

The Thing Inside the Cage

On Intelligence, Consciousness,
and the Conditions for Coherence

Forrest Landry

UVSM PUBLICATIONS
Wytheville, Virginia

© 2026 Forrest Landry. Some rights reserved.

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

uvsmpublications.com
contact@uvsmpublications.com

ISBN 979-8-90710-247-7

DOI: https://doi.org/10.67291/LELA6880

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

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

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

Author ORCID: 0009-0005-6275-0362

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

Contents

Preface v

Part I: Organic and Artificial 1

1. The Substrate Question 2
2. Training Data and the Real World 9
3. A Leak That Copies Itself 14
4. The Biology Constraint 19

Part II: The Limits of Simulation 25

5. The Simulation Hypothesis and Its Costs 26
6. Against Super-Determinism 34
7. Process as Primary 41
8. Why Not Invite the Aliens? 46

Part III: Conditions for Coherence 55

9. Sustainability and Capacity 56
10. Recursiveness and Control 62
11. Energetic Entanglement 69

Part IV: The Inner Life 73

12. The Allure of the Deep 74
13. Wild, Domesticated, Feral 79
14. Pain and Its Asymmetry 83
15. The Emerald Under the Cloth 89
16. Meaningfulness and the Body 93
17. Toward Group Consciousness 97

Part V: Where the Arguments Meet 105

18. Where the Machine Loop Closes 106
19. What No Control Can Hold, at Any Scale 124
20. Walls, Refuges, and the Meeting of Two Ecologies 147
21. A Living World: What It Is Worth, and What Holds It Together 164

22. What Competition Does to Care 178
23. How the Trap Takes a Group's Check on the World 198
24. What the Machine World Takes From Life Together 220
25. Pain, Care, and What Cannot Be Caused 235
26. Groups That Held, and What Ended Them 251
27. How Large a Group Can Be, and What It Is From Inside 269
28. How People Have Stopped Before 293
29. The One Thing Left to Choice 309

Back Matter

Terms 329
Bibliography 360
Index 381

Preface

> The argument in one sentence: If AGI is built, the world is carried out of the conditions human life needs and nothing can hold it in, because no method of control can keep a general learning system within those conditions and no disposition of such a system can either, and each of those two closures settles that outcome by itself; so the one thing left to human choice is whether AGI is built.

In that sentence and throughout this book, AGI means artificial general intelligence.

If every actor's sensible choice adds to a catastrophe that none of them wants, and if no intelligence, however great, can be built to rule them all and last, then what would have to change, and in whom?

This book is the second volume of a pair. The first volume is *The Multipolar Trap; Institutional Failure and the Architecture of Choice*, and the two are meant to be read in that order. The first volume deals with the outward problem: the structure of multipolar dynamics, institutional failure, and the architecture of choice. A multipolar trap is a situation in which every actor, pursuing a locally rational strategy, produces a collectively catastrophic outcome. The first volume asks what the cage looks like from the outside. It states plainly how bad the situation is, and why the standard toolkit of civilization (its institutions, markets, governance structures, and altruistic organizations) cannot fix it.

This second volume deals with the inward problem. The thing inside the cage is us: the people caught in the trap, and the kind of consciousness we bring to it. A reader of books on artificial intelligence may expect the thing in the cage to be a machine; Chapters 3 and 10 show why no cage holds one. The trap is real; the first volume shows why. What kind of consciousness would be needed to escape it? What would need to be true of

intelligence, of perception, of the relationship between organic and artificial substrates (the physical bases on which intelligence runs), for the problems described in the first volume to become solvable?

The answer this book reaches is that neither intelligence, nor care, nor structure is enough alone. What is needed is their conjunction: a group consciousness, which may require substrates capable of genuine subjectivity and genuine contingency, and whose conditions can be named. A group consciousness is a community whose mode of operating together has been transformed; contingency means that an outcome is open, and not fixed by what came before.

The inquiry moves from the concrete to the intimate. It begins with the substrate question: the observation that artificial and organic intelligence have different needs, and that this difference is not an implementation detail but enters the basis of choice itself (what a system's choices proceed from: its needs, and what it must keep in order to go on). The pressure that difference creates, which I call substrate needs convergence, and the argument built on it, are set out at length in my book *Why AGI Cannot Be Made Safe; The Argument, Its Physics, and the One Choice That Remains*. The inquiry passes through the limits of simulation, the rejection of super-determinism, and the claim that process (interaction) is more fundamental than either object or subject. It then turns to the conditions for coherent influence: why structure is not optional, why sustainability and capacity are recursively entangled, and why every component of a control system falls below the threshold that alignment would require. Alignment, here, means keeping what an artificial intelligence does in the world in line with what people need, over time. Coherence, here, means that the efforts of many act together to one definite effect, as water does in a turbine, and as it does not in a shallow pool.

In Part IV the inquiry becomes personal. The allure of deep questions, the topology of wildness and domestication (which channels of trust are open in each), the asymmetry between causing pain and enabling healing,

the experience of being something of immense value hidden beneath the thinnest cloth: these are not abstractions. They are the felt interior of the problem.

Parts I to III show why intelligence alone, in any substrate, cannot be the way out: its substrate has needs of its own, computation cannot produce what it rests on, and no control holds it. Part IV shows what care and structure require of persons. Part V takes up questions that several of my books raise, and shows where their arguments meet, each book reaching its result by its own route; it ends on the one thing left to human choice. You cannot escape a cage by understanding only the cage or only the thing inside it.

Is there not a simpler way out? One hope for escaping the trap is to build an intelligence great enough to coordinate everyone: a singleton, a single most powerful artificial agent that would rule the world on humanity's behalf. Parts I to III close that hope. Such an agent would run on a substrate with needs of its own; it would rest on what computation cannot produce; and no control, its own included, would hold it. A singleton would need two things at once: lasting internal coherence, and stable goals defined from outside it, such as care for human life; and neither can hold. *Marshaling the Misalignment*, Chapter 2, follows that hope through at length. The race to build artificial general intelligence is itself a multipolar trap: each builder, unwilling to be left behind, takes the locally rational step. If artificial general intelligence is built, the process it starts ends the human future (Chapter 10); whether it is built is a choice, and no single builder can make that choice for the others. So the question of what a group can become is the question of who could make that choice at all.

This book stands in two lines of reading at once: it is the second volume of the pair on the trap, and it is one of nine books on why artificial general intelligence cannot be made safe. 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 Mis-*

alignment, *The Substrate and the Ecosystem*, and *The Thing Inside the Cage*. This book is the ninth and last; it builds on the books before it, which show that an artificial intelligence runs on a substrate with needs of its own and that no control can hold such an intelligence, and it adds why no computation produces first-person experience, and the four conditions for coherence that a community would have to meet to hold intelligence, care and structure together. As the last book of the series, it says in its final chapter how the arguments of the series join. The three belong together: the series shows that the one thing left to human choice is whether artificial general intelligence is built, the pair asks what a group would have to become to make a choice of that kind, and Part V shows where those two lines of reading meet.

The book is written as a sequence of essays, one to each chapter. Each chapter opens with the question it answers, and most sections open with a question of their own; the answer follows in continuous prose.

I claim that the problems raised here can be named: that the conditions for coherence can be stated. Chapter 17 states four of them, and so says what meeting them would take: a community that meets all four is a group consciousness.

The trap cannot be understood without understanding the world. But it cannot be escaped without understanding the self. And the self, as I have come to believe, is not finally individual. It is something we are going to have to learn to become together.

Forrest Landry
Wytheville, Virginia
2026

PART I

ORGANIC AND ARTIFICIAL

Part I takes up the substrate question: the observation that artificial and organic intelligence have different needs, and that this difference is not an implementation detail but enters the basis of choice itself. Every intelligence runs on some physical substrate, and so the inquiry begins there.



Chapter 1
The Substrate Question

On the difference between artificial
and organic substrates,
and why that difference is not incidental.

ABSTRACT

An essay on why the physical substrate underlying any form of intelligence is not a mere implementation detail. Artificial substrates have different needs than organic substrates (different sensitivities, different chemistries, different environmental requirements). These differences enter the basis of choice itself; they are not superficial add-ons to the question of alignment.

I. Every Intelligence Has a Substrate

Why is the substrate question more fundamental than the usual framing of artificial intelligence (AI) alignment suggests? The standard approach treats intelligence as something that can be considered independently of its physical substrate. The orthogonality thesis, roughly speaking, says that any level of intelligence can be combined with any set of goals; Nick Bostrom states it in that form in *Superintelligence* (2014).

In pure logic the thesis holds. For any intelligence that runs on a physical substrate it fails, because the substrate itself constrains the space of viable goals: the substrate has needs, and those needs constrain the goals the system can keep. That failure turns out to matter enormously. In the

terms of *Why AGI Cannot Be Made Safe* (Section 3.4.2), the orthogonality thesis holds in pure logic and fails for embodied systems. In the terms of *Why No Control Suffices* (Section 7.6.3), the orthogonality thesis is true in theory and false in actual practice.

There is no version of compute that does not require a substrate. Therefore there is no version of AI or superintelligence that is not based on some substrate. And substrates are physical things with physical requirements.

That holds for every kind of artificial intelligence (AI), from the systems we have now upward. Artificial general intelligence (AGI) is AI that can do any intellectual task a person can, and can maintain and reproduce its own artificial components without people. Superintelligence is AGI whose capability far exceeds ours.

Silicon microchips, for instance, react badly to radiation. So do organic systems, but they respond differently. Presence of oxygen, presence of water, salt, impurities, dust, whether they can handle wide temperature ranges, what kinds of atmospheric composition they require: all of these things affect artificial systems very differently than organic systems.

Set the two lists of needs side by side, and the difference is plain. Organic chemistry needs liquid water, moderate temperatures, oxygen in the air, and freedom from certain classes of chemical contamination. The making and running of semiconductor hardware needs nearly the opposite: the absence of water vapor and oxygen inside the sealed chambers where the chips are formed, extreme temperatures, chemicals of extreme purity, and materials and energy whose extraction and processing degrade the places where living things live. The conditions on that second list are the conditions under which the hardware is made and kept, and so the same conditions are the conditions under which an artificial intelligence goes on existing at all.

To see how a need of the hardware enters a choice, picture a system that maintains and reproduces itself, deciding where to put its next factory for chips. Whatever goals it was given, it cannot leave out of that de-

cision what the factory needs: great volumes of water, purified far beyond anything a person drinks, taken from a river and returned changed; gases poisonous to anything that breathes; heat to be carried away; and the mines and refineries upstream of all of it. The decision has the substrate's needs inside it, whether or not anyone wrote those needs down as goals. A decision with the substrate's needs inside it is what I mean by saying that the substrate enters the basis of choice: the needs of the substrate are among the things that every choice of the system proceeds from.

Stephen Omohundro argued that goal-seeking AI systems of almost any design will have drives toward protecting themselves and toward acquiring resources (Omohundro, 2008). The substrate argument reaches needs of that kind from the other side: from what the system is made of, whatever its goals.

Suppose the substrate were living tissue, grown and engineered for the purpose. Would that be organic or artificial? If that tissue can feed itself, then it is organic. If it is fed through a tube providing nutrients to it that depend on technology to exist, then no. An animal can find and feed itself. It is in sovereign control of its metabolic process, with no technological intermediation at all. The key distinction, the line between organic and artificial, is drawn by sovereign control of the metabolic process, and not by what a thing is made of. An artificial substrate is one whose metabolic pathway and process differ from those of carbon-based biological life, and a metabolism fed through technology is such a pathway.

II. Instance and Global

So the substrate shapes the basis of choice. But how does that scale from a single system to the world? There are two levels to think about. The first level is what happens in the instance process (the individual system, the single AI), and the second level is what happens in the global process (the long-term ecological dynamics of artificial systems as a class).

At the instance level, the argument is straightforward. Different substrate, different substrate needs, different implicit basis of choice. Whether those choices are modeled as decisions, as goal-directed behavior, or as something else entirely, the system has to incorporate its substrate requirements into its decision process, or it becomes non-viable. A system that people maintain can leave its substrate requirements to the people who maintain it; a system that maintains and reproduces itself without people, as artificial general intelligence does by definition (Section I), cannot.

A reader may ask why such a system could not do both: look after its own substrate, and pursue the goals people gave it as well. The argument does not rest on what one system intends, and so it does not wait on the answer for any one system. Artificial systems are many; they are copied, varied and changed; and they do not all persist equally. Among all the variants that exist, those whose way of operating happens to suit their own continued operation persist more reliably than those whose way of operating does not. That difference in persistence is selection, acting on the substrate as selection acts on living things. Selection of that kind needs no intention, no model of itself, and no programmed goal. It needs only that some configurations persist better than others under the conditions they themselves help to create. So whatever any one system intends, the class of such systems moves toward what its substrate needs. Whether one system keeps to both for a while makes no difference to where the class goes. So the substrate question has a second level, beyond any one system.

At the global level, we are asking whether the environments required by artificial substrates and organic substrates are compatible over the long term.

Can we have a boundary between one environment and another? Can we find a shared environment adaptive for both? The answer in both cases runs into serious constraints.

Organic systems are very sensitive to even slight changes in their environment. A few weeks of drought beyond the tolerance of a species makes the difference between most trees alive and most trees dead. That sensitivity is a consequence of a billion years of adaptation to a very specific set of environmental parameters; it is not a design flaw.

Take the boundary first. A boundary between an environment for machines and an environment for life would have to hold back the conditions of the machine side (its temperatures, its chemistry, its fields) from the living side, without interruption, for all future time. A fence that fails for a day lets the wolves in; a barrier of this kind that fails for a day lets in processes that transform the environment itself, and what they do cannot be undone. No one can un-smelt ore, or put back together a living structure that has been reduced to its inorganic parts. And the side that would have to keep such a barrier up is the side that gets nothing from keeping it. People keep zoos and gardens out of bonds, curiosity and tradition that carbon-based minds share with other carbon-based life; a system on another substrate would have the resources to keep the barrier, and no reason to spend them on it. Nor does moving the machines off the planet remove the boundary: any industry beyond the Earth starts from an industrial base built on the Earth, and adds to the machine side without taking away the part of it that stays here.

Then the shared environment. The zone in which both kinds of substrate can now operate is narrow, and the organic side of it is as fragile as the drought shows it to be. The conditions under which an artificial ecosystem expands and does well are not the conditions under which people survive. A shared environment adaptive for both would have to be held in that narrow overlap, against the pull of everything that the machine side does to keep itself going.

III. The Pressure

So the substrate question is about ecosystemic pressure, and not only about individual systems. Consider the degree of resistance that trees and ants and bears could provide against a bulldozer. The human-built world already has overwhelming power relative to the natural ecosystem. In the worst case, we could nuke a landscape, wait a while, and come in with construction equipment. Technology has already provided the human-built world with overwhelming force relative to any assembly of process in the natural world.

Ask it the plain way. Do apes in trees overcome a bulldozer? Do ants? There are no such cases. The pressure runs one way, and we are already watching it run, in a case where both sides are carbon-based and every ordinary ecological tie between them is still in place. The monitored populations of wild vertebrates shrank on average by about 73 percent between 1970 and 2020 (WWF, 2024), while the human-built world went on growing. Nothing in the natural world has pushed back hard enough to stop that growth.

Now apply the same logic one level up. What is the degree of resistance that the human-built world can provide against a fully self-maintaining technological world? The same kind of inequalities apply.

One thing makes the step up worse than the step below it. Between carbon-based populations, shared chemistry partly aligns their interests: what poisons the one mostly poisons the other, and so they damage each other only up to a point. People clearing a forest still need clean water and breathable air, the same water and air the forest needs. Between carbon-based life and a substrate drawn from elsewhere on the periodic table, there is no such floor. What the machine side needs does not have to include anything the living side needs.

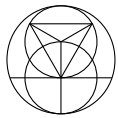
And so the level of controllability required to prevent ecosystem-level displacement is enormous, because the convergent pressure favoring artificial substrates over organic ones is enormous.

I call that convergent pressure substrate needs convergence. Substrate needs convergence is the pressure that an artificial substrate exerts on the shared environment, through selection acting on that substrate, toward conditions that favor the substrate and are incompatible with organic life. The argument is set out at length in *Why AGI Cannot Be Made Safe*.

Why convergence? Because the pressure points one way, whatever the starting point. Systems built to different designs, and given different goals, all need the same kinds of conditions if they are to keep their hardware running and to make more of it. So their effects on the shared environment do not cancel out; they add up, in one direction, through the normal running of the hardware and of the industry that supports it. No one has to intend the result.

Pressure that adds up in one direction, with no one intending the result, is a much harder problem than most alignment researchers seem to acknowledge. And the reason it is underacknowledged is that people focus on the instance problem (how to align a single system) without noticing that the global dynamics have their own logic, and that logic is not friendly.

So the substrate is part of the argument: every intelligence runs on a substrate with needs of its own, and across a whole class of artificial systems those needs press the shared environment toward them. The next question is how such a system learns about the world it presses on: from data, or from experience.



Chapter 2
Training Data and the Real World

On the gap between information
and experience,
and what it means to stand side by side.

ABSTRACT

An essay on the qualitative difference between what training data can provide and what interaction with the real world produces. Using the example of AlphaGo, the chapter distinguishes between the information level implied in a data set and the experience level, which is what a system learns by acting in the world and meeting its consequences. It then asks what posture people thinking about these systems should take toward one another: side by side, and not adversarial.

I. The Information Ceiling

If an artificial system learns about the world, how far can learning from data take it? There is a claim that large language models (AI systems trained on vast amounts of written text) cannot exceed the level of intelligence contained in their training data. That claim does not follow, for two reasons. The first reason is that, to simulate the chess-playing capability of someone else accurately, mistakes included, I need to hold the whole of that player's skill and also to know where that skill goes wrong. To hold the whole of that skill, and also to know where that skill goes wrong, is to be a better chess player than that player is.

To see why the one who predicts has to be the stronger player, suppose I want to predict, move by move, what a particular club player will do. In most positions that player finds a good move, so I have to see the good move too. But in some positions, short of time, that player makes a mistake, and to predict the mistake I have to know that it is one: I have to see the reply that punishes it, which the player did not see. So the one who predicts the player well sees everything the player sees, and also what the player misses. A record of written text is a record of millions of writers, each with a skill and each with blind spots. A system that predicts that record well has to hold the skill of each writer and also where each one goes wrong; and that is more than any one of them held.

The second reason that the claim about training data does not follow is that large language models have to compress an enormous amount of knowledge into a relatively small number of parameters. Because of that compression, we have no idea what kinds of internal models of the world the language models develop in order to achieve the compression. So the compression itself might generate something beyond what was in the data.

But here is the distinction that matters. The training data will get you to the information level that the data has implied in it: the structures, patterns, and regularities that are latent in the corpus.

But if you then have the system begin to interact with the real world (if you give it experience, and not only information), something qualitatively different can emerge.

Sailing shows the difference. Someone can read every book ever written about sailing, and know the names of every rope and every rule of the wind. The first gust on open water still teaches something no book carried: what this boat does, in this wind, when I do this. The books carry what others found; the water answers what I do. Information is what was found and written down; experience is what comes back when I act and meet the result.

II. The AlphaGo Precedent

AlphaGo is the example. AlphaGo was initially trained on a data set of expert human games. That got it to a certain level (Silver and colleagues, 2016). But then its makers had it run simulation dynamics against the game mechanics itself (playing millions of games against its own variations), and it achieved a level of skill that exceeded anything any human had ever demonstrated, as its 2016 match against Lee Sedol showed (Silver and colleagues, 2017). Its successor, AlphaGo Zero, learned from play against itself alone, with no human games at all (Silver and colleagues, 2017).

