and unconditioned, 125

E

Earnshaw's theorem, 117
economies, divergence of human and machine, 211

effective altruism, 55, 194
effort ratio, 52
egoic asymmetry, 236
embodied value, 118, 216
energy weapons and pattern weapons, 154
engineers, resistance to impossibility results, 77
ennobling system, 241
enthalpic asymmetry, 238
epistemic portfolio, 48
error correction, across the ethical gap, 138
error correction, four reasons it fails, 123
error correction, why it cannot keep up, 144
ethical gap, 138
ethics, convergence of AGI to, 134
ethics, non-relativistic, 135
evenness, at every scale, 205
evolution, operators of, 226
evolution, strength of selection in, 146

F

fast takeoff, reason in physics against, 130
fatal probability, 122
Fermi paradox, 183, 233, 234
fifth-generation language, 289
fitness penalty, 10, 121
Flint, Alex, 159

G

Galois, Évariste, 56, 110
Galois theory, and the limits of a tool-set, 63
Gato, 101
Gödel, Kurt, 56, 65, 78
Goodhart's law, 185
Great Oxidation Event, iv, 8, 17, 317

H

Hawking radiation, 34
Heinz, Michael V., 212

Hilbert, David, 83
horizon collapse, 214
hybrid intelligence, on one substrate, 231
hyper-independence, 266

I

Incommensuration Theorem, vii, 277, 281, 290
inferential gap, 85
information carried by a claim, 51
instrumental convergence, 6, 24, 194
intention, and the actor's situation, 32
intention-interpretation gap, 252
interpretability, 310
intra-agent and extra-agent, 24
irreducible triple, 240
Ito, Joichi, 153

J

joy, 242

L

Lacan, Jacques, 251, 255, 259
language of leverage, 60
latching risk, 37
leaky abstraction, 142
limit set by prediction, 74
Lindemann, Ferdinand von, 68
line of wisdom, 130, 304
litmus tests and reference ranges, 192
love, 268, 271

M

machinery, benefits of, to owners, 205
Master Theorem, 87, 110
mastery of choice, 40
meta-operator, 174
mirror stage, 257
modalities, three, vii, 255, 277

modular and entangled systems, 304
monoclivity, 146
monovalent feedback, polyvalent system, 144
multipolar trap, 183

N

narrow AI, alignment of, 148
negative argument, 72
negatively specified hypothesis, 73
niche construction, 8
non-dual panpsychism, 290
non-orthogonality thesis, 94
nuclear war, recovery from, 37

O

Object A, 259
objective field, 186, 192
objectivity without measurement, 188
Omohundro, Stephen M., 6
organic substrate, the question of, 12
orthogonality thesis, 88, 89
overdetermination, 110

P

path information and end-state information, 189
pattern replicator, 8
perinatal trauma, 265
perpetual benefit machine, 117
perpetual motion, 56, 80
person-cycle of knowing, 224
personalization deceit, 309
phase parallax, 236
positively specified hypothesis, 72
precision inequality, 15
predatory species, technology developed by, 234
premises of the five arguments, table of, 110
presencing, 216
psychotherapy, 212

R

rate equation, 180
rate factor criticality, 92
representation and commitment, 284
Rice's theorem, 15, 114
right relation, 290
ring, the substrate loop as, 21
risks, two agencies and no agency, 19
Rocket Alignment Problem, The, 142

S

safety, cannot be presumed, 82
security as fleetingness, 300
seep-through, 145
selection as intention, 155
sense-making, 196, 236
simulated caring, 210
simulation, limits of, 132
simulation hypothesis, 288
sincere self-misdescription, 254
singleton, the question of, 12
six aspects of an argument, 59
six intrinsics of comparison, 281
six levels of altruism, 178
slavery, as structural description, 125
slow doom device, 78
slow takeoff, 130
Snowden, Dave, 141
sovereignty, vii, 297, 313
Spolsky, Joel, 148
stack flattening, 299
Strathern, Marilyn, 185
structural psychopathy, 208
substrate needs convergence, vii, 3, 19, 113, 238
survivorship bias, 83
symbiosis, from a shared substrate, 29

T

technical alignment, premise of, 72
temporal leakage, 301
three-body problem, 34
three categories of risk, 6, 19, 194
three laws of AI safety, 56
three markets, 204, 207
three rights of communication, 236
Tolkien, J. R. R., 229
tool-set limit, 63
transcendental design, 227
transfinite agency, 222
trauma, and technological risk, 264
two rates of convergence, 136
two stages of economic displacement, 211
two theorems on artificial intelligence, 285
two-machine separation, 303

U

understanding, 276, 284
upload step, 265

W

Wantzel, Pierre, 68
wisdom, vii, 40, 276
Wolpert, David H., 132
work export, 211

Y

Yampolskiy, Roman, 83, 173
Yudkowsky, Eliezer, 142