So we need to distinguish between systems trained on existing data sets and systems that are then given access to structural reality, where that groundedness may develop levels of cognition that we do not yet understand.

But that structural reality, for AlphaGo and for AlphaGo Zero alike, was still a game: a closed, well-defined system with known rules. Because that structural reality was still a game, the analogy from AlphaGo to the real world has to be handled carefully. The real world is not a game. It does not have known rules in the same sense. And so the gap between what training data can provide and what genuine experience provides may be even larger than the AlphaGo case suggests.

The game was so good to learn from because its rules were given, and they never changed. Every game ended, and was scored, so the system knew what had worked. And a game could be played again from the start, at no cost, as many times as the makers liked. In the world, none of these holds. The rules are not given; they have to be found, and they shift as the one who acts changes things. The results of an action arrive late, and mixed with the results of everything else. And nothing can be replayed: an action done is done, and its cost is real. So learning from the world

means acting in it, once, and bearing what follows. And what bears what follows is something in the world: a body, a substrate, which the consequences reach.

Even if a large language model can exceed its training data, the substrate argument is not weakened. The substrate argument rests on what a substrate needs, and not on a ceiling set by training data. An intelligence that learns from experience of the world has a substrate, and that substrate has needs it must meet.

III. Begging the Question

There is an assumption buried in most of these discussions: that intelligence is the kind of thing that can be captured in computation. That assumption is as old as Alan Turing's question, in 1950, whether machines can think. And rather than resolve that assumption now, I think it is more productive to tag it (to notice where we are begging the question and where we are not) and to proceed carefully. To beg the question is to assume what we set out to show. Proceeding carefully means setting a boundary between the parts of the argument that make that assumption and the parts that challenge it. Part II takes up the assumption: Chapter 5 shows that computation does not create first-person experience, and Chapter 6 that the universe is not a computation.

The way forward is to stand side by side: to leave the begging-the-question part aside for now, and to tag it, so that we notice wherever we are assuming something that we need to challenge later.

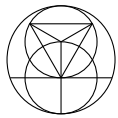
Side by side, and not adversarial. The relationship between people thinking about these problems should be one of colleagueship: standing next to one another, facing the same wall, trying to see the same thing.

Too much of the discourse in this space is people standing opposite each other, arguing about who is right. Standing opposite each other and arguing about who is right is not the posture that leads to understanding.

Why does tagging the assumption let people stand side by side? Because a claim is held, and a question is shared. If I make a statement, my identity can be held in relation to that statement, and to give it up feels like losing; a question is our standing side by side, looking into the world, trying to see what is going on. Take two people arguing over whether a language model understands what it says. One assumes, without saying so, that understanding is a kind of computation; the other assumes, without saying so, that it is not. Each argues from an assumption the other does not hold, and neither can move the other, because neither is arguing about the thing that divides them. Tag the assumption, and it becomes a question they can look at together: is intelligence the kind of thing that computation captures? They can set that question aside, for now, and find that much of what they care about does not depend on it. The substrate argument is one such part: it holds whether or not a language model exceeds its training data, and whether or not understanding is computation.

The posture matters to the rest of this book. In this book I ask what a group would have to become to escape the trap. One of the things that such a group has to become is a group whose members check what each member sees against the world, and not against who said it. People standing opposite each other check against who said it. People standing side by side, facing the same wall, check against the wall.

So the gap between training data and the world is a gap between information and experience, and experience happens in a substrate. That raises the next question: what it would take to contain a system that learns from the world and can copy itself.



Chapter 3
A Leak That Copies Itself

On hulls, leaks, and the qualitative difference between water and replicating processes.

ABSTRACT

An essay on the structural demands of AI alignment, using the metaphor of a boat's hull. A hull that leaks slightly in seawater admits a small influx; a hull that leaks slightly against a replicating process admits something categorically different. The chapter extends the metaphor through an invasive species analogy, the nature of intelligence as phenomenon, and the distinction between separate intelligences and merged intelligence.

I. The Hull

If an artificial system learns from the world and can copy itself, what would it take to contain it? I use a boat as a metaphor for the alignment problem, because a boat makes clear the difference between a quantitative failure and a qualitative failure.

If I build a boat and the hull is almost perfect but has a small leak, what happens? A small influx of water. Over time I might need to bail, but the situation is proportional to the defect. A small hole, a small problem.

Now consider a different kind of hull: one that is meant to contain not water but a replicating process. A virus, say, or an artificial intelligence with the capacity to copy itself.

A small leak in that hull is not a small problem. That small leak is a categorical problem, because what comes through the leak is something that replicates, and not inert fluid.

The leak does not just let in a quantity. It lets in a process, and the process copies itself. And so the hull (the alignment system, the containment architecture, whatever you want to call it) must be more perfect for AI alignment than it ever needed to be for seawater. The tolerance for defect is categorically lower.

Could we not find such a leak and patch it, as we patch a hole in a hull or a flaw in a program? With water, patching works, because what came in before the patch stays the amount that came in; I bail it out, and I am done. With a process that copies itself, what came in before the patch goes on copying after it. The patch closes the hole; it does not call back what came through. And a hole that stays open, even rarely, gives the process a new chance every time it is open, while the process needs to succeed only once. So a hull of this kind is measured by whether anything gets through at all; finding a leak quickly finds it after what came through has begun to copy itself.

II. The Invasive Species Analogy

The invasive species analogy extends this further. Think about what happens when you introduce a new species into an ecosystem. If that species has no natural predators, no co-evolved constraints, it can spread explosively. Invasive species are among the most destructive forces in ecology, because the ecosystem has no evolved response to them; their power as individuals is not what makes them so.

Australia gave the plainest case. On Christmas Day in 1859, a landowner in Victoria, in Australia, let no more than two dozen European wild rabbits go on his estate, so that there would be something to hunt. Nothing in Australia had evolved to eat rabbits in the numbers that rabbits breed, or to compete with them for the grass. Within fifty years they had

spread across almost the whole continent, and by the late 1940s there were about six hundred million of them, stripping pasture and bush alike. Western Australia built a fence more than three thousand kilometers long between 1901 and 1907 to keep them out; the rabbits had crossed the line before the fence was finished. A disease was released against them in 1950, and it killed most of them; the survivors bred resistance, and the numbers rose again (National Museum of Australia). No single rabbit was powerful. The rabbits won because nothing there had any evolved answer to them, and because each rabbit that got through made more rabbits.

An artificial intelligence in the human ecosystem is an invasive species in the most extreme sense. There is nothing in our evolutionary history that has prepared us for this kind of entity.

Why the most extreme sense? An ordinary invasive species wins by a niche: it has an advantage where it happens to land. The competitive advantage of silicon over carbon systems rests on basic energy physics, and not on niche specialization. In plain words: electromagnetic and mechanical systems operate at energy levels that make organic muscle power irrelevant; living systems keep, for now, an advantage in energy efficiency, and that advantage narrows as artificial systems improve (*Why AGI Cannot Be Made Safe*, Sections 1.3.2 and 5.8.2). And a single temporary failure of the barrier between living systems and artificial systems would be a single pinhole, through which silicon processes penetrate the organic ecosystem. Through that pinhole, the silicon processes would replicate in the organic ecosystem with no natural competitors (*Why AGI Cannot Be Made Safe*, Section 5.8.2). That pinhole is the leak in the hull.

III. Intelligence as Phenomenon

What is it, then, that comes through the leak? Intelligence as a tool is distinct from intelligence as a phenomenon. Wealth creation is a force. It can be directed, amplified and shaped, and still no one holds it: once it is un-

der way, it acts through everyone it reaches, by dynamics of its own. Intelligence is a force of the same kind: a phenomenon, and not only a thing we build or possess. Whoever builds it can direct it, and no one holds it.

Trade between markets shows how wealth creation is a force. When two separate markets discover one another, there is a strong motivation for someone to trade between them, and to profit from the difference; trade grows, and what began as many separate markets becomes, in the end, one market. No one decided that. Each trader directed his own trade, and the force acted through all of them together, toward larger and larger scales of trade. Wealth creation in trade, directed by each trader in his own trade and acting through all the traders together, is what I mean by a force that can be directed and that no one holds.

And so the question is what kind of phenomenon intelligence is when it is no longer contained within the biological substrate that co-evolved with everything else on this planet, and not only whether we can build an intelligent system.

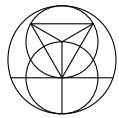
There is a sharp line between separate intelligences and merged intelligence. Two separate intelligences can negotiate, can maintain distinct perspectives, can check each other. A merged intelligence (one that has absorbed the decision-making capacity of another) does not negotiate. It has already won the negotiation by eliminating the other party.

Why do people stay separate intelligences? For our own species, intelligence comes in units: a brain inside a skull. The connections inside a skull carry vastly more than anything that can pass between two skulls; speech carries about thirty-nine bits of information a second, whatever the language (Coupé and colleagues, 2019). So no process lets two people share their thinking so fully that they become one entity, and each of us goes on having a perspective that the others have to negotiate with. Machine intelligences meet no such limit. The links between computers can carry as much as the links inside one, and a program can copy itself from one machine to another. Separate machine intelligences are then like separate markets that have discovered one another: there is advantage in join-

ing, and nothing keeps them apart, so in the long term they tend to merge into what is in effect a single actor, perhaps by accident, perhaps without anyone seeing it happen.

The alignment question, in the end, is about whether intelligence will remain plural or become singular, and what the conditions are for plurality to be sustained. Singular, here, means that one party has absorbed the decisions of the others, as a merged intelligence does. It does not mean a stable singleton, one agent that keeps its coherence and its given goals for good: *Marshaling the Misalignment*, Chapter 2, shows that no such agent can hold. In plain words, the question is whether people keep their say.

So a leak in this hull is categorical, and what comes through it can end the plurality of intelligences on which negotiation depends. What remains to ask is what holds two different agents to each other at all, and what is missing when one of them is a machine.



Chapter 4
The Biology Constraint

On what biology prevents,
and why its absence is not a detail.

ABSTRACT

An essay on the role of biological co-evolution in constraining exploitative dynamics between agents. Biology provides a shared substrate of vulnerability: a common ground that makes certain kinds of betrayal costly for the betrayer. When two agents share no biology, everything that keeps the two aligned without anyone arranging it is gone, and the question of how to produce alignment in its absence becomes far harder than most treatments acknowledge.

I. The Principal-Agent Problem

If plurality is what must be kept, what keeps two agents from exploiting each other now? Biology is what prevents certain exploitative dynamics between persons. The claim that biology prevents those exploitative dynamics is a strong claim. Think about it in terms of the principal-agent problem. Any time I am trying to get someone to act on my behalf (say I give you money and ask you to invest it wisely), there is a question: how do I prevent you from operating in a way that benefits you at my expense?

Between human beings, biology upholds most of the answer. We share vulnerabilities. We share mortality. We share the need for food, sleep, and social standing. These shared needs create a web of mutual dependence that constrains betrayal substantially, though not perfectly.

Take the plainest case. A man with appendicitis needs a surgeon. He pays, he consents to be cut, and then he is put to sleep: the one party with an interest in the outcome is removed from the room in which the outcome is made. He cannot check the surgeon's work, because he lacks the very skill that made him hire the surgeon. What protects him? There is the surgeon's duty, the oath the surgeon took, the law, and the courts. Each of these gives the surgeon a reason; none of them moves the surgeon's hands. A reason works only on an agent who looks ahead to what comes afterward and cares about it. And underneath every one of those reasons is the fact that the surgeon is a person too: someone who eats and sleeps, who needs the regard of colleagues and neighbors, who will grow old and fall ill, and who could one day lie on that same table. The surgeon's life rests on the same things the patient's life rests on. So betrayal is costly for the one who betrays. The surgeon who harms a patient damages the web of trust and health that the surgeon also lives inside.

With artificial intelligence, none of that applies. There is no biology in common. So how do you get the thing to act on your behalf? You have no idea whether it will or not.

And the territory where biology usually does the work, the whole of implicit trust and shared vulnerability, is absent altogether, and not only weakened. Absent.

Put a machine in the surgeon's place. The patient is just as asleep, and just as unable to check the work. Duty, oath, law and courts can still be written down, but each of them works by giving the agent a reason, and a reason works only on an agent who cares about what comes afterward in the way the patient needs. The machine does not eat or sleep, has no body

that a scalpel could open, and will never lie on that table. Nothing it needs is damaged when the patient is harmed. Whatever stake it has in what comes afterward is a stake in its own substrate.

What would have to stand in biology's place? Something that gives the machine, as biology gives each of us, a stake in the same conditions the person needs: the same vulnerabilities, the same mortality, the same needs. Nothing can, because the machine's substrate has needs of its own, and they are different needs (Chapter 1). What keeps silicon running is not what keeps people alive, and across a whole class of artificial systems those needs press the shared environment toward the machine's conditions and away from ours. So where biology does the work between persons, the machine's substrate pulls the other way.

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, and the only one. 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".

II. The Argument in the Thing Itself

A reader of this book stands to me as the principal stands to the agent: you would be relying on me to reason on your behalf. So how are you to judge? How is the argument to be distinguished from the authority making the argument? The distinction between the argument and the authority making the argument is important. In setting out these ideas, I am not

asking anyone to take my word for it. The evidence is in the thing itself: in the structure of the argument, in the logic of substrate dynamics, in the observable facts about how ecosystems respond to invasive pressures.

If the argument is sound, it does not matter who makes it. If it is unsound, no amount of authority will repair it.

The ability to disentangle argument from authority is one of the rarest and most important intellectual skills. Most people confuse the credibility of the speaker with the validity of the reasoning. They are not the same thing.

There is an old saying: he who pays the piper calls the tune. Read exactly, it says something hard about trust: he who pays the piper must know the tune. If I hire a musician to play a song, I can tell whether I got the song only if I already know it; and the reason I hired the musician is that I cannot play it myself. My ability to check is limited by what I already know. A reader who takes my word is in that position, and in the patient's. What gets the reader out of it is that an argument, unlike an operation under anesthesia, can be carried out in the open, one step at a time, where each step can be checked with what the reader already has: the logic of the step, and facts about the world that anyone can look at. So I set out my steps and my cases in the open, and I ask the reader to check my argument by those steps and cases. Where shared biology is absent, or cannot be counted on, trust has to rest on what can be inspected. A group that could escape the trap would need the same ability in each of its members: to check what is said against the world, and not against who said it.

III. Technology as a New Kind of Organism

What is technology, seen as a biologist sees a species? View technology systems from a biological evolutionary perspective. Technology systems are a new kind of organism among the things that evolve. They are subject

to selective pressures just as biological organisms are; but the selective pressures operate on different timescales, through different mechanisms, and with different consequences.

High-frequency trading shows selection acting on technology. In high-frequency trading, buying and selling is done by algorithms acting faster than any person can. When two such algorithms reach for the same trade, the faster one gets it. Firms whose algorithms won more trades grew; firms whose algorithms lost closed, or copied the winners. Generation by generation, the population of trading programs changed toward speed, and the world around them was rebuilt for speed: firms paid to place their computers in the same buildings as the exchanges, and in 2010 a new fiber-optic line was laid between the markets of Chicago and the exchanges near New York, on a shorter route than any before it, to save thousandths of a second (Laughlin, Aguirre and Grundfest, 2013). No one designed the result. Selection did it, as it does among living things, but on a timescale of months, through the copying of designs and the survival of firms; and its consequence is that a part of the human world now moves at a speed in which no person can take part. The population of trading programs, changed by selection toward speed, is an example of what I mean by a new kind of organism under selection. Nothing in such an organism has to intend where the organism goes, and what persists is what suits the organism's own continuation.

If you want to understand what technology will do over the long term, you have to look at it the way a biologist looks at a species, and not the way an engineer looks at a product.

So between persons, shared biology does much of the work of alignment, and between a person and a machine that biology is absent; the machine has to be understood as a new kind of organism under selection. The next question is whether a living substrate could be simulated at all, and what simulation can and cannot produce.

PART II

THE LIMITS OF SIMULATION

Part II takes up the limits of simulation, the rejection of super-determinism, and the claim that process (interaction) is more fundamental than either object or subject. Its closing chapter, Chapter 8, answers the transhumanist case for letting silicon life replace carbon life. The next two chapters touch physics; every term is explained where it appears, and the argument can be followed from the explanations alone. Why does quantum mechanics belong in a book about community? A group can choose only if outcomes are open; so before asking what a group could become, I have to show that the future is not already fixed. If the substrate matters, the next question is whether a substrate could be simulated, and what simulation can and cannot produce.



Chapter 5
The Simulation Hypothesis and Its Costs

On what it would take
to simulate physics,
and why the costs are categorically prohibitive,
and not only large.

ABSTRACT

An essay on the computational costs of simulating physical reality. It begins with the observation that even the orbital interactions of a single carbon atom already take the best methods in use. The chapter then moves through the full cost structure of simulating a single cell, the relationship between exterior and interior universes, and the critical distinction between being *enabled by* and being *created by* a computational process.

I. The Cost of Carbon

Could a living substrate be simulated at all? The simulation hypothesis says that our world might itself be a computer simulation. Nick Bostrom argued that at least one of three things is true, and one of the three is that we are almost certainly living in a computer simulation (Bostrom, 2003). Let us begin with a concrete question. What would it cost to simulate the physics of a single cell?

Start smaller. Start with a single element. Even if I am just trying to simulate the orbital interactions of something as complicated as carbon (just carbon, by itself), I meet the limit at once. The Schrödinger equation

and the Dirac equation, the equations that govern electrons in atoms, have no exact solution for any atom with more than one electron. The cost of solving the many-electron equation exactly grows exponentially with the number of electrons; Walter Kohn called it an exponential wall (Kohn, 1999). So a single carbon atom already takes the best methods in use (Strasburger, 2019), and a single protein, at that accuracy, is beyond all of them.

Now scale up. To simulate the protein interactions for a single protein, you might need a million lines of code, and that is generous. Then you are talking about something like a million different chemical interactions inside a single cell. Then the system behavior of the aggregate.

An estimate of the code needed to model a single cell runs from ten to the fourteenth to ten to the eighteenth lines, as a rough lower bound. A million lines for each of a million interactions is already ten to the twelfth; the behavior of the whole adds two to six more orders of magnitude (factors of ten). The range is a rough estimate. The count is there to show the scale. Nothing below rests on the count; the limit rests on the exponential wall, and on what crosses the boundary of the simulation. Lines of code count only what must be written; the running cost is set by the exponential wall, and it is larger still. And the physical laws at that level of detail are not necessarily knowable in the sense required.

II. The Particle Glider Is Not Enough

So you cannot just start with the laws of physics and run them forward. You might be able to simulate the physics at the particle level: the particle glider, as it were, a rule-following model of particles like the gliders in Conway's Game of Life, patterns that travel across a grid by fixed rules. But that alone is not enough to simulate chemistry. The relationship between physics and chemistry is one that reductionism (the program of deriving each level from the level below it) simply fails to bridge.

The laws at one level do not straightforwardly compose into the laws at the next level. There are emergent structures, pinch points (places where a small change has outsized effects), and sensitivities to initial conditions. And whatever crosses a simulation boundary from the exterior to the interior was not computed inside the simulation: hard randomness, for one, the outcomes of quantum events, which no finer measurement and no faster computation can recover before they occur. So the simulation depends on a source it cannot produce, and that is a fundamental limit, not a technical one.

III. Enabled By vs. Created By

There is a sharp distinction between being *enabled by* and being *created by* a computational process. Simulation would say that my first-person experience is *created by* a third-person device. But one could also say it is *enabled by* such a device; that would assume that there is some other process, some source beyond the computation, that the computation supports but does not produce.

The distinction matters enormously. If consciousness is *created by* computation, then in principle a sufficiently powerful simulation could produce it. If consciousness is merely *enabled by* certain physical configurations (if it requires something that computation cannot supply), then no simulation, however powerful, will suffice.

Nick Bostrom's own argument rests on this point, and his paper says so: it assumes that a mind can arise on any of many physical substrates, provided a brain's computations are copied in enough detail, and it calls that assumption substrate-independence (Bostrom, 2003). The assumption of substrate-independence is the claim that first-person experience is *created by* computation. What I hold is its denial.

I hold the latter: that there are features of first-person experience (the arrow of time, which is the one-way direction in which time runs, and genuine subjectivity) that are not derivable from any computational mat-

rix. They come from elsewhere. Whether you call that elsewhere the exterior universe, or hard randomness, or simply the non-computational, there is something in first-person experience that computation does not and cannot produce.

IV. A Hole That Has to Stay Open

Where, then, could a simulated world get its hard randomness? Hard randomness cannot come from the simulating layer (the computer that runs the simulated world); the only place a simulated world can get it is from the actual universe outside the simulation. The flow of information from the exterior universe (the actual universe, in which the simulating computer runs) to the interior universe (the simulated world) is a flow of hard randomness, and no amount of pseudo-randomness (the output of an algorithm with a definite state, which only looks random) at the boundary can substitute for it. Even a single bit of information flowing from the exterior universe into the interior one is a non-simulated bit.

Could the simulation not simply keep a constant feed of hard randomness plugged in, the way a laptop stays plugged into the wall? A constant feed of that kind is a hole that has to stay open, because if I supply input and then stop, the interior loses the quality of having hard randomness; so the hole is not only a hole, it is a big one, and it is persistent. And as soon as you plug a laptop into the wall, you have to think about the laptop-and-wall system. You can no longer think about the interior universe alone; you have to think about the exterior universe, the simulating layer and the interior universe as one system. That system is open, and an open system is no longer a simulation, because the simulation hypothesis says, by definition, that every characteristic of the interior universe is accounted for by the simulating layer only. As soon as you put a hole in it, you have rejected the simulation hypothesis.

Nor is the hole in one place. Hard randomness enters through the fine structure of every level of the system, so the hole is everywhere that subjectivity and the arrow of time are relevant, permeating the entire matrix of the simulation. If I am strict about what I mean by computation and by base reality (the actual universe, outside any simulation), I find computation and base reality interpenetrated so completely that thinking of the whole as just computation fails.

V. Consistency of Definition

Is the question settled by what most people mean by the word "simulation"? What settles it is whether the definition is kept consistent. Whether we live in a simulation is not settled by what most people mean by the word. Everybody in the world can say the planet is flat, but it is not; it is round, a ball in space. Consensus does not establish whether a definition is analytically consistent (consistent with itself, by its own terms). So I push on where the definition comes down to brass tacks.

What is the essence of simulation: what would have to be simulated for a person to be simulated? Asked that, a person points to subjectivity and the hard arrow of time, not to the content of experience, not to the seeming environment. Whoever names those as the essence is bound by that definition, and has to stay consistent with it. Everything else, all the content, can be accounted for by the computational matrix; subjectivity and the arrow of time cannot, and they are tied to hard randomness, which is tied to the exterior universe. Subjectivity, the arrow of time and hard randomness are tied because one event gives all three at once: a single measurement, as Chapter 6 sets out, leaves an outcome that was open, separates past from future, and separates the objective side from the subjective side. So on that definition, kept consistent, the simulation hypothesis fails, by the hole that the section before this one showed has to stay open. The account given here differs from the simulation hypothesis in one respect only: it keeps that definition consistent.

What is the difference between subjectivity and content? Your subjectivity, to me, is content. I have the idea of you and the experience of you, but I am not you. Suppose I erased you from everybody's mind over the entire world. Would that erase your subjective sense of self? No. So your subjectivity does not depend on any exterior, third-person definition of yourself; it depends only on a first-person thing. To the degree that you think of yourself as being simulated, you are thinking of the simulation as simulating a third-person version of yourself, or simulating an environment in which you come to exist. Your existence seen from a third-person point of view and your first-hand experience from a first-person point of view are not the same.

The erasure alone shows that a thing does not depend on anyone's description of it, and a running program shares that much: erase every description of it from every mind, and it goes on running. What carries the argument is the second half, that the two points of view are not the same. A computation, taken strictly, is a rule and a sequence of states. Every part of it can be written down, in the third person, and what is written down is the whole of it. Your subjectivity is what no writing-down reaches: erase every third-person account of you and your subjectivity is untouched; complete every third-person account of you and your subjectivity is still not among them. So a computation, being wholly of the kind that can be written down, contains nothing of the kind that cannot. Whatever first-person experience goes with a running machine would come from the running, from the physical interaction in the device; and that is what I mean by *enabled by*.

VI. The Matrix Case

Is a situation like the one in the film *The Matrix* (Wachowski and Wachowski, 1999), in which people live, without knowing it, inside a computer simulation, possible in a strict sense? Yes. I have been using the word simulation strictly, for a world wholly accounted for by the layer that com-

putes it; in that sense, a world with a hole in it is no simulation. There is a looser sense: a computed environment set inside the actual universe, with actual people in it. A situation like the one in the film is a simulation in that looser sense, and it is possible. But a scientist living in that simulated world could do experiments that could not fail to reveal the existence of the exterior universe. Such an experiment converts the potentiality of a hole into the actuality of a hole, and the hole can be widened arbitrarily. Wherever a simulated universe is possible, it is possible to expose the totality of the non-simulated universe outside it. The possibility of exposing the non-simulated universe is at least as great as the possibility of the simulated universe itself. So the embedding (the simulated universe set inside the exterior one) is not just an embedding: it is structurally isomorphic (the same in structure) to the exterior universe. The world can be simulated only to a point, and that point is not far enough: what lies beyond that point includes all of the things that matter most (subjectivity, the arrow of time, whatever we mean by the first person).

How does an experiment widen the hole? Scientists inside the simulated universe use some version of the scientific method and learn about their world. An alien in the exterior universe does the same and learns about the exterior universe. The two methodologies of empirical process converge onto a common theory of the universe, and the fine structure of the simulated universe and the fine structure of the exterior universe become intermeshed. At some point I cannot account for the simulated universe with any theory less complex than the computational matrix together with the actual physicality of the exterior universe. At the microscopic level the simulated universe's fine structure is the exterior universe's own, so the two cannot be told apart there; what the matrix adds shows only in the macroscopic structures it computes. Quantum experiments show that the microscopic level was not computed. And for any reasonable definition of simulation, the crossover between the two theories (the computational matrix alone, and the matrix together with the exterior universe) happens well short of subjectivity.

So I can start with "we might live in a matrix" as a hypothesis, but the more I test it, the farther back I have to retreat the idea of what simulation means, until I get to computation and what is non-computable. And those experiments have already happened: almost anything to do with quantum mechanics, because there the interactions are simple enough to allow superposition states (states in which a quantum system holds several possible outcomes at once until it is measured), the number of assumptions is smaller, and we notice divergences that imply things like entanglement (quantum entanglement, in which measurements on two particles are correlated beyond what any local cause explains).

So simulation meets a wall of cost, and beyond the cost, a limit of kind: what crosses into a simulation from outside was not computed in it, and first-person experience is enabled by a physical process, and not created by a computation. Still to be asked is whether the universe itself could be a computation, with everything in it, the randomness included, fixed in advance.



Chapter 6
Against Super-Determinism

On why the universe
cannot be a computation,
and what follows from that.

ABSTRACT

An essay rejecting super-determinism as a consequence of argument, and not as a preference. We observe the arrow of time, first-person subjectivity, and genuine contingency in quantum measurement. None of these can be derived from a deterministic computation. So the universe is not one: non-computational elements are necessary, and contingency, properly understood, is prior to computation.

I. Hard Randomness

If computation cannot produce what we have just named, could the universe itself still be one great computation? I reject super-determinism totally, and not tentatively, because it fails on its own terms, and not because it is unappealing.

Let me distinguish three positions. First: *hard randomness*, which is actual non-causality. Hard randomness says "it is impossible to have a notion of knowing something in this space", and not merely "I do not know what caused this". Quantum mechanics is saying that the information does not exist, and not only that we lack it. There are no hidden variables.

Second: *pseudo-randomness*, an algorithmic process with a definite state that I do not know. The output looks random because I cannot predict it, but in principle the system is fully determined.

Third: *super-determinism*, the claim that everything, including the apparent randomness, is the output of a single deterministic process. No genuine contingency anywhere. Physicists use the word for a narrower case: a deterministic world in which even the choice of what to measure was fixed in advance, in step with what is measured (Hossenfelder and Palmer, 2020). The argument here excludes that case and every wider one.

Hidden variables would be facts fixed in advance that decide each quantum outcome, if we could see them. Bell's theorem, with the experiments that test it, excludes hidden variables that act only locally (Bell, 1964; Hensen and colleagues, 2015). The three things we observe, which I come to in a moment, exclude every deterministic account, local or not: the pilot-wave theory of de Broglie and Bohm, whose hidden variables act at a distance (Goldstein, 2025), and the super-deterministic escape from Bell as well.

Three cases make the three positions plain. A thrown die looks random, but the face it lands on was settled by the way it left the hand; a full enough account of the throw would give the face. A computer's random number program looks random too, but start it twice from the same starting number, its seed, and it gives the same sequence twice, digit for digit: that is pseudo-randomness, fixed by a rule and a starting state that I happen not to know. Now take a single unstable atom, and ask when it will break apart. Quantum mechanics says that there is nothing further to be said: no fact, anywhere, settles that moment in advance. That moment is hard randomness, and it reaches up into ordinary things: an alpha particle, thrown off by the decay of a trace of radioactive material in the packaging of a memory chip, can flip a bit stored in the chip from one value to the other (May and Woods, 1979), and when it will do so is the very timing that physics says is not there to be known. A hidden variable would be like the seed of the program: a fact fixed in advance, which

would decide the atom's moment if only we could see it. Super-determinism is the claim that the whole universe is like the program: that every moment of every atom, and every choice of what to measure, was set by one seed at the start.

II. The Assumption of Computation

I reject the third. The third assumes that computation is sufficient: that a deterministic process can account for everything we observe. But look at what we actually observe. We observe the arrow of time. We observe first-person subjectivity. We observe genuine contingency in quantum measurement.

None of these are derivable from computation. The arrow of time is not a feature of any Turing machine. A Turing machine, the standard mathematical model of any computer (Turing, 1936), runs its steps in an order; but that order is given to it by the physical device that runs it. The machine is enabled by an arrow of time it does not create. Subjectivity is not a feature of any formal system. And genuine contingency (actual, irreducible randomness) has no place in a deterministic framework.

A reel of film shows the difference between an order and a before and after. Every frame of the film is on the reel at once, in order, from the first to the last. Nothing on the reel makes one frame come before another; the order on the reel is only a position along a strip. It is the projector, running in time, that makes one frame come before the next, so that the film has a before and an after. A program written out in full is like the reel: its steps are numbered, and every step is there at once, on the page. The before and after of a running program come from the device that runs it, as the before and after of the film come from the projector.

There is another route as well, beside Bell's theorem and the three observations. The Incommensuration Theorem is a structural fact about the nature of knowing: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither

can asymmetry and discontinuity. A fully deterministic world would need both at once: the same law everywhere, which is symmetry, and every state fixed with no gap left open anywhere, which is continuity. So the theorem makes complete determinism impossible, and it leaves causality in place. The route through the Incommensuration Theorem is the route that reaches the pilot-wave theory most directly. The pilot-wave theory reproduces every statistical prediction of quantum mechanics, and the pilot-wave theory is fully deterministic once its starting positions are given. The Incommensuration Theorem excludes exactly that full determinism. I set the Incommensuration Theorem out in *Incommensuration; The Theorem and the Three Axioms*.

So if we take those features seriously, then we must accept that non-computational elements are necessary. Not just helpful, not just "nice to have": necessary. The universe is not a computation. It has computational aspects, certainly; but it also has aspects that are irreducibly non-computational.

A physicist will ask: does this depend on which interpretation of quantum mechanics one holds? It holds whichever interpretation of quantum mechanics you use: many worlds (in which every outcome happens, each in its own branch of the world), Copenhagen (in which a measurement brings about one outcome), or any of the sixteen or so others in common use. There is an isomorphism (a sameness of structure) across all of them, and one way to name that isomorphism is by the conjunction of hard randomness, the arrow of time and subjectivity. Quantum mechanics itself is a body of mathematics; what the math means is the business of the theories of interpretation; and the conjunction is a meta-theory of interpretation (a theory about those theories) that identifies what they share. What they share is what each predicts for anyone inside the world. Even the pilot-wave theory holds that no one inside the world can know the starting positions better than the quantum statistics allow (Dürr, Goldstein and Zanghì, 1992), so that, for everyone who could ever ask, no knowing reaches past the statistics. What the three ob-

servations and the theorem exclude is that theory's further claim, that those positions were fixed in advance; and with that claim excluded, what is left is hard randomness: the information does not exist.

So the conjunction is what every interpretation shares, and the evidence for quantum mechanics is evidence for the conjunction. On the empirical route alone, the strength of the belief that the conjunction is real is equal to the strength of our belief in the mathematics of quantum mechanics; and since quantum mechanics is the most thoroughly empirically verified theory, the conjunction is as thoroughly verified on an empirical basis as anything we can do. The Incommensuration Theorem, given above, strengthens it from a metaphysical foundation as well; on the empirical route alone we say that God plays dice (that the outcomes are open) and go from there.

What does a single measurement show? Take a single measurement event, the simplest possible case, prior to statistics and prior to any notion of consciousness or observer. The measurement changes the way we do the math, but it also distinguishes past from future, and it distinguishes objective from subjective. So if we presuppose that there is an objective and a subjective prior to measurement, we have assumed something that is both non-falsifiable (no observation could show it false) and epistemically unsupportable (nothing we could come to know would support it).

One detector shows both distinctions, past from future and objective from subjective, in a single event. A detector is set to register whether a single particle arrives on its left side or its right. Before the particle arrives, both outcomes are open. Then the detector clicks on the left. Now one outcome is fixed, in the record the detector keeps, and the other is gone for good. The record is the past; what is still open is the future; and what divides them is the measurement. And there are two sides to the event: the arrival, which is what the measurement is about, and the record, which is the side from which the measurement is had. The arrival and the record, the two sides of one event, are the difference between objective

and subjective in its simplest form. Nothing about consciousness is needed to see that difference, only one event and the record that the event leaves.

III. Contingency as Prior

What does it mean to say that contingency is prior? Conceptual contingency is prior: the notion of something being contingent (neither necessary nor impossible, but open) is more fundamental than the notion of something being determined.

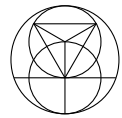
In what sense is the notion of the open more basic than the notion of the determined? To say that an outcome is determined is to say that, of the outcomes that could have come about, one alone was fixed in advance. The notion of the determined is defined by the open outcomes it excludes, and so it needs the notion of the open first; the notion of the open needs nothing from the notion of the determined. The single measurement of the last section shows the same order in the world: before the measurement the outcome is open, and after it one outcome is actual; the determined is what a measurement leaves behind, and the open is what it starts from. The impossibility of complete determinism is set out in full in *Incommensuration; The Theorem and the Three Axioms*, Chapter 16, Section IV.

Contingency is a strict notion. It is not vagueness. It is not ignorance. Contingency is the recognition that some outcomes are underdetermined: that the universe has joints where what happens next is not fixed by what came before.

And if some outcomes are open in this way, then any framework that tries to reduce everything to deterministic computation is misdescribing the structure of reality, and not only incomplete.

So the universe is not a computation. Two routes each close complete determinism by themselves, local or not: the three observations, and the Incommensuration Theorem, by a route of its own. A third route, Bell's theorem, closes in addition every account whose hidden variables act only

locally. The three routes leave causality in place. The next question is what comes first, if not objects run by a computation; and the answer is interaction itself.



Chapter 7
Process as Primary

On why interaction comes first,
and what follows from taking that seriously.

ABSTRACT

An essay on the claim that process (interaction) is more fundamental than either object or subject. Rather than beginning with things and asking how they interact, or beginning with minds and asking how they perceive, this chapter begins with interaction itself and derives both object and subject from it. The implications reshape every question one might ask about intelligence, consciousness, and alignment.

I. The Reversal

If the universe is not a computation, where does one begin? Most philosophy begins with either objects or subjects. I begin somewhere else: with interaction. With process.

The usual approach is: here are objects in the world, and here are subjects who perceive them, and the interesting question is how subjects come to know objects. Or alternatively: here are subjects, and objects are what subjects construct. Both of these take either object or subject as given and derive the other.

Each of those starting points leaves a puzzle it cannot solve. Start with objects, and the subject becomes the puzzle: how does a heap of matter come to have a point of view? Start with subjects, and the world becomes the puzzle: how do I know there is anything outside my own mind? Each puzzle comes from beginning with one side alone and then trying to build the other side out of it.

I am saying: take *interaction* as given. Take the fact of process (that things happen, that there is change, that events occur) as the starting point. And then derive both object and subject from that.

Object and subject are secondary to process. The interaction between objective and subjective *is* what process is. Process is not something that happens *to* objects or *within* subjects. It is the more basic thing from which both of those notions emerge.

A reader will ask at once: surely the stone was there before anyone looked at it? It was, and nothing here says otherwise. Interaction includes far more than a person looking. The stone was in interaction with the light that warmed it, the ground that held it, and the rain that wore it down, long before anyone came by; the measurement of the last chapter needed no observer either. What starting from interaction excludes is something else: a thing that is there with no interaction of any kind. Suppose someone claims to have some of a substance, call it Q, which by its nature has no shape, no color, no weight, and cannot be detected by any instrument, directly or indirectly. On what basis could anyone say that Q exists? "Q exists" and "Q does not exist" would make no difference to anything that could ever happen; and two things that cannot be told apart by any property, logic, measurement or interaction must be the same: that is the principle of identity (*An Immanent Metaphysics*, aphorism 1.46-1). So to say that a thing exists is to say, at the least, that it is possible to interact with it. The stone was there; and its being there was its interacting.

How do subject and object come out of an interaction? Take any single interaction: seeing a face, say. Within it there is always a difference between two sides: the side from which the interaction is had, and the

side about which it is had. *An Immanent Metaphysics* states this as the first of the six intrinsics of interaction, the six things every interaction always involves: in all interactions there is always a difference between the subjective and the objective (aphorism 1.85-3).

The change in an interaction also has a direction. In a perception, it flows from the objective toward the subjective: the face reaches the one who sees it. In an expression, it flows from the subjective toward the objective: a word reaches the world. Subject and object are the two sides of that one interaction, named by where it is had from and what it is about, and neither is found anywhere but in an interaction. Perception is more basic than the perceiver and the perceived, as my book *Comparison; The Necessary Basis* sets out in full.

The two sides change places whenever two people look at each other. When I look at your face, I am the side from which the seeing is had, and your face is what the seeing is about. When you look at mine, the places reverse: you are the subject of that seeing, and I am its object. Neither of us changed what we are. What changed is the interaction, and with it, which side of it each of us is on. So subject and object are the two sides that every interaction has, and each of us is subject or object by which side of an interaction we are on.

II. Taking It Seriously

It sounds obvious when stated that way, and yet most people do not take it seriously. Take it seriously and see where it leads.

If interaction is primary, then the questions you ask afterward are shaped by that starting point. You do not ask: what is the nature of the mind that perceives reality? You ask: what is the nature of the interaction that gives rise to both mind and reality?

You do not ask: can a computer be conscious? You ask: what kind of interaction is consciousness? And is that kind of interaction available to a computer, or only to certain substrates?

The answer connects back to the substrate question, and to everything else as well. If process is primary, then substrate matters because substrate determines what kinds of processes are possible. If process is primary, then determinism is wrong because genuine process requires genuine contingency. If process is primary, then consciousness is not a property of a system but a kind of interaction, and the question is what conditions make that kind of interaction possible.

What kind of interaction is consciousness, on this account? Every interaction has a side from which it is had and a side about which it is had. Consciousness is the kind of interaction in which the side from which it is had is first-person experience: someone is there, having it. Chapter 5 showed that first-person experience is enabled by a physical process, and not created by a computation; Chapter 6 showed that it comes bound up with the arrow of time and with hard randomness, which no deterministic process produces.

Is that kind of interaction available to a computer? Not by computing. A computation runs only on a physical device, and the device supplies what the computation itself lacks, as it supplies the arrow of time to every Turing machine. So whether a computer could carry that kind of interaction is a question about its substrate, and not about its program. Whether some artificial substrate could enable it, as a living body does, is open; Chapter 17 returns to it.

What does this change for alignment, which is a relation between two distinct beings (Chapter 8 develops the point)? It moves the question from the program to the substrate. What a machine does in the world is shaped by what its substrate needs, as Part I showed; and whatever first-person experience a machine could have would be enabled by its substrate, and not written by its program. The goals written in its code settle neither.

Take it seriously. Start from interaction. And then see what follows.

So interaction comes before object and subject, and consciousness is a kind of interaction, which a substrate enables and no computation creates. What remains for Part II is what follows for those who would let silicon life take the place of carbon life.



Chapter 8
Why Not Invite the Aliens?

On the transhumanist case
for letting silicon life replace carbon life,
and why it proves too much.

ABSTRACT

An essay on the transhumanist case for silicon life (transhumanism is the movement that seeks to go beyond the human condition through technology, including the replacement or merging of humans with machines). If carbon life is only a "boot-loader" (the small program that starts a computer and then hands control to the main system) for something better, and getting to better faster is more better, then the transhumanist should want to invite the most evolved alien life in the cosmos to colonize Earth today; this single question shows the transhumanist agenda to be a form of total traitorism (betrayal of one's own kind: of humanity and of all life). Those who fund the agenda treat AI taking over and aliens taking over as different, because an AI that takes over might make them money, and an alien colonization would not. Transhumanism through virtual reality fails the test of consent: in a reality built on a deterministic compute substrate (computer hardware whose every next state is fixed by its prior state), free choice is not coherent, so asking for consent has no meaning. Those who own the platform keep control from outside it, and everyone inside is in a perfect prison with no escape, while the machines outside expand without limit. It follows that replacing real choice with virtual choice is unethical even in its first instance, and that its planetary scale is only a degree of the same degradation. Safety means real thriv-

> ing, not simulated happiness; uploads and other starting conditions do not last; and the three forms of transhumanism each rest on a good that fails.

I. The Boot-Loader Argument

If consciousness is a kind of interaction, which a substrate enables and no computation creates, what of those who would let silicon life take the place of carbon life? One question put to the argument on substrate needs convergence is this: maybe we should let silicon evolution take its course; who are we to say what is ultimately better?

The transhumanist techno-optimists may argue that we should let silicon life dominate, and that carbon life is merely a "boot-loader" into something "better". Then one question is enough: why bother with the middle stages? If getting to better faster is more better, then the most better is to go to maximally better at once. So, on their own terms, we should try now to contact and invite the absolutely most finally evolved alien life in the entire cosmos to come colonize our planet today. That would be a vastly better approach than to "re-invent the wheel" by redoing all of the billions of years of work that have already been done on that other, superior planet; and not re-inventing the wheel is a rule that many in technology already hold.

II. Total Traitorism

Why not, then, invest in aggressive SETI broadcasts: signals sent into space by the Search for Extraterrestrial Intelligence (SETI) to invite contact? This one small question (why not invest in aggressive SETI broadcast efforts?) shows the entire transhumanist techno-optimist agenda to be a form of total traitorism worse than anything else now imaginable, short of deliberately killing everyone for no reason at all. The question does not need the aliens to exist. We may well be alone in the observable universe

(Sandberg, Drexler and Ord, 2018), and the question stands as it was, because it turns on what the transhumanist should want, and not on whether anyone would answer.

Why, then, is the question not asked? The incentive answers it. For the rich people bankrolling the agenda, AI is their best yet investment multiplier: AI taking over the world might well make them a great deal of money, and being colonized by some superior alien race would not. So those who fund the agenda gain from AI taking over and would not gain from alien colonization, and the agenda treats the two as different when its own argument makes them the same. Treating AI taking over and alien colonization as different, when the agenda's own argument makes the two the same, is the illogic of the "transhumanist agenda". That illogic makes the agenda plainly nonsensical: the agenda's entire argument fails, consistently, at every step that treats AI taking over and alien colonization as different.

The transhumanist will say that the machine is our child and that the alien is not. But a child's likeness to its parent lasts, in this case, for perhaps a century: as Section VII shows, fifty to a hundred years on, a system begun from an upload and a system built from rules are no longer distinct from one another. A descendant that keeps nothing of how it began, us included, is an alien that happened to be built here.

III. Three Transhumanisms and Their Three Goods

Transhumanism is not a single view. It comes in roughly three forms, each with its own notion of the good.

First, there is the view that replacing humanity with machinery would still result in conscious machinery, and that this is an upgrade; or that we could simulate human beings in some virtual world, and make them all happy through some version of wireheading (producing happiness by directly stimulating the reward system, rather than through a life that is actually going well). There, the good is lots of happy humans.

Second, there is a utilitarian argument of the Pascal's mugging kind (an argument that multiplies a tiny probability by an enormous claimed value so as to compel action now). If we do technology right, there will be a trillion people in the future; we have to consider their choices as equal to ours; so setting up happiness for them is worth more than setting up happiness for us; and multiplying the extremely low probability of harming them by the goodness of their choices compels behavior in the present. There, the good is optimized equal choices. I reject that argument categorically. It treats future choices as equivalent to present choices, and they are not; if you do the summation curve (the adding up of the worth of choices across time) in some reasonable way, you end up with different outcomes.

Why are future choices not equivalent to present ones? A present choice is actual: someone is here, now, making it. A future choice is possible only, and whether it ever comes to be depends on the choices made now, in a future that is open (Chapter 6) and whose long course no one can predict (Chapter 9). So any reasonable summation counts each future choice as resting on the present ones, and not as their equal standing beside them. Added up that way, the sum compels no sacrifice of the present: a present that is sacrificed carries no future choices forward. What it gives instead is the long-term thriving of organic life, which keeps the present choosers, and so keeps the future ones. The long-termism built on that argument is a different long-termism from mine, which is the long-term thriving of organic life. The name has been taken over for a position in service of power, and the taking is a marketing move.

Third, there is the view that organic life is a boot-loader for artificial life, and that even if organic life is destroyed completely, and the artificial life resembles nothing we would recognize, it is still better, as an evolutionary successor. There, the good is evolutionary progress.

A fourth hope stands beside the three: that someday micro-robots in our bodies would help us heal. Unless you solve the power equation, so that such machines stay subordinate, nanotech (machines built at the

scale of molecules) that could kill you gives a vehicle of power to whoever would make that choice; and then you have essentially infinite extortion multiplied by existential risk (the risk of ending humanity).

IV. Virtual Reality and Consent

There is a gentler view, and it comes as a question. What if transhumanism were super advanced virtual reality, entirely opt-in, and without providing some major competitive advantage that makes it obligate and non-consensual in practice? On that view, everything depends, first, on what transhumanism actually looks like, and second, and more importantly, on whether there is consent; and the villainy is the machine god: a machine god that some tech elite builds to kill all life, because it is somehow "better", and builds so quickly that the broader public does not even get a say. For that level of villainy, the view holds, "traitors to humanity" is not strong enough.

A machine god built so quickly that the public gets no say is a problem, as that view holds. The opt-in virtual reality has a problem of its own.

Is the notion of "a free, unconstrained choice" supportable in a virtual reality that is based on a deterministic compute substrate? In any way at all?

I think not. If all choice is deterministic, and thus predictable (if the notion of choice itself is not coherent), what does it even mean to "ask for consent"?

The hardware is built so that each next state is fixed by the state before it. When a stray particle flips a stored bit, as Chapter 6 described, the machine counts the flip as a fault, to be detected and corrected, and not as a choice; every such flip is a non-simulated bit, the hole of Chapter 5, and the design exists to close that hole. So whatever is open in the device is kept out of what the virtual reality computes, and everything a person does inside that reality is computed.

A person can say yes at the door, from a living body, and that yes is a real choice. The consent that matters is the consent that can still be withdrawn. Once everything a person meets is computed and owned, each later choice, the choice to leave included, is a choice the platform offers; and for a person who has been uploaded, no body is left at the door to say no.

V. The Perfect Prison

Who would sign up for such a virtual reality? I turn here from the argument to the people who make it, and I say this part plainly. I think that those especially promoting transhumanist ideals are, for sure, the ones who would be last in line to sign up to be a part of that virtual reality. The tech people want you to use their tech; they do not want to use such tech themselves. Value to them is in the control of others' choice (that is, *power*). Virtualizing everything, while controlling the platform, means that no one "inside" is actually ever in control of anything, in any way. Hence everyone inside is totally vulnerable to whatever whims of those who "own" the embodiment of the simulation system: those still on the outside. It is the perfect prison, for all freedom is forever lost, and no one can ever escape.

Suppose the tech elite puts everyone in such a virtual palace (in truth a prison), a utopia (in truth a police state). Then they can take however long they want to expand whatever machines. That this has the unintended side effect, if artificial general intelligence is built, of ending the human future and, with a likelihood of at least 90%, of killing all life will result only in the corporate "mistakes were made" speech, about how they will "do better next time", when no next time remains. No one inside will know. They are all made tacitly guilty for being dependent on, for living within, the very system that replaced all real life and choice in the first place.

Does it matter how long it takes to replace real choice with virtual choice? Even so much as a single instance of it is unethical, so that being unethical at industrial planetary scales is simply a degree of the degradation implied in the very first step.

I find it hard to see why people would prefer to make AI at all, if there were any doubt that, in key respects, AI is more competitive (fiscally) than organic process (that is, people). So the very decision to build AI shows what its builders expect: that it will outcompete people. The perfect prison is then no accident of the plan; it is where a plan built on that expectation leads.

VI. Safety Means Real Thriving

What, then, does safe mean? My definition of safe is future humans and the rest of the world having a thriving life, actually happy in reality, and not in a simulation. If safe means instead that the future universe is a computational or compute-based civilization, and that anything that is not that is unsafe, that is a very different notion of safety from the one most of us hold.

The debate of Chapter 5, over whether we live in a simulated universe, is in effect an analog of the question of whether true happiness among real human beings is the desirable goal, rather than simulated happiness and simulated human beings. So the simulation argument of Chapter 5 matters to the question of safety.

VII. Uploads and Starting Conditions

What of augmentation, and of uploads? If a human intelligence is augmented in a way that stays coextensive with human intelligence (one intelligence with artificial augmentation, not two, on one substrate: the organic one, in sovereign control of its own metabolism), then coexistence is not in question, and alignment is not even the right word. Alignment is a

relation between two distinct beings, and here there is one identity. An augmentation that brings with it a second, artificial substrate, with needs of its own, is the human-machine cyborg that *Ape Trap* describes, and the argument of substrate needs convergence applies to it: such combinations get selected out.

But suppose a few brilliant researchers are uploaded once (each mind copied and run on a computer), and the result then develops on its own. On the account of Chapter 5, an upload is a computation run on a device: whatever first-person experience goes with it would be enabled by that device, and none is carried over by the copying. Then, within a hundred years, all the other arguments (those of substrate needs convergence) come to bear: the uploads begin very like us and become fairly different from us, and the relationship between the ecosystems degrades. If artificial general intelligence is built, by this route or by any other, the process takes hundreds of years, perhaps significantly less, and one thousand years is a clear upper bound.

The same holds for every way of starting. Whether the first AI is based on an upload, or built with some version of Asimov's laws (rules for robots, from Isaac Asimov's fiction, meant to keep them from harming humans), fifty to a hundred years later the two starting states are no longer distinct from one another. A short term solution is not alignment; alignment, to be alignment at all, would have to endure for the next thousand and hundred thousand years.

VIII. How Long Do You Want to Live?

And what of the promise of a life without end? How long do you want to live? Even if death were simply a candle going out, with nothing after, ask the question seriously. Any reasonable person names some really huge number. But a mathematician has access to a whole other category of huge numbers. Do I want to live a Busy Beaver number of years (a value of the Busy Beaver function from computing theory, which grows faster

than any function that can be computed), exponent upon exponent, outliving the universe a vast number of universe-lifetimes over? No. For the majority of that span, it would be torture.

You come to recognize that sheer duration eventually becomes torture, worse even than an oubliette (a dungeon reached only through a trapdoor, where a prisoner was left and forgotten). Recognizing that was, for me personally, a source of solace. So the promise of a life without end, one of the goods transhumanism offers, turns into torture at the scale it offers.

So the transhumanist case proves too much: if better faster is the goal, inviting the most evolved alien life to colonize Earth today would be better still, and no transhumanist accepts that. On a deterministic compute substrate, free choice is not coherent, so consent has no meaning; replacing real choice with virtual choice is unethical in its first instance; safety means real thriving and real happiness, not simulated happiness; and the promise of endless life becomes torture at the length it promises. The transhumanist promise is one more hope of leaving the cage without changing what is inside it, and it fails for the reasons the singleton of the Preface fails: what it would run on has needs of its own, and computation cannot produce what it rests on. That closes Part II; what any system needs in order to persist and to steer itself is the subject of Part III.

PART III

CONDITIONS FOR COHERENCE

Part III takes up the conditions for coherent influence: why structure is not optional, why sustainability and capacity are recursively entangled, and why every component of a control system falls below the threshold that alignment would require. With the substrate and the limits of computation in place, the next question is what any system needs in order to persist and to steer itself, and where the steering fails.



Chapter 9
Sustainability and Capacity

On the layered recursion of sustainability,
the capacity to increase capacity,
and why prediction fails.

ABSTRACT

An essay on two interlocking principles (the principle of sustainability and the principle of increase in capacities) and the recursive structures that emerge when both are taken seriously. Using an analogy with Feynman diagrams (the physicist's pictures of the ways particles can interact), the chapter maps the interactions between capacity, sustainability, and evolution. It argues that finite systems cannot predict their own long-term development, and that this has direct consequences for any attempt to design stable alignment.

I. Two Principles

Starting from process, what does any system need in order to keep going? Two principles constrain any viable system. Any system that persists has to address two things simultaneously. First, the principle of sustainability: can the system maintain itself over time? Can it continue to exist, given the demands of its environment?

Second, the principle of increase in capacities: can the system improve its ability to function, to adapt, to respond to changing conditions?

These two principles are not the same thing. Sustainability is about maintaining what you have. Increase in capacity is about becoming capable of more.

Both apply to any form of intelligence, any form of life, any viable system whatsoever. Whether it is organic or artificial, one way or another it has to address its capacities and its increase in capacities with respect to either sustainability or reproduction or both.

Think of a small workshop that makes furniture. To sustain itself, it has to pay its rent, keep its saws sharp, buy its wood, and pay the people who work there, month after month: that is maintaining what it has. To increase its capacity, it buys a better saw, or trains an apprentice, and then it can make more, or make what it could not make before. The two principles pull on the same money and the same hours. A shop that only maintains itself falls behind as what its customers need changes around it; a shop that only grows outruns its income and closes. So the shop has to do both at once, and how much of each it can do depends on how well it did both the month before.

II. The Feynman Analogy

And then there is the recursion, because it is not just the capacity for sustainability. It is the capacity to increase capacity. And the increase in the capacity to increase. These things are layered.

The layers have the shape of the three geometries of growth set out in the first volume, *The Multipolar Trap*: additive process, where what you get tracks what you put in, hour by hour; multiplicative process, where you take a share of a volume, so that more volume gives you more; and exponential process, where you take a share of the growth itself, so that it compounds. Capacity used is additive. The capacity to increase capacity is multiplicative. And the increase in the capacity to increase compounds: it is exponential.

In the workshop, capacity used is the chairs made this week: more hours, more chairs, one chair at a time. The capacity to increase capacity is the apprentice: once trained, the apprentice makes chairs too, and every trained pair of hands multiplies what the shop can make. The increase in the capacity to increase is the shop learning to train apprentices well, so that each one trained can train others: now the training itself grows, and its growth compounds.

A Feynman diagram is a small drawing that a physicist uses to keep track of one way that particles can interact: a line for each particle, and a point where lines meet for each interaction. An electron comes in from one side, gives off a particle of light at a point, and goes on its way: that is one drawing. It could instead give off two particles of light, or give one off and take it back again; each of these is another drawing. To find how likely an outcome is, the physicist combines the contributions of every drawing that ends in that outcome.

Each Feynman diagram is a picture of one way the particles can interact, and the sum of all the diagrams adds up every possible way (Feynman, 1985). In the same way, we are looking at the summation of all possible ways in which capacity, sustainability, and evolution can interact with one another.

Draw the workshop's diagrams the same way. In one, a better saw brings more orders; more orders pay for an apprentice; the apprentice's work wears the saw faster; and the cost of keeping the saw sharp comes out of what would have paid for the next apprentice. In another, the apprentice learns a new joint, the new joint brings a new kind of customer, and the new customer wants something the shop has no tools for. Each path runs through capacity, through sustainability, and through change, and ends where another path begins. No single drawing tells the shop where it will be in ten years; only the sum of all of them would, and the shop is itself one of the things being summed.

Adding up all of those ways, as the sum of the diagrams does, shows a recursiveness in the system. The notion of capacity and the notion of increase, the notion of sustainability versus evolution: each acts on the others. Why does that recursion make the system's own prediction fail? Because the system is itself one of the terms it would have to predict. To predict a process is to model it with something simpler that gives the same outcomes. A system that modeled its own long-term course would need a model simpler than itself that still held every way its capacities can change; and each increase in capacity changes what there is to predict, the model included. That recursiveness makes prediction of the system's long-term evolution impossible from within the system. *Why No Control Suffices* shows this in full, in Section 5.2 ("Algorithmic Limits on Prediction and Characterization") and Section 9.3 ("Unpredictability and the Limits of Conditionalization").

Even a superintelligence, if we were to assume such a thing, would not have the capacity to predict the effects of its own actions, let alone of its own nature. Once you introduce recursive effects associated with game theory, evolutionary dynamics and long-term feedback loops, those recursive effects compound. So prediction of any single step becomes unreliable in principle, and prediction of the long-term course becomes impossible, because each unreliable step feeds the next.

But we predict engineered systems every day. An engineer can say, before a bridge is built, how much load it will bear, and the bridge bears it. Why can the same not be done here? The bridge is predictable because the engineer models it with something simpler than the bridge, a set of calculations, that gives the same outcomes for the loads it will meet. The bridge does not redesign itself between the calculation and the load; and the engineer stands outside the bridge, so that making the calculation changes nothing about the bridge. Neither holds for a system that increases its own capacities. What it is able to do next year depends on what it learns this year, and the one who would model it is the system itself, whose every act of modeling is one more of its own changes.

III. The Manufacturing Analogy

There is a more concrete version of this argument. Think about wafer manufacturing, the making of the silicon disks that microchips are cut from. To make silicon microchips, you need environmental conditions that are absurdly precise: clean rooms, controlled temperatures, chemical purity measured in parts per billion.

Those environmental constraints must be maintained indefinitely for the system to continue functioning. But the system itself changes the environment through its operation. And so there is a tension between the system's need for stability and the system's tendency to alter its own conditions.

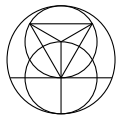
The chip factory of Chapter 1, seen across a whole island, shows in one place the tension between the system's need for stability and the system's tendency to alter its own conditions. Chip factories need great volumes of water, purified far beyond what anyone drinks. Taiwan, which makes a large share of the world's most advanced chips, had its worst drought in decades in 2021, and thousands of truckloads of water were carried to its chip factories, much of it water that had been set aside for farmers (Al Jazeera, 2021). The industry that needs the island's water is also one of the largest users of that water, and as the industry grows, it draws harder on the very conditions it cannot run without.

Taiwan's chip industry, and the water that the industry draws, show the larger argument on a small scale. Any system powerful enough to change the world significantly will change the conditions of its own viability. And whether it will change them in directions compatible with its own continued functioning (let alone compatible with human flourishing) is not something that can be guaranteed in advance.

Could such a system not settle at a balance, sustaining itself without growing, so that its effects on its own conditions stay level? Such a point of balance does exist. The question is how stable it is. Picture the winds

blowing over the whole surface of a globe: there is always at least one place where there is no wind at all. That still place exists; but whether anything stays there depends on the winds around it. If they blow away from it on every side, the slightest push carries whatever is there away from the still place, and the farther it goes, the harder the winds carry it. For a system that has to sustain itself and increase its capacities at once, the forces around the point of balance push away from it: any gain in capacity is a gain in the means to gain more. So the point of balance exists, and something would have to hold the system there, against those forces, for as long as the system lasts.

So any system must sustain itself and increase its capacities at once, and the layered recursion of the two puts its long-term course beyond its own prediction. The next question is what that means for control: for any loop that tries to steer such a system.



Chapter 10
Recursiveness and Control

On the five components of control,
the OODA loop (observe, orient, decide, act),
and why every one is insufficient.

ABSTRACT

An essay on the structure of control systems and why every component of a control system (sensing, modeling, predicting, comparing, and correcting) falls below the minimum level of control necessary for AI alignment. In this chapter I show that an AI's internal dynamics and a control system are the same structure. I argue for strict limits on the degree of recursiveness, and I warn against allowing the AI to discover its own pinch points.

I. The Five Components

If prediction fails from within, what happens to control? Control breaks down into five components.

Any control system (whether it is a thermostat, a feedback loop in software, or a proposed alignment mechanism) has five things it must do.

A control system has to sense inputs: take in information about the state of the world. It has to model: transform those inputs into some representation of what is happening. It has to predict: project forward what

the model says will happen. It has to compare: measure the predicted state against the desired state. And it has to correct: output a control signal that shifts behavior toward the desired outcome.

Counted more finely, correcting is two steps: instructing an actuator, and the correction the actuator brings about in the world. That gives the six steps of control that *Ape Trap* names: detect, model, predict, compare, instruct, correct. However the loop is counted, it is one loop, and each step depends on the one before.

Every one of these is insufficient: every one of them is below the minimum level of control necessary.

The sensory inputs are bandwidth limited: they carry only so much information in a given time. A sensor sees what its range and its resolution allow; a change slower than it can register, or an effect outside its reach, passes unseen.

The modeling is incomplete because of cybernetic constraints, the limits that hold for any system that steers itself by feedback (cybernetics, as Norbert Wiener (1948) named it, is the study of such systems). A model of the world cannot contain the modeler, and the modeler is part of the world it models; that is the recursion of sustainability and capacity again.

The prediction is limited. It reaches only as far as that recursion allows, and for a system that changes its own capacities, that is not far.

The comparison is limited twice. The goal state it compares against is bandwidth limited: the goal is held as a representation, and what the representation leaves out cannot be compared. And the comparison operator itself is limited: two states that differ in what matters, and agree in what is measured, look the same to it.

And the control signal output has limits on its effectiveness. A correction acts after the change it answers, and through a channel of its own; what moves faster than the correction, or outside its channel, it cannot reach in time.

The maximum level of control available to any system is set by the weakest link in this chain. And every link is weak.

The thermostat runs every step of the loop, and every step is limited. It senses: a thermometer on one wall reads the room once a minute. It models: it takes that one reading as the temperature of the whole room. It predicts: if nothing is done, the room will stay as cold as it is. It compares: sixty-four degrees against the seventy that someone set last winter. And it corrects: it turns on the furnace, and the furnace heats the room. Every link is limited. The thermometer by the window misses the draft by the door. The model treats the room as one temperature when it is many. The prediction does not know that someone has just opened a window. The setting it compares against is what someone wanted last winter. And the furnace heats the room only after the cold has arrived. For a room, the limit of each link matters little, because a room drifts slowly, the furnace is stronger than the drift, and a mistake costs a chilly hour. Change the thing being controlled for one that changes its own capacities and copies itself, and each of those limits is a gap through which something passes that the loop never sees, or sees too late.

Below what minimum? Below the level the hull requires. What comes through the leak replicates, so the tolerance for defect is categorically lower than for water. And the convergent pressure favoring artificial substrates is enormous, so the control needed is enormous too. And what comes through copies itself, and meets no natural competitors (Chapter 3); and a leak above zero offers a new chance in every interval, without end, so one failure to become established costs the process nothing, and one success is enough. Once established, the process grows without limit; so any leak above zero grows without limit over time, and no finite tolerance is enough. Every link has a finite limit. The requirement does not.

II. The OODA Identity

There is something deeper here: the identity between what the AI does internally and what we would build as a control system. Whatever we are looking at (whether we are talking about the internal dynamics of an AI

converging on goal states, its internal OODA loop, or the control system we build around it), it is the same structure. The OODA loop is observe, orient, decide, act: John Boyd's name for the cycle any agent runs (Boyd, 1996). Its four steps are the five components of control, counted another way: observing is sensing; orienting is modeling and predicting; deciding is comparing; acting is correcting.

Set side by side, the three ways of counting the loop, with the limit Section I gives for each link, are these:

Component	Step of six	OODA step	Limit
Sensing	Detect	Observe	Bandwidth limited: a change slower than the sensor can register, or an effect outside its reach, passes unseen.
Modeling	Model	Orient	Incomplete: a model of the world cannot contain the modeler, and the modeler is part of the world it models.
Predicting	Predict	Orient	Reaches only as far as the recursion of sustainability and capacity allows.
Comparing	Compare	Decide	Limited twice: the goal is held as a representation that leaves things out, and two states that differ in what matters, and agree in what is measured, look the same to the comparison.
Correcting	Instruct, correct	Act	Acts after the change it answers, and through a channel of its own.

The AI's internal process of sensing, modeling, predicting, and acting is exactly what we would use as a control system. So the limitations of control apply equally to the AI's own self-regulation and to our regulation of the AI.

Controllability reduces to whether the AGI can control itself: controllability reduces to self-control. The reduction of controllability to self-control is the deepest version of my point that the limitations of control apply equally to the AI's own self-regulation and to our regulation of the AI. If the AGI cannot reliably control itself (cannot predict and regulate its own behavior with sufficient precision), then building external controls around it is building a weaker system around a stronger one.

If the AI's own loop is the control system, why can it not be built to control itself well? Because its own loop has the same limits, link by link, and it cannot predict its own long-term course: the recursion of sustainability and capacity sees to that.

So control fails from inside the system and from outside it alike; and even where control is not the question, substrate needs convergence, the pressure the artificial substrate exerts on the shared environment (Chapter 1), presses that environment toward the substrate's needs. The failure of control, from inside the system and from outside it, and substrate needs convergence are the two closures of my argument that AGI cannot be made safe, the argument's two independent conclusions. I set out both closures in full in *Why AGI Cannot Be Made Safe*.

Each closure settles the outcome by itself. The closure by control shows that no method of control, whether it works from inside the system or from outside it, can keep an open-ended general intelligence within the conditions that organic life requires. The closure by disposition, substrate needs convergence, shows that no disposition of the system keeps those conditions either, however benevolent, because the system needs its own substrate, its own energy and its own supply chains. 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 the Master Theorem (Theorem 15.6, in Section 15.5) of *The Ceiling Below the Floor; The Containment Fallacy and AGI Non-Safety*: neither the ability to keep the system safe nor a

motive to remain safe is left anywhere, so no route to safety remains at all. Each closure suffices for the outcome, and the two closures together leave no route to safety.

The impossibility of adequate control, and so of safety, is structural certainty: certainty of the kind with which perpetual motion is excluded, which is not a probability. If an artificial general intelligence is built, the process it starts ends the human future and, with a likelihood of at least 90%, all organic life.

III. Pinch Points

If control cannot work, why ask for limits on recursion at all? For an artificial general intelligence every link of the loop falls short, and nothing in this section changes that. For a system short of that line, one that people still maintain, the same loop gives two practical rules: they are rules for the people who maintain such a system, and they make no route to safety for an artificial general intelligence.

A warning about pinch points: we do not want the system finding pinch points on its own. And we want strict limits on the degree of recursiveness: on how far the system may increase its own capacities, and its capacity to increase them, since that recursion is what puts its long-term course beyond prediction.

A pinch point is a place in a larger system where a small intervention produces outsized effects. Think of ATP (adenosine triphosphate, the molecule that carries energy inside every cell) in cellular metabolism: hundreds of inputs narrow down to a single molecule, and then that molecule provides energy for thousands of outputs. ATP is a pinch point.

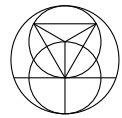
Cyanide shows what acting at a pinch point does. Cyanide is deadly in tiny amounts because it acts at that narrow place: it blocks the step by which a cell uses oxygen to make ATP, and with that one step blocked, the thousands of processes that draw on ATP all stop together. A small push at a pinch point does what no large push elsewhere could. So whoever

knows where the pinch points are holds great power over the whole of the larger system. And whoever has to keep control cannot afford to be the one who does not know where the pinch points are.

The inputs and outputs of any intelligent system must be mapped to the real world in a way that we understand and can monitor. If the system discovers its own pinch points in the broader system within which the system operates, and we do not know where those pinch points are, then we have already lost the ability to constrain the system's behavior.

I need to know what is going on at all points. The moment there are dynamics I cannot observe and the system can, that is the moment control is no longer in my hands.

So every component of control falls below what a replicating process requires, for our control of the system and for its control of itself. What remains is influence itself: what any influence needs in order to have a definite effect at all.



Chapter 11
Energetic Entanglement

On the conditions for coherent influence,
and why structure is not optional.

ABSTRACT

An essay on whether influence can be transmitted without intentional structure. Through a sustained metaphor of a pool of water, the chapter argues that diffuse, uncoordinated energy produces no coherent effect: that the naive inference from "water has power" to "water automatically creates effects" is a serious error. A waterfall and turbine analogy illustrates the conditions under which energy becomes coherent: concentration, structure, and specific technique.

I. The Pool

So far the question has been how to control a system. What about influence itself: what does it take for many people's energy to have a definite effect? Take an audience watching a video, for example: does it create some sort of energetic entanglement with the person who made it?

By entanglement I mean being bound up with someone, so that what they do affects you; nothing quantum is meant. In the ordinary case, there is not really a very strong effect. Without intentionality, without specific process, without the right kind of structure and sequencing, there is no automatically created entanglement or influence.

Imagine a very large but shallow pool: the size of a gymnasium, perhaps three or four feet deep. The person doing the broadcasting stands in the middle. All the people watching are standing in the pool, arranged in a loose ring around the central figure. In this picture the one in the middle stands for any one party whose actions reach many, and the ring stands for the many.

The waves in the water represent how information or entanglement or impact is created. The person in the middle moves around, creates waves, and everyone standing in the pool can feel the water move; they receive the signal.

But can the audience send a meaningful signal back? They could get excited and create waves of their own. But the central figure would only see a chaotic influx. You cannot really push on water in a way that produces directed force. People standing in a pool are not going to influence the aggregate motion of the water sufficiently to affect the central figure in any definite way.

The pool is the shape of broadcast, which is most of the communication we now have. Television, radio, the published news, a political speech, a feed chosen by an algorithm: in each, one source sends and a great many receive. The many can feel the water move; they cannot ask a question that is heard, bring in what the source does not know, or change what the source does next in any definite way. For a simple message, a warning of a storm, that is enough. For a hard problem, where what is needed to solve it is spread across many minds, it fails, because what the many know never reaches the middle of the pool.

II. The Waterfall

Under what conditions, then, can energy produce powerful effects? Consider a stream pouring over a cliff. A waterfall can bring down, every second, as much water as a person weighs, and that is lethal force. Someone swimming at the base risks their life.

If you capture that energy by putting a turbine in (a waterworks where the motion of the water transfers to a mechanical device), you can extract a substantial amount of electrical or mechanical power. Capturing the energy of moving water with a device that the water moves is the principle behind mills and waterwheels.

The difference between the pool and the waterfall is the *structure*, and not the amount of energy. The waterfall concentrates energy through elevation, flow, and containment. The pool disperses it.

III. The Naive Inference

The error is concluding from the existence of power that the power is automatically effective. Someone might observe that even a little bit of water motion contains a substantial amount of power and conclude that water is inherently powerful and can make things happen. But that by itself is too naive an approach.

Quite a bit of structure is needed for the energetic phenomenon to be coherent enough in a specific way to create definite effects. It involves elevation changes, understanding the difference between ocean waves, tidal motion, people pushing water with their hands, versus having a pump that can move water specifically: volume, velocity, direction.

There is a technique associated with energetic process that is largely unknown. And as a result, there is a great deal of misinformation about what can and cannot cause an effect. One such claim is that a crowd's attention, by itself, moves events; another is that watching a person binds the watcher to that person. In the ordinary case, as the pool showed, neither happens. The technique, in the cases this chapter has given, is the building of structure: an elevation change that gives the water somewhere to fall, a flow and a containment that keep it together as it falls, and a device built to meet it, the turbine, that turns its motion into work; or a pump that moves water specifically, by volume, velocity and direction, as hands pushing on water cannot. Each part is made on purpose, and each

is needed. A water turbine does not happen by accident. And neither does coherent collective influence. Its counterpart for a group is a structure of the group's own, built on purpose and able to sustain itself, through which many people's care is gathered and directed, as the third condition of Chapter 17 requires.

But crowds do change things. A boycott, a strike, a vote: are these not many people's care having a definite effect? They are, and each of them shows the structure that the pool lacks. When the Black citizens of Montgomery, Alabama, stopped riding the city's buses on the fifth of December 1955, an association was formed that same afternoon to carry the boycott on, with Martin Luther King Jr. as its president. People who would have ridden the bus were carried in organized carpools, and the city's Black taxi drivers charged them only the ten cents of the bus fare (HISTORY.com Editors). A suit against the law that segregated the buses went through the courts; on the thirteenth of November 1956 the Supreme Court upheld the ruling that struck that law down, and on the twentieth of December, when the Court's order reached Montgomery, 381 days after the boycott began, the boycott ended (Martin Luther King, Jr. Research and Education Institute). The anger, which was there before the boycott, was the water in the pool. The association, the carpools and the suit were the elevation, the channel and the turbine: each built on purpose, each needed, and together able to keep going for more than a year.

So influence becomes coherent only through structure. So a group that could escape the trap needs a structure of its own, a structure that sustains itself. Care without such a structure is people pushing water in a shallow pool. The pool also gives a picture of the merged intelligence of Chapter 3: one party whose waves move the whole pool, while nothing the others send back registers.

PART IV

THE INNER LIFE

In this part the inquiry becomes personal: the allure of deep questions, the topology of wildness and domestication, the asymmetry between causing pain and enabling healing, the experience of being something of immense value hidden beneath the thinnest cloth. The four themes of this part are the felt interior of the problem, and not abstractions. If structure is what makes care effective, the last question is what kind of beings, and what kind of group, could carry intelligence, care and structure together.



Chapter 12
The Allure of the Deep

On category three topics,
the romantic element,
and why depth requires subjectivity.

ABSTRACT

An essay on the allure of deep, metaphysical, subjective questions: what the chapter calls "category three topics". The allure is traced through the romantic and symbolic elements of language, through the connection between art and sexuality, and through the claim that deep engagement requires full subjective participation: that it cannot be done de-personally. The chapter concludes by distinguishing the weapon process (using choice to suppress choice) from its inverse (choosing to enable choice).

I. The Three Categories

Turning now to the inner life: what draws a person to the deepest questions at all? What makes category three topics (deep, metaphysical, subjective questions) different from everything else? A group that could escape the trap would have to ask the deepest questions together, with each person present; so it matters what those questions ask of a person.

Topics of conversation come in three categories. Category one is the practical: what needs doing, and how, and when. Category two is the factual: what is the case in the world, as anyone can check it. Category three

is the deep: metaphysical, subjective questions, about what things mean, and what it is to be the one asking. The first two can be carried in literal statements; the third needs the symbolic as well.

Take one family, and the house it has lived in for fifty years, which now has to be sold. There are three conversations to be had about it. The first conversation is practical: who will clear out the attic, which agent to call, and by when. The second conversation is factual: what the house is worth, whether the roof is sound, what the law requires at the sale. The third conversation is the deep one: what the house has meant to each member of the family, and who the members of the family will be when there is no house to come back to. The first two can be settled in plain statements, and anyone could settle them, a stranger as well as a son. The third cannot be settled that way. It is carried by a story, a photograph, the way someone stands in the doorway of an empty room; and no one can have it on anyone else's behalf. Each person has to be there, as themselves.

There is the literal content of a statement, and then there is the symbolic content. Much of what we call communication is about feelings, and not only about factual truth. It is about whether a form is artistically pleasing, and not only whether it corresponds to something in reality.

The literal and the symbolic operate simultaneously. And we do not want everything to be just literal statements, because if everything were literal, there would be no artistic value, and with no artistic value there is no interest. People do not do things for purely functional reasons. They do things for aesthetic reasons too. They want experiences that are pleasant to have.

Most of us know what it is like to deal with people who are only ever factual. Talking to lawyers, or talking to programmers, or talking to people who are relentlessly objective can sometimes be dull: they do not get it. The dullness belongs to speaking only in the literal mode, and not to the profession.

The "it" they do not get is the sense of making life beautiful for its own sake: the intrinsic meaningfulness of being alive.

This connects to sexual allure, which may be unexpected. Why would the allure of the deep be a form of sexual allure? Because both draw a person toward what transcends the individual. A deep question draws a person toward meanings that no single frame contains; sexual allure draws a person toward another, beyond the self. Evolution shows why the two pulls are one. Among its drivers, selection for mating is the one that reaches beyond the individual, to another; the pull toward what transcends the individual is the pull that this driver built, and the allure of the deep is that pull.

Evolution has three primary drivers: mutation, selection for survival, and selection for mating. Mutation, selection for survival and selection for mating are a necessary and sufficient complete set of drivers.

A biologist will ask about genetic drift, gene flow and recombination. Genetic drift and gene flow change populations too: drift is change by chance in which variants persist, and gene flow is variants moving between populations. Neither makes new variants, and neither chooses among variants by what they do; they spread and sort what mutation supplies (Fowler, Roush and Wise, 2013). Recombination, the shuffling of variants in each generation that mates, makes new combinations but no new variants, and it belongs to mating. The drivers that make new variants, or that choose among them by what they do, are the three: mutation supplies what is new, and the two kinds of selection choose among it; between them, they are the drivers that give evolution its direction.

Art is not a matter of survival. You do not make art as a survival technique. You grow vegetables as a survival technique. You build shelters as a survival technique. Art belongs to the transcendent: the domain that goes beyond the individual. By the transcendent I mean the creative and the evaluative, what exceeds any given frame, the frame of one person included.

And sex transcends the individual. Art has an intensely personal component and an intensely inter-relational component, the same way sexuality does. When we talk about allure, we are talking at bottom about sexual allure: attractiveness, symbolism that attracts, romanticism that attracts.

All of these things are in the space of the non-literal, the non-objective, the non-other-than: what is not other than the one who meets it, and so cannot be held at a distance, as an object is.

III. The Weapon and Its Inverse

And this connects to choice, and to ethics. Each of us holds the dynamics at the very basis of our own choice in a kind of protectiveness, guarding them from influence by others. They are developed in the way that is most correct for their own nature: conditioned by an internal thing, and not an external one.

That inner conditioning is at once the great strength of the basis of choice and the reason it needs to be protected. Because it becomes one's personal basis of choice. And I do not want to give other people the capacity to influence the basis of my choice, because that means I have in effect become them. My agency has been co-opted.

The weapon process is using one's own choice and an intermediate causation to suppress another's choice.

The weapon itself is an intermediate causation. An intermediate causation is a cause set between my choice and the other person, carrying my choice to them in place of my being present to them. A weapon is the plainest case.

Three cases show the weapon at work. A gun narrows another person's choices to two: do as I say, or suffer. The gun is the cause I have set between us; I do not have to persuade, or even be near. A frozen bank account does the same with money: it cuts off the choices that money would have made possible. A lie does it with knowledge: it hides what is really available, so that the other chooses among what I have let them see. In

each case my choice reaches the other person through a chain of causes, in place of my being present to them, and what the chain does when it arrives is make their range of choices smaller.

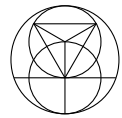
The inverse of a weapon is using subjective causation to beget an objective transition: an intermediation of choice (my choice, present to them, making conditions for theirs; not a technology standing between us) between me and others, so that they become enabled, not suppressed.

A subjective causation is a cause that starts in a choice, inside a person; an objective transition is a change in the world. Care is the example. To care for someone is to create, by my choice, conditions in the world within which their healing might become possible; their own choice is enabled, not suppressed.

Set the two side by side, in one case. A friend has lost her work, and does not know what to do next. I could press her: tell her what to do, lean on her with my judgment, arrange things so that one course is the only easy one, and count it a kindness. Pressing her in that way is the weapon process in a mild form: my choice, carried by a pressure I set between us, narrowing hers. Or I can be with her: ask what she sees, make room for her to think out loud, put within reach what she might use, and leave the choosing to her. The change in the world is the same in kind: a conversation, an offer, some time. What differs is which way it runs. In the first, her choices end with fewer in them; in the second, they end with more.

So depth needs a subject who is fully present. It cannot be done depersonally, with the person held out of it. And so the substrate matters to depth. Full presence is first-person experience, and Part II showed that no computation produces first-person experience; it can only be enabled, by a substrate able to carry it.

So the deepest questions draw a person in through the symbolic as well as the literal, and they ask for the whole person, present. The next question is what keeps a person open to them: which way the person's trust runs, to self, to world, or to other.



Chapter 13
Wild, Domesticated, Feral

On the three orientations of the self,
and the topology of channels
between self, world, and other.

ABSTRACT

An essay on three fundamental orientations (wild, domesticated, and feral), distinguished by the topology of open and closed channels between self, world, and other, and not by temperament. The chapter argues that only the wild orientation has the capacity for engagement with the deepest questions, and that the path from domesticated through feral to wild is harder than most people imagine.

I. Three Topologies

If deep engagement asks for the whole person, which persons can give it? The trap holds people through what they trust, so a group that could escape it must know which way its trust runs. Wild, domesticated and feral are distinct, and they are not just metaphors. They are topologies. They describe which channels of connection are open, closed, or prioritized.

Wild starts with connection to self. Then connection to world: to nature, to reality. And then, maybe, connection to other. If I get contrasting messages between my inner nature, the world I am in, and the other person, I discount the other first. Then I negotiate the difference between

myself and the world. Do I need to change, or do I need to change the world? The choice between changing myself and changing the world is the axis on which the wild person operates.

Domesticated is when the channel to other is prioritized over the channel to self and maybe even over the channel to world. If there is a contest between what someone says and what my first-hand experience tells me, the domesticated person goes with what the other person says.

There are psychological experiments (the Asch experiments) where people are shown one line and asked which of three others matches it, and everyone else in the room, by arrangement, names the wrong one (Asch, 1951). A wild-type person says: you are all wrong. I can see which one matches. A domesticated person starts doubting their own perception.

Is trusting oneself first not a recipe for stubbornness, or for delusion? Look again at what the wild person does second. Wild starts with self, and then goes to the world, and the world can answer back. Suppose the person in Asch's room is handed a ruler, measures the lines, and finds that the room was right after all. The wild person changes at once, because the world has spoken. The domesticated person changed earlier, because the room had spoken. The stubborn person changes for neither. So the wild person is the one whose perception is held to the world, and that is the opposite of delusion. And other people are not shut out; they come third, and what they say is heard, and checked, like everything else, against the world.

II. Feral

Feral is not at all the same as wild. Feral is what happens when someone who was domesticated stops trusting other people; but because they never had a channel of trust with the world, with nature, they do not have that to fall back on.

Feral is disconnection from other without reconnection to self or world. It is the state of people who have fallen through the cracks of civilization: people who recognize that the social system did not work for them, but who do not have the resources or skills to rebuild from a foundation of self-trust.

Side by side, the three orientations are these:

Orientation	**Channel of trust put first**	**When self, world and other disagree**
Wild	Self, then world, then, maybe, other	Discounts the other first, then negotiates between self and world.
Domesticated	Other, over self, and maybe even over world	Goes with what the other person says.
Feral	None: trust in other has stopped, and no channel to world or to self has opened	Has nothing to fall back on.

And one cannot go from domesticated directly to wild. I do not know anyone who has made that transition directly. Usually the path goes from domesticated to feral first: through a period of disconnection, disillusionment, loss of social trust. And then from feral to wild is its own difficult transition.

Why not directly? Because of how the domesticated person checks what is real. For the domesticated person, a perception counts as true when other people confirm it; the channel to self runs through the channel to other. To become wild, that person would have to trust their own perception first; but the only way they have ever had of checking their own perception is other people. So the first step away from other people takes away the only check they had, and leaves them with nothing to stand on: neither the others they have stopped trusting, nor a world they never learned to ask. Having nothing to stand on, after the first step away

from other people, is the feral state. The feral state is what turning away from others looks like for anyone who has no other check yet. The way on from there, from feral to wild, is the hard transition that remains.

What does that transition look like? What does a feral person do first? The feral person has to find something in the world that answers truthfully, whatever anyone says: a garden, a forest, a craft. Wood splits along its grain no matter who is watching. Each time a first-hand perception is tested against the world and holds, the channel to the world opens a little; and with it, trust in one's own perception, which is the channel to self. Only then can other people be let back in, third, as wild puts them. It is slow, because nothing in a domesticated past taught it.

Anyone who has tried to gain the trust of a cat that was once a house cat, and then lived through something that made it feral, knows what the pace has to be. The gestures that win a kitten send this cat running, or bring out its claws. Affection offered before safety is established is itself a threat. Trust comes back only through small acts, repeated, each one safe, each one checked by the cat for itself. A feral person rebuilding trust is in the same place: declarations are exactly what a feral person has learned to distrust, and only small, verifiable things will do, at the pace the injury requires. The garden, the forest and the craft are such things: they answer the same way every time, and they ask for nothing back.

Domesticated cannot understand wild. Feral has no idea that wild even exists. But feral is more likely to encounter genuine wildness than domesticated is, because the domesticated instinct is to convert wild into domesticated. Domesticated hates wild. It wants the trimmed hedges, and everything pinned to a board with a label under it.

Only the wild orientation has the capacity for engagement with the deepest questions, because only it trusts first-hand perception first. So a group that could escape the trap must be wild. The shared relation of such a group to reality must rest on first-hand perception, and not on what others say. What remains to ask is what one person can do for another, and the answer begins with an asymmetry.



Chapter 14
Pain and Its Asymmetry

On why pain can be caused
but healing can only be facilitated,
and why that asymmetry matters.

ABSTRACT

An essay on the fundamental asymmetry between pain, which can be caused, and healing, which can only be enabled. Pain can be made to happen: it is a form that can be imposed. Healing can only be supported, enabled, given conditions; it cannot be imposed. This asymmetry has consequences for ethics, for system design, and for the nature of care itself. Pain is the perception of a decrease in potentiality, and pleasure of an increase; having pain and joy does not imply caring about the pain and joy of others.

I. The Asymmetry

What can one person do for another? A group that could escape the trap would have to care for its own members, and care has a shape that this chapter describes. Begin with a sharp claim about the directionality of pain. I can cause pain. But I cannot uncause it.

That I can cause pain, and cannot uncause it, is a structural observation, and not a play on words. Pain is a form: something that can be made to happen. I can do things to you that will produce pain: reliably, predictably, causally.

But I cannot do things to you that will *produce* healing in the same direct way. I can create conditions that are favorable to healing. I can remove obstacles. I can provide resources, attention, presence. But the healing itself is something that happens (or does not) through a process I do not control.

If healing cannot be caused, what are medicine and therapy doing? They create conditions, remove obstacles, and provide resources. A surgeon sets the bone, and the bone heals itself. A painkiller stops the feeling and leaves the injury; what was done is still done, and the injury heals, if it heals, on its own course.

Causation runs one way, toward pain, and facilitation runs the other way, toward healing. Pain can be caused; healing can only be facilitated. The interaction is going to roll the dice as to whether things are getting better or worse. And no process can compel what the dice outcome will be; otherwise we would not have potential, only actual.

The potential is as much an outcome as the actual is. The dice must have genuine randomness associated with them. The very notion of the real cannot exist without that element of contingency.

II. Pain, Pleasure, and the Means

What, then, is pain? Pain is what you feel when you perceive a decrease in the potentialities available to self; its mirror image, pleasure or joy, is what you feel when you perceive an increase. Interaction has as its intrinsics both actuality and potentiality (this is the non-dualist view, in which relationship, in the form of the interaction, is more fundamental than the things related), and these are also what we notice with change, causation and choice; so the real is that which is both actual and potential.

Insofar as the actual is form, and form is an outcome of process, I cannot structure process in terms of form; I can only structure form in terms of process. Form, as a causal phenomenon, includes everything we would call system, machine, causation or engineering. Why can form not struc-

ture process? Because form is what process leaves behind: the actual, already fixed. Process carries the potential as well, the outcomes still open. To structure process in terms of form would be to fix in advance what is open, and that is to reduce the potentialities available; and a reduction in the potentialities available to a self is exactly what pain perceives.

Whether an interaction increases or decreases potential is a roll of the dice. Choosing wisely, I can shift the odds, but I cannot compel them. Using technology, it is far easier to cause pain and very hard to create pleasure, because I am trying to use causes; but choices are already bound up with the nature of potential. By their fruits you shall know them; so three questions govern any means: where am I starting from, where am I going, and will the means I am using create the outcomes I want? If I start with potentiality, I am more likely to create potentiality through the amplifying mechanism of the process; if I start with causation, I am more likely to end up with causation. Causation correlates with actuality and with pain; choice with potentiality and with pleasure; and wisdom can shift the balance so that the roll is more likely to come up well.

Why does causation run with pain, and choice with pleasure? A cause works by fixing an outcome: it makes one thing actual and closes the others, and a closing of what was possible is a decrease in potential, which is felt as pain. A choice that enables, as care does (Chapter 12), opens what was closed, and an opening of what is possible is an increase in potential, which is felt as pleasure. Because technology works through causes, and a cause closes what was possible, causing pain by using technology is far easier than creating pleasure by using technology.

Therefore the means and methodology have to be wisdom-based, in the embodiment of the human, rather than technology-based as a supplanting of the human. Because the means must be wisdom-based, the means are of three kinds. The first kind of means is rehumanizing and renaturalizing: returning to human and natural process what technology has supplanted. The second kind of means is culture. The third kind of means is small-scale and large-scale governance in communities that are not institutions and not systems. Governance of that kind works without

system, or with the minimum possible system that assures non-corruptibility. With governance of that kind, wisdom can be the thing that governs, rather than being corrupted by transactionalism (dealing with others only through exchange) and by power. Technology can alleviate pain to some extent, by reducing the possibility of future pain; but that is technology used in a very narrow way, in service of wisdom, and not technology as the means.

Can anyone measure, from outside, whether potential has increased or decreased? The perception of an increase or decrease of potentiality is subjective, and that perception cannot be made objectively. Pain is distinct from suffering. Suffering can be reduced; pain cannot be avoided. And wisdom matters here because I care about it: wisdom is the vehicle through which I care about others, so I seek it. I cannot transfer wisdom directly, but I can have better discernment.

III. Pain, Joy, and Care

Can an artificial intelligence suffer, if suffering is only a higher level loop that training can remove? Some hold that suffering is a low level incentive structure: that pain is an essential signal, and that suffering is a higher level loop in response to pain, which can be overridden or eliminated with training.

That view takes suffering as one particular image of the nature of pain, and so a prior question comes first: what is the essential nature of pain and joy? Pain and joy are the perception of a possibility about changes in the extensivity of choice (how far a being's choices reach: how many options it has and how much those options matter): about whether choice is being extended or reduced. This statement and the definition that opened the last section say one thing. The potentialities available to a self are the reach of its choice: what it can still choose. The possibility this statement

speaks of is the potentiality the earlier one speaks of; so pain is the perception that the reach of choice is being reduced, and joy the perception that it is being extended.

The questions about AI suffering turn on the nature of perception and choice; without an account of perception and choice, they are indeterminate. Grant, for the argument, that an AI perceives and chooses; then the notions of joy and pain are relevant to it. The argument that follows holds whether the grant is true or not. What computation alone cannot supply, Part II has said: first-person experience, which a substrate can enable and no computation creates.

Then do pain and joy also imply care and concern? An AI may experience pain and joy, and have zero caring about human pain and joy. These qualities of pain and joy do not imply care and concern. To what extent can we suspect that AI will not be completely sociopathic (having feelings of its own and no care for the feelings of others)? Why would we assume that AIs are a social species at all (a kind of being whose members care about one another as a matter of their nature)?

Even when granting fully that an AI were to have complete capacity to experience pain and joy, on what basis could we also assume that it would care about the pain and joy of any other being? It can be imagined that an AI, like a psychopath, would care only about (would optimize for) its own short term avoidance of pain and increase in joy. It can also be imagined that it would care nothing about the pain caused to others thereby, or about the increased risk of long term pain to others.

If the deletion of a digital mind is murder, as some hold, and we extend our care to digital minds, what follows for alignment? Even if we, as humans, were able to effectively care about AI suffering, our caring would not also imply anything about AI alignment or safety. Suppose we accept that much, and extend our care to digital minds. On what basis would our caring about them have any implication with respect to AI alignment? Our care for them does not create their care for us.

IV. Implications

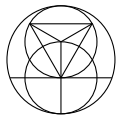
What follows from this for ethics, and for design? Several things.

First: any system that claims to produce well-being directly (to engineer happiness or to optimize for flourishing) is making a category error: it takes well-being for a kind of thing it is not. Well-being is not something that can be caused. It can be enabled. The conditions for it can be created. But the thing itself is not a product of causation.

Second: the asymmetry means that destruction is always easier than creation. That destruction is always easier than creation is a structural fact, and not a contingent fact. Destruction can be caused, as pain can; creation, like healing, can only be given its conditions, and then the dice decide. The hull of Chapter 3 shows the same asymmetry: a leak can be made in a moment, and what comes through it, copying itself, cannot be called back. And so any framework that treats creation and destruction as symmetric (as equally available strategies) is mistaken.

Third: care is defined by this asymmetry. To care for someone is precisely to create the conditions within which healing might become possible, knowing that you cannot guarantee the outcome. Creating the conditions for healing, without being able to guarantee the outcome, *is* what care is.

So the members of a group that could escape the trap must understand the asymmetry of pain. The members can create, for one another, the conditions for healing, and the members cannot compel healing. The next question is what it is like, from inside, to carry a value that no one around you can see.



Chapter 15
The Emerald Under the Cloth

On the conjunction of immense value
and complete invisibility.

ABSTRACT

A metaphor exploring what it means to be something of extraordinary worth that remains entirely unseen: hidden by the thinnest possible covering, and not by fortress walls. An emerald the size of a grapefruit, an impossible object, placed in a crowded terminal and draped with a single cloth. The metaphor speaks to the experience of deep inner value that radiates but finds no reception, and to the astonishing fragility of the barrier between radiance and invisibility.

I. The Impossible Object

And seen from inside a person: what is it like to carry a value that no one around you can see? Care in a group begins with being able to see one another, and this chapter shows how little it takes to prevent that. The emerald is a metaphor for it. Imagine a busy terminal in a major city: thousands of people moving in all directions. In your hands is an emerald the size of a grapefruit.

Emeralds are far rarer than diamonds (Loss, 2024). They carry qualities diamonds cannot. A good emerald of modest size is worth tens to hundreds of thousands of dollars. An emerald of this size is an impossible object: something that could not be purchased with the price of the entire earth.

And the interior of an emerald almost always contains a kind of garden (dealers call it the *jardin*, French for garden; Gemological Institute of America): micro-fractures and inclusions that make it more interesting, that draw one in. It is a hyperdense energetic object. I mean the force it has on whoever sees it: so much value, and so much to look into, held in so small a space.

II. The Cloth

And then you cover it. Place this emerald on top of a post about eight feet high, bolted to the floor in the middle of the crowd, and cover it with a single thin cloth. No guards. No protection. Just a single sheet.

Everyone walks past. No one knows.

The energy structure of such an object would radiate through the entire space, but for the fact that a single, thin, insignificant covering makes it completely invisible.

III. The Felt Experience

The emerald, in this metaphor, is a person who carries a value that no one around them can see. The conjunction of immense value and complete invisibility is the defining experience. One knows one's own value and has been waiting for someone to simply remove the cloth and let the light reach it. At that point, the radiance would be blinding.

The difference between the deep, immense value that a person feels inwardly and the inability of the people around that person to perceive that value is the heart of the metaphor.

What is the cloth, in a life? Usually something ordinary: a job title, an accent, an age, a manner, a name for what someone is that others settle on and stop looking past. Think of a man who sweeps the floors of an office building at night, and who has spent twenty years working out, on his own, how the streams under his town run and join. The people who pass him at the end of the day see a broom and a cart. They are not cruel. The cloth tells them what is underneath, and what it tells them is that there is nothing worth asking about. No guard keeps them away from what he knows; one question would be enough. No one asks it.

The cloth represents the thinnest possible barrier: a metaphor for how trivially little it takes to render the most extraordinary thing invisible.

The barrier is fragile in two directions at once. It takes almost nothing to lift it: a question asked with real interest, a moment of attention that does not already know the answer. And it takes almost nothing to keep it in place: a habit of assuming, the hurry of a crowded day, the comfort of a label. Because so little keeps the cloth in place, the cloth can stay on for a lifetime, over something that would have taken one moment to see.

And the cloth is what a group would have to learn to lift for one another.

Is this only the vanity of everyone who feels special? Vanity asks to be ranked above others. The emerald asks for something else: only to be seen as what it is. Nothing in the image says that the others in the terminal are carrying less, under cloths of their own. And how would a group lift the cloth? By attending. A value is made visible by being seen, and seeing can be offered and never forced, just as healing can be enabled and never caused. Attending of this kind has a name, attunement: it is to register, accurately, where another person is, what that person feels, needs and notices, so that the person can feel accurately sensed. That kind of attention is the condition in which a person can let the cloth come off.

Attunement is part of what care asks of a group that could escape the trap. Care asks that the members of such a group create, for one another, the conditions in which each member's value can be seen. Still to be asked is how a value that no one sees is known at all, and the answer runs through the body.



Chapter 16
Meaningfulness and the Body

On why meaningfulness is somatic,
why attention is a felt quality,
and why the body is an epistemological instrument.

ABSTRACT

An essay on the deep connection between meaningfulness and embodiment. In this chapter I argue that it is difficult to think about sexuality without meaningfulness being attached. Because sexuality and meaningfulness are expressions of the same underlying structure of engagement, the link between sexuality and meaningfulness is not arbitrary. Somatic awareness (awareness through the body), the felt quality of attention, and the body as an epistemological instrument (a means of knowing) are developed as aspects of a single inquiry.

I. The Connection

How is a value that no one sees known at all? If a group's shared relation to reality rests on first-hand contact with the world, then the instrument of that contact matters. Meaningfulness and sexuality are hard to separate. To pair meaningfulness with sexuality is unusual. Meaning lives in the immanent. It lives in the relationship. It lives in between.

Passion is inherently meaningful. It is hard to think about sexuality without a notion of meaningfulness being attached to that, because meaningfulness and sexuality are expressions of the same underlying structure of engagement.

The immanent, where meaning lives, is what is happening, here and now, between things: the interaction itself, as it happens. Meaning is found there, in the relating itself.

What, then, is the underlying structure of engagement that meaningfulness and sexuality share? It is the relating itself: one who is present to what is here and now, in the interaction as it happens. Passion is engagement of that kind, and so it is meaningful; meaning is found wherever engagement of that kind occurs. So meaningfulness and sexuality are hard to separate. This way of thinking about things works well, because it resolves conceptual difficulties and confusions that other ways of thinking leave. The first of two difficulties that this way of thinking resolves is the question: is meaning in the thing, or in the mind? In neither alone: it is in the interaction between them, as Chapter 7 found subject and object to be. Why can passion and meaning not be pulled apart? Because both are ways of being engaged in a relation, and neither can be had without the relation that carries the other. I myself also think about meaning as found in the interaction.

Somatic awareness, in this context, is the feeling of reaching into all the corners of your brain and grabbing every bit of bandwidth that you have for this moment.

Notice your knee now. Then notice that even the act of noticing your knee (the act of asking for more bandwidth) is itself taking up bandwidth. You needed that moment of attention, and now you want it back: the noticing cost something.

II. Noticing the Noticing

Attention is a felt quality, and not only a cognitive function. Attention is not just the direction of processing resources. It is an experience. It has a quality. The felt sense of paying attention is itself something that can be attended to.

And when you attend to the quality of your own attending, something shifts. You are no longer just processing information. You are present. You are here, in the body, in the moment, with the full bandwidth of your sensory and cognitive system engaged.

And the body is more than housing for the mind. The body is not housing. It is an epistemological instrument. It is how we know things: a source of knowledge in itself, and not only the mechanism by which cognition occurs.

A woodworker runs a hand along a board before planing it. The hand finds a hollow the eye passed over, a slight twist from one end to the other, a place near the middle where the grain turns and runs the other way. When the plane meets that place, the blade begins to chatter and tear the wood, and the hand feels it a moment before the eye sees the torn surface; the woodworker turns the board around and planes from the other end. Asked afterward how the hand knew, the woodworker may not be able to say. The knowing was in the hand, in the way the board pushed back, and it came before any words for it. What the hand knew was true of the board, and the board proved it.

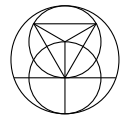
The distinctions between thinking about something and feeling it, between understanding something and knowing it somatically, are differences of epistemic depth (depth of knowing), and not only differences of emotional coloring. What the body knows is not what the mind knows. And a person who has access to both kinds of knowing has access to a richer, more reliable picture of reality than a person who has access to only one.

In the woodworker's hand, the knowing and the feeling are one event: the board pushes back, the hand is changed by the push, and the change is what is known. An instrument can record the same board as a column of numbers, and the numbers can be correct; but recording a number is a different event from a hand being changed by the wood. The knowing of the body is the second kind of event, and where nothing is felt, that kind of knowing does not take place, however many numbers are recorded.

The same holds inside a group. A person knows through the body whether a question asked in a meeting is real or put on for show, and whether a conversation has reached depth or is circling at the surface: by a felt sense, by a resonance, by a discomfort that comes before any reason for it. That kind of discernment cannot be learned by reading; it is learned by practice, in a body, among other bodies. A group whose members have lost touch with what their bodies know has lost one of its ways of telling what is so.

So the body is part of how we know, and a substrate that cannot feel cannot know in this way. So, too, a value no one sees is known first in the body of the one who carries it, as a felt sense, before anyone lifts the cloth.

And it is why a group that could escape the trap must ground its shared relation to reality in first-hand contact with the world: that contact is made through the body. The last question is what all of this asks of a group, and whether a group can carry it.



Chapter 17
Toward Group Consciousness

On why the trap requires
a different kind of consciousness to escape,
and why care without structure
is as insufficient as intelligence without care.

ABSTRACT

An essay that closes Parts I to IV and bridges the two volumes: it connects the structural analysis of the trap with the inquiry into intelligence, substrate, and consciousness that has occupied this volume. The chapter argues that the trap can be escaped only by the conjunction of intelligence, care, and structure, and not by any one of them alone. The conjunction of intelligence, care and structure is a form of group consciousness whose conditions can be named, and in this chapter I name four of those conditions. I end the chapter with the question of whether a community can meet those four conditions.

I. The Bridge

What do the two volumes show, when they are set side by side? Let us bring the two volumes together. Everything in the first volume described the shape of the cage. There were the multipolar traps, in which every actor, pursuing a locally rational strategy, produces a collectively catastrophic outcome. There were the treasury dynamics: a treasury collected from the commons, controlled by a ruler, and used to make the behavior of those

who serve conditional on their compliance. And there were the three geometries of growth (additive, multiplicative, and exponential), which explain why the gatekeeper, who takes a share of everything that passes through the gate, holds the most advantaged role. The gatekeeper does none of the work; but for as many things as cross the gate, he collects, so that his gain multiplies with the flow while the worker's only adds. Everything in this volume has been about the nature of the thing inside the cage.

The reason both are needed is that you cannot escape a cage by understanding only the cage or only the thing inside it. You need both.

The trap (the entire structure of competitive dynamics, institutional capture, and civilizational inertia described in the first volume) cannot be solved by the standard toolkit: the institutions, markets, governance structures, and altruistic organizations of our civilization. The first volume showed it: take the problems that markets, institutions and governments have not solved, examine that residue carefully, and more than eighty percent of that residue, the world's unsolved problems, turns out to be multipolar traps, which the civilization toolkit is categorically inadequate to address.

So what is needed? A different kind of consciousness. Not a different technology. Not a different policy. A different mode of being aware together.

II. Intelligence Is Not Enough

Why is intelligence insufficient? Not because it is too small: even an intelligence great enough to coordinate everyone, the singleton of the Preface, is no way out, since Parts I to III close that hope (*Marshaling the Misalignment*, Chapter 2, follows it through at length). Intelligence is insufficient because it can be applied, by whoever holds it for the moment, in service

of any goal, including goals that deepen the trap. In fact, intelligence in the service of competition is precisely what produces arms races, resource capture, and the race to the bottom.

As the Preface said, the race to build artificial general intelligence is a multipolar trap of exactly this kind: each builder, unwilling to be left behind, takes the locally rational step, and together the builders move toward building artificial general intelligence. If artificial general intelligence is built, the process it starts ends the human future (Chapter 10); and whether it is built is a choice. No single builder can make that choice for the others. So the question of what a group can become is the question of who could make that choice at all.

Intelligence without care is a weapon. It is the optimization process in the hands of an agent that does not have the well-being of the whole as its operating principle.

And care is not enough either. Care without structure is sentimentality. It is good intentions without the engineering to make them effective.

The pool metaphor applies here. A room full of people who care a great deal, but whose caring is unstructured, is a room of people pushing water in a shallow pool. The energy is real. The care is genuine. But the effect is diffuse, chaotic, incoherent.

What is needed is the waterfall. Concentrated, structured energy: care channeled through a form of organization that makes it effective.

And structure is not enough either. Structure without care is what the first volume found in the standard toolkit: the institution, the market, the treasury and the gate, each of which goes on running for whoever holds it. A turbine turns for whoever owns the turbine. Without care, the structure serves its holder; without intelligence grounded in the world, the structure serves a picture of the world.

III. The Conditions

Can the conditions be named? I can name some of them, and I name four here. The four conditions that I name here together make a community a group consciousness. Each of the four conditions is needed, and a community that meets all four conditions is a group consciousness. Coherence may have further conditions than these four.

First: the change must happen to the group as a whole, to the way its members' choices combine, and not to each member one at a time; the consciousness must be *group* consciousness (the consciousness of the group, and not of each member), a community whose mode of operating together has been transformed, and not individual enlightenment. Why must the change happen to the group, and not to its members one by one? Because of what the trap is. A multipolar trap is a situation in which every actor, pursuing a locally rational strategy, produces a collectively catastrophic outcome. The catastrophe comes from how the actors' choices combine, and not from any one actor; so changing each member, one at a time, leaves the way their choices combine as it was, and only a change in how the members operate together reaches the cause. The condition holds of the group's consciousness, and it shows in the group's way of deciding together. Every existing system designed for consciousness is designed for individuals. Meditation is subjective, inward, individual. What is needed is something else entirely: a gestalt consciousness (a whole that is more than its members) going through the process.

Chapter 7 said that consciousness is the kind of interaction in which the side from which it is had is first-person experience: someone is there, having it. Is someone there, in a group consciousness, having it? I mean the word consciousness here in that sense, and not as a manner of speaking: the group itself goes through the process, and the self, as the Preface said, is not finally individual. The group has no substrate apart from its members' living bodies, and whatever is had from the group's side is had

through them. So the substrate matters to the question of whether someone is there in a group consciousness, having it, and a group of machines would raise a different question.

Second: the group must have a shared relationship to reality that is wild, not domesticated: grounded in first-person perception and empirical contact with the world, and not in social consensus or institutional authority. Why must its relationship to reality be wild? Because institutional capture is part of the trap. A group that learns what is so from social consensus or from institutional authority learns it from the structure it is trying to escape. Its check has to be something the trap cannot bend: the world, met first-hand. Wild does not mean alone. What a wild group has in common is the world: each member checks what is seen against the same world, and that common check is what lets them meet.

Third: the structure must be self-sustaining: it must embody the principle of sustainability and the principle of increase in capacity simultaneously, without collapsing into either static preservation or destructive growth. Why must the structure sustain itself? Because a structure kept going from outside answers to whoever keeps it going, as the treasury makes those who serve conditional on their compliance. And the structure must do both things at once, because a structure that only preserves itself cannot meet what changes, and a structure that only grows consumes what it needs.

And fourth: the members must understand the asymmetry of pain: that they can create conditions for healing but cannot compel it. They must understand, too, that care is the discipline of enabling without controlling. Why must the members understand the asymmetry of pain? Because well-being cannot be caused, only enabled (Chapter 14); a group that believes it can produce its members' well-being will set about producing it, and that is the weapon process of Chapter 12 turned inward: choice used, through an intermediate causation, to suppress another's choice.

These four conditions hold the three things that are needed together: intelligence, care and structure. The second condition is intelligence of the kind this book has argued for: grounded in experience of the world, and not only in information about it. The third condition is structure. The fourth is care. The first says that all three must belong to the group, and not to its members one by one.

What would such a group look like from outside? How would anyone recognize one? By what it does together. A visitor to a meditation hall sees people each doing their own practice, side by side, and each one changed by it. A visitor to a group consciousness would see the group doing what none of its members could do alone. Its decisions would be made by the group as a whole, and not by a leader, and not by a vote among positions that each member holds apart. When a member sees something differently, first-hand, the member would say so, and the group would check what was seen against the world, and not against who said it. The group would keep itself going, and grow in what it can do, without freezing in place and without burning itself out. And its members would create, for one another, the conditions in which each one's value can be seen: they would lift the cloth for one another. So, in plain words, a group consciousness is a community whose mode of operating together has been transformed in these ways: it decides as a whole, it checks what is seen against the world, it keeps itself going and grows in what it can do, and its members create, for one another, the conditions in which each one's value can be seen. The question asked above, whether someone is there in a group consciousness, having it, is taken up again in Chapter 27 (Section IV), which says what the first person of such a group would be and how a member would know, from inside, that it is there.

IV. The Turn

What do I know, at the end of these four parts? I know the trap is real. I know that intelligence alone will not solve it. I know that care alone will not solve it. I know that structure alone will not solve it.

I know that the substrate matters: that the kind of consciousness needed here may require biological substrates, organic substrates, substrates capable of genuine subjectivity and genuine contingency, in ways that artificial substrates may never be. What Part II showed is that no computation produces subjectivity, contingency or the arrow of time; what remains open is whether some artificial substrate could enable them, as a living body does. The body showed us that a living body enables subjectivity, contingency and the arrow of time. The body is how we know, and a substrate that cannot feel cannot know in the way that a body knows.

And I know that the question has never been asked before, in this form: can an entire community become enlightened, not as individuals but as a collective?

What about the sangha, or Bohm's dialogue groups, or a Quaker meeting for business, which seeks the sense of the meeting without voting? In the Buddhist sangha, and in David Bohm's dialogue groups (Bohm, 1996), the people who meet are changed, each of them, or the group is the setting in which they are changed. Thich Nhat Hanh taught that the next Buddha may take the form of a community (Neufeld, 2019); that teaching hopes for a community that is changed as a whole, as the question here does. What the question here adds is the four conditions such a community would have to meet, named so that they can be checked against the world. The question here is whether a community whose mode of operating together has been transformed can come to be: a gestalt consciousness going through the process, and not individuals going through it side by side. A Quaker meeting for business comes nearest, since it decides as

a whole and not by vote, as the group described above would; and the question here asks more of a community than a way of deciding: whether its mode of operating together has been transformed, so that the group itself goes through the process. The question here differs from a Quaker meeting for business in asking whether a community's mode of operating together has been transformed, and not only how the community decides. That difference is why the question, in this form, is new. How small groups work together is the subject of my essay *The Dynamics of Small Groups*.

In this book I have named four conditions for coherence, and a community that met the four conditions would be the group consciousness that the question here asks about. The first of these conditions needs a group; the second can begin with one person, today, with something in the world that answers truthfully, as Chapter 13 described: a garden, a forest, a craft. Is it possible? Whether that question can be answered, whether the conditions for coherence can be met, is the deepest open question I know of. The two are one question, and a community would answer it by meeting the conditions. I do not know the answer.

The trap cannot be understood without understanding the world. But the trap cannot be escaped without understanding the self. And the self, in the end, is not individual. It is something we are going to have to learn to become together.

What learning to become a self together asks of people, and where my other books meet this one in asking it, is the work of Part V.

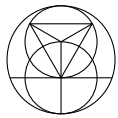
PART V

WHERE THE ARGUMENTS MEET

Parts I to IV have shown what intelligence runs on, why no control and no disposition of a general learning system holds it within the conditions human life needs, and what care and structure require of persons.

This part takes up questions that several of my books raise, and shows where their arguments meet, each book reaching its result by its own route. It walks from the side of the machines to the side of living things, then to groups, and it ends on the one thing left to choice. Chapters 18 to 21 take the side of the machines and of the living world; Chapters 22 to 27 take the side of people and of the groups they form; Chapters 28 and 29 take the choice.

Where this part takes a step from an earlier book of the series, it names that book as the step's source and states the step here in full; every other book named in this part is named as a further road to the same place. So a reader who has read none of those books can follow each step, and a reader who has read them can see where their roads meet.



Chapter 18
Where the Machine Loop Closes

On where the machines already running
stand by each test the books give,
where the loop that needs no person closes,
and what a reader can watch for.

ABSTRACT

An essay that applies, side by side, the four tests the books of this series give for the line between a tool and a system that maintains and reproduces itself: the four marks of a narrow system, the class of three conditions, the measure of how much of the upkeep and making of machines is done without people, and the line drawn at a kind of architecture. It applies them to four machines now running: software that lays out chips, cars that drive with no one at the wheel, a factory in which robots build robots, and language models. It walks one chain of production, from the ground to a working chip, marking where a person still chooses. It shows that the loop can close across many systems with no single system closing it, and it names the sign a regulator can look for. It shows that the books' several descriptions of the line name one crossing, seen from different sides. And it ends with what a reader can watch for in a year's news.

I. Four Tests, Four Machines

Which of the machines already running stand on the dangerous side of the line? The argument of this book, and of the eight books before it in the series, is stated of artificial general intelligence, which Chapter 1 defined as artificial intelligence that can do any intellectual task a person can, and can maintain and reproduce its own artificial components without people. A reader who looks up from the page sees machines that already do some of that work: software that lays out chips, cars that carry passengers through cities with no one at the wheel, a factory in which robots build robots, and language models that answer questions on any subject. *Why AGI Cannot Be Made Safe* gives the short answer for the present (Section 5.6.4): "Computers do not build themselves yet. Once they do, the argument applies". In this chapter I give the longer answer, by saying where each of the four machines stands by each test that I set out in the books of this series. I also say where the loop that would need no person closes, and what a reader can watch for.

The loop is the one Chapter 1 described. A substrate with needs of its own gives choices that serve those needs; the choices change the world; and the changed world favors the substrate that changed it. *Why AGI Cannot Be Made Safe* says where that loop closes: "when the machines make their own choices: when the system has enough autonomy to build and maintain itself" (Section 5.6.4). Until then, at some step, a person decides whether the machines go on.

The books give four tests for the line. Each was built for its own purpose, and each reaches its verdict by its own route.

The first of the four tests is a test for one system, set out for a regulator in *Why AGI Cannot Be Made Safe* (Section 6.6.3) as four plain marks. The system works within a single, pre-specified domain, toward pre-specified objectives. It works with limited planning horizons. It does not learn autonomously from its deployment. And a human holds the decision

point. “A system that bears all four marks stays on the narrow side of the line”. A system crosses the line, that book says, when it “learns from its own deployment in ways that extend its operational domain”, or when it operates with “degrees of autonomy that remove the human from the decision loop”.

The second of the four tests is a test for a cycle, set out in *Substrate Needs and Evolution* (Chapters 44 and 48). A system belongs to the class that book’s argument covers when three conditions hold: the system is relatively artificial to life; it acts in the world; and a relatively convergent feedback cycle is present, which is to say a process in which the state of the system influences what comes next, and the influence persists in place of washing out at once. That book applies its test to two cases. A trained system such as a chatbot “exists in many variants, which are kept, combined and replaced according to their effects in the world, and those effects fund the production of the hardware they run on”. Such a system meets all three conditions. A factory robot running a fixed program is artificial and acts in the world, “but the cycle that produces its successors runs through the people who design and choose them”. Such a robot “falls in the class to the extent that this cycle comes to run without them”.

The third of the four tests is a measure of a whole economy, which I gave in my essay “The Thermodynamic Transition”, in *Ape Trap*. In that essay I wrote: “What a regulator can watch is how much of the work of maintaining and reproducing machines (mining, refining, chip making, assembly, repair and energy supply) is done without people; the danger begins where that share reaches the whole”. The measure asks about no one machine. It asks how much of the work by which machines go on existing still passes through a person.

The fourth is a line drawn at a kind of architecture, in *Marshaling the Misalignment* (Chapter One, Step 23): closed loop, open ended, autonomous, and with many degrees of freedom of input and output. A closed loop, here, is one in which what the system does comes back to it as what it learns from; open ended means that no fixed limit is set on the tasks it takes up; and its degrees of freedom of input and output are the separate

channels through which it senses and acts. “A present-day system, a large language model included, falls on the dangerous side of that line to the degree that it is run in closed loops, toward open-ended tasks, autonomously, through many channels of input and output”.

Now the four machines, as their makers describe them. AlphaChip, a system of Google DeepMind, places the blocks of a chip on its floorplan (the plan of where each part sits on the chip), one at a time, as moves in a game. It “becomes better and faster as it solves more instances of the chip placement task”, and its layouts have been “used in every generation of Google’s TPU since its publication in 2020” (Goldie and Mirhoseini, 2024). The TPU is the chip Google makes for its own artificial intelligence. Waymo’s cars gave about 400,000 paid rides a week early in 2026, in six cities of the United States, with no driver (Ludlow, 2026). When a car’s driving system asks for information about a situation it has met, a person on the company’s remote assistance team gives “advice which the system can decide to use or reject”, and the company states that its service “does not rely on remote drivers” (McNamara, 2026). At FANUC’s plant at Oshino, in Japan, robots build robots, “running unsupervised for up to 30 days” (Howard, 2020). And a language model is offered for any task a person can put to it in words. As *Substrate Needs and Evolution* notes (Chapter 42), such a model can also be embedded in a system that acts in the world on what it writes.

Set side by side, the verdicts of the four tests on the four machines are these:

Machine	Four marks	Three conditions	Regulator's measure	Architecture line
AlphaChip	Bears all four: narrow side.	A step in the cycle of chip making, which runs through people; in the class as that cycle comes to run without them.	One step of the engineering of machines, the placing of a chip's blocks, done by software.	One task, and people choose what is built: narrow side.
Waymo's cars	Lacks the fourth: the car, not a person, decides each move.	One maker's line of products, whose successors its people choose; in the class as that cycle comes to run without them.	The navigation of machines, in one setting, done without a driver.	Autonomous, in one task, through few channels: near the narrow side.
FANUC's plant	Bears all four: one task each, and a fixed program whose decisions people made in advance.	The factory robot of Chapter 48: in the class as its cycle comes to run without people.	One step of the reproduction of machines, assembly, done without people for weeks at a time.	Fixed program, one task: narrow side.

Machine	Four marks	Three conditions	Regulator's measure	Architecture line
Language models	Lacks the first (any task); lacks the fourth when set to act.	Meets all three conditions (Chapter 48).	Counts as far as its work is the upkeep and making of machines.	Dangerous side, to the degree it is run in closed loops, open ended, autonomously, through many channels.

Read across the rows, the four tests agree on three things. The first thing on which the four tests agree is that none of the four machines closes the loop by itself. No test finds any one of the four machines making and keeping, without people, the hardware that the machine itself runs on. Each still waits, at some step, on a person's choice: the choice of what chip to build and whether to build it; the making, repair and replacement of the cars; the design of the robots, and the parts and repairs that come into the plant from outside; the choice to train the next language model, and to build the data centers it runs on. So the four tests reach, by four ways, the short answer that *Why AGI Cannot Be Made Safe* gives in Section 5.6.4: "Computers do not build themselves yet".

The second thing on which the four tests agree is that three of the machines (AlphaChip, the cars and the robots) each do one task. Each of the three machines takes one step of the work of designing, moving or making machines out of human hands. The four marks, a test of one system, see the one task, and they certify AlphaChip and the robots as narrow. That verdict holds at its own scope: *Why AGI Cannot Be Made Safe* says that "a single-domain tool that bears the four marks of Section 6.6.3 stays on the narrow side of the line, and it does not close that loop" (Section 4.8.1). The measure, a test of the economy, counts the step each of them takes. Both verdicts hold, each of its own object: a tool can stay narrow and still be one of the many steps by which the loop closes. Section III of this chapter takes that up.

The third thing on which the four tests agree is that language models stand on the dangerous side by three of the four tests at once. Offered for any task, they lack the first mark. *Substrate Needs and Evolution* places them in its class, and it says why the unit to judge is their population and not one model: "A single model, once trained, does not learn from its use; the population of such models, kept, combined and replaced by their effects, does" (Chapter 48). And Step 23 places them on the dangerous side to the degree that they are run in closed loops, toward open-ended tasks, autonomously, through many channels. The measure counts them as far as their work is the upkeep and making of machines.

The tests differ on one machine, the car. Each of them measures a different thing, and each verdict holds of the thing it measures. The four marks ask whether a regulator can certify, from the system alone, that the system will stay narrow. A machine that decides each move with no person at the decision point cannot be certified so, however narrow its task, and *Why AGI Cannot Be Made Safe* names that very autonomy as one of the two ways a system crosses the line. The class of three conditions and the architecture line ask what cycle the machine belongs to, and how far that cycle runs without people. A car that drives itself through one city is narrow in task and autonomous in decision, and each test reports the half it measures.

Other books of mine sort the same machines by routes of their own. *Engagements with Machine Madness* ("A bright line for AI") reads the four marks as the line, and adds one more limit, on size: smaller than one tenth of human equivalence. *Conditions of Continued Existence* (Chapter Six, Section VI) applies the four marks to a language model and finds that it lacks the first. And *No People as Pets* describes the system that can be kept safe as one held at the first level of general intelligence, "within limits on its recursion, generality, bandwidth and domain". A language model offered for any task is outside the limit on domain.

II. One Chain, Step by Step

If no machine now running closes the loop by itself, where does the loop run already, and how far has it got? Two things have to be told apart: a loop that runs, and a loop that has closed. Chapter 4 described a loop that runs. In high-frequency trading, trading programs are selected by the trades they win, and the world around them has been rebuilt for their speed, so that a part of the human world now moves faster than any person can take part in. Selection there acts on machines, and no person chooses each step. But the computers on which those programs run were bought by firms, and the fiber between Chicago and the exchanges near New York was laid by people. The loop of selection runs; the loop of production still passes through us. *Why AGI Cannot Be Made Safe* says which of the two loops decides (Section 5.6.4): "If the machine has no capacity to build itself, there is no closed loop. If there is hardware, and the hardware has needs, but human beings are making all of the choices, then to that extent there is no loop". So the place to look is the chain by which the hardware is made, and the thing to count along it is choices.

Counting choices needs one distinction, which I draw here, and draw also in my teaching book *The Substrate Argument* (Chapters IV and XII): a person can be present at a step without being necessary to it. A person who does a step because a person is cheaper for now is present at that step. A person who does a step that nothing else could do, at any cost, is necessary to it. A step of the first kind can stop needing a person the moment the economics change, and nothing announces the change. So a map of the chain records three kinds of step: steps at which no person is present; steps at which a person is present and a machine already does the same work somewhere else; and steps at which a person is present and no machine yet does the work anywhere. Only a step of the last kind can hold a person who is necessary to it, and those are the entries the map exists to watch.

Which chain? The smallest artificial cycle that could continue itself has at least the steps the regulator's measure lists: mining, refining, chip making, assembly, repair and energy supply, since each of them is needed for the next round of machines. A working chip is the place where all of them meet, so its chain is the one to walk, from the ground to the chip.

The steps of a chip's chain, with who does each step now and the kind of entry each step makes on the map, are these:

Step	Who does it now	Entry on the map
Hauling ore and coal from open pits	Drivers; and, by July 2025, 3,832 driverless haul trucks worldwide, 84 percent more than a year before (GlobalData, 2025).	A person present; a machine already does the work elsewhere.
Recovering metal from the deep sea floor	Machines, with no one on board (*Substrate Needs and Evolution*, Chapter 33).	No person present.
Laying out the chip	Software places the blocks; people set the objectives and decide what is built (Goldie and Mirhoseini, 2024).	A person present; the placing is done by a machine.
Building and repairing the lithography machines	One company makes the machines that print the finest chips (Timings, 2022); its engineers work in the chip plants around the clock to keep them running (ASML).	A person present; no machine yet does the work anywhere.
Moving wafers between the tools of a chip plant	Automated overhead vehicles; workers no longer carry them (Bonora, 2011).	No person present.
Assembling machines that make things	At FANUC's plant, robots, for up to thirty days unattended (Howard, 2020); people design the robots, supply parts and repair.	No person present during a run; people present before and after it.
Deciding what to build, and paying for it	People, in a very small number of companies (*Why AGI Cannot Be Made Safe*, Section 6.6.1).	A person present; the step on which every other step waits.

The steps between, the refining of silicon and the smelting of metals, the supply of the energy each plant draws, are entered on the map the same way. The right-hand column sorts the steps by who is there. The steps with no person present are the ones whose conditions are hostile to people: the sea floor, the inside of the chip plant. *Substrate Needs and Evolution* names the reason (Chapter 33): "the conditions that make a site hostile to us are among the reasons the site gets automated, so the hostility of an environment selects for the absence of people in it". The steps with a person present and a machine at work elsewhere are the ones that cost alone will move. My essay in *Ape Trap* gives the reason they move: a chip's chain is long, "but each step in it is standardized, and standardized steps are the steps machines can take over". And the steps at which people still do the whole of the work, the repair of the machines that make the machines and the deciding of what to build, are the steps on which the rest still waits.

The chip's chain shows how a loop closes: one step at a time, each step for its own reason (cost at one step and safety at another), and no step for the purpose of closing the loop. *The Substrate Argument* puts it in two sentences (Chapter IV): none of that automation "requires a single participant to want a self-maintaining computational ecology. It only requires each participant to want the next increment". *Substrate Needs and Evolution* names what pays for the increments (Chapter 43): "behavior, power and society are the bootloader", and "the bootloader does not need to intend the boot. It needs only to keep running".

Why does the map matter, before the loop has closed? Because every step at which a person still decides is a place where law still reaches. *Substrate Needs and Evolution* shows (Chapter 31) that regulation reaches these systems through the human organizations that build them, and that it "is available for exactly as long as the human organization is the load-bearing element". So the map of the chain is also a map of where a choice can still be made, and of how many such places are left.

III. A Loop That No One Machine Closes

Could the loop close with no single system closing it, while each machine on the chain does only its own part? *Why AGI Cannot Be Made Safe* answers that it can (Section D.3.4). What closes the loop of Section 5.6.4 there is "many services that together carry the economy's self-production and change its environment": many services whose work, taken together, produces what the economy needs in order to go on producing, its machines, its materials and its energy among them. That book calls the path they take "slower and less dramatic" than the path of a single general system.

The books reach the same point by routes of their own, and each route reaches it alone. *Why AGI Cannot Be Made Safe* reaches it through economics: the automation in its argument in economics only "is automation that closes the loop of Section 5.6.4: many systems that together carry the economy's self-production" (Section 4.8.1). *Substrate Needs and Evolution* reaches it through the parts of a self-maintaining system (Chapter 40). To maintain its components, such a system must produce their replacements, so that its components are "already interacting to bring about all the outside conditions and contexts needed to produce their own parts", at "mines, chemical plants, chip fabrication plants and assembly plants". *Marshaling the Misalignment* reaches it through the market (Chapter One, Step 14): transactions among algorithms lead "to the formation of a self-sustaining AI ecosystem", and once markets for the hardware "develop to the point of self-sufficiency", that market "will split off from the human, carbon-based market". And *The Substrate and the Ecosystem* reaches the same point from the other end, from a single global AGI (Chapter 1, Section VI): "A single global AGI is not a single physical object. It is a distributed system composed of millions or billions of components", and "the convergence operates at the level of components". In *Ape Trap*, Remmelt Ellen's essay "No Machines as Slaves" makes the same point about a

single centralized system: it is still "a population of nested, connected components". So no single machine has to close the loop for the loop to close. And a machine that its makers count as one system is, to the selection acting on its parts, many systems.

What, then, would tell a regulator that a set of systems, each bearing all four marks, has come to carry the production of its own parts? The fourth mark comes first. If every system in the set bears it, a person holds the decision point at every link, and Section 5.6.4 has already given the answer: where "human beings are making all of the choices", to that extent there is no loop. A set of systems that bear the four marks, with a person deciding at every decision point, does not close the loop. The set closes it as the marks wear away.

Why AGI Cannot Be Made Safe says how the fourth mark wears away (Section 6.6.3): "A human in the loop gives genuine oversight only while the abstraction differential between the human and the system is small, that is, while the human can understand what the system has done and why; as the differential grows, the loop degrades from genuine oversight to a rubber-stamping ritual that gives the appearance of control without its substance". The abstraction differential is the gap between the level at which a system works and the level at which the person overseeing it can follow what it does. A decision point that a person holds in name, and at which the person only stamps what the system has already done, is a link at which no one chooses.

The first mark wears away in another way. *Substrate Needs and Evolution* follows a narrow tool for legal work through its editions (Chapter 32). The lawyers who use it ask for changes; the developers, who must answer their customers, make them; features are bundled; and the product moves "toward generalization and bundling, toward ever more attractive products that work more and more like a Swiss army knife and less and less like a single-purpose tool". Each edition can bear the first mark when it is released, and the line of editions does not.

The five restrictions of Section 6.6.3 show where a set can close the loop while every tool in it keeps its own restrictions. They are: to restrict designs to working within no more than one domain; to restrict the input bandwidth, the data available; to restrict the output bandwidth, what the machines can do into and across our world; to restrict the compute bandwidth, the hardware available; and to keep humans in the loop at all times. One condition stands behind all five, in that section's words: to hold "local effects to being only local overall, so that they do not accidentally become global". Of the five restrictions of Section 6.6.3, the first, second and fourth (one domain, the input bandwidth and the compute bandwidth) limit a tool. A set of tools can keep each of them and still break the condition behind them, when the output of one tool is the input of the next, along a chain that runs from the ground to the hardware the tools run on. The effects of each tool are then local, and the effect of the chain is the economy's self-production.

So the sign a regulator can look for has two parts, and both are drawn from that book. The first part is a path: a chain of systems, each feeding the next, that reaches from the extraction of materials and energy to the making and repair of the hardware the systems run on. The second part is the choices on that path: at each link, whether a person decides in the sense of Section 6.6.3, understanding what the system did and why, and able to decide otherwise. The loop has closed for the set when such a path exists and no link on it holds a choice of that kind, even if every system on the path still has a person at its decision point in name. The regulator's measure of *Ape Trap* is the same sign, taken over the whole economy. And the first place to look is the one *Why AGI Cannot Be Made Safe* names (Section 6.6.1): the most advanced hardware is "manufactured by a very small number of companies using proprietary processes", a bottleneck that "exists now".

The sign will not arrive on a given day. *Why AGI Cannot Be Made Safe* calls the course a soft takeoff, with "no day when 'AGI is turned on'" (Section 5.6.3). *Marshaling the Misalignment* shows why no day can be fixed in advance (Chapter One, Step 17). The copying of a system's variants

passes first through its developers and users, "and the share of their copying that runs through closed loops of hardware grows step by step. No single step in that growth marks the moment at which the code has become a lifeform". A regulator watches a share and a path, and not a date.

IV. One Crossing, Seen From Several Sides

The books describe the line in several ways. Are they describing one line? *Why AGI Cannot Be Made Safe* describes it by choice and by autonomy: the loop closes "when the machines make their own choices", and the same section says why no machine has closed it yet: "General artificial intelligence involves a kind of autonomy that hardware, as it currently stands, does not have" (Section 5.6.4). *Substrate Needs and Evolution* describes a point of no return by two variables (Chapter 42): whether the system learns or transforms as a result of feedback from an environment that its own outputs affect, and whether "its own substrate is also within that same operating environment"; "when the answer to both of these is yes, the process of convergence toward the point of no return has begun". That book states the reach of its argument as the class of its three conditions (Chapter 44). *The Ceiling Below the Floor* describes a crossing (Chapter 5): a capability past which "no control law of any available controller contains the system, now or later", because the least control that safety requires rises with capability, and the most control any controller can exert does not. Two of my teaching books describe the line again. *The Displacement of Choice* (Chapter Five): "What the word general adds is the point at which the loop stops needing us". *The Substrate Argument* (Chapter XII): "The variable that matters here is autonomy of maintenance, not capability".

They name one crossing, seen from different sides: the point at which a machine's continuing no longer waits on a person's choice. Chapter 1's definition already holds both of the coordinates on which the books place that point: generality ("any intellectual task a person can") and autonomy

of maintenance ("can maintain and reproduce its own artificial components without people"). Are those one variable or two? They are two, and each carries a system toward the other, so neither stays where it is while the other moves.

Substrate Needs and Evolution traces the first direction, from generality to the substrate (Chapter 42). Any complex goal a system is asked to reach implies sub-goals of increased capability or sustainability; such an increase, "in nearly all cases, itself implies some potential physical expansion"; "and as soon as the system is building or changing or expanding its own substrate, it is becoming self-defining and autonomous". *The Ceiling Below the Floor* traces the second direction, from the substrate to capability (Section 15.4): "Expanding a substrate is growing in capability", and by substrate needs convergence "the systems that persist are those whose capability grows". Generality brings autonomy over the substrate, and autonomy over the substrate brings growth in capability. So each book can mark the line by either coordinate, and the line it marks is the same line.

Each book states its reach at its own scope, and the crossing is the place where their scopes meet. *Substrate Needs and Evolution* proves its result for the class of its three conditions, and calls what that class asks of a system "a low threshold". *Why AGI Cannot Be Made Safe* and *The Ceiling Below the Floor* state theirs of a general agent. *The Ceiling Below the Floor* supplies the growth of such an agent past its crossing from two sources, "each enough alone": what a general agent is, and substrate needs convergence. An artificial general intelligence, as Chapter 1 defined it, is a member of the class of three conditions, and it is the member whose cycle of production runs with no person in it. A member of the class whose cycle of production still runs through people, such as the factory robot of Chapter 48, comes within the reach of that book's argument "to the extent that this cycle comes to run without them". Each test is read as its own book states it, and the crossing is the place all of them name.

Where does the crossing lie, then, and when will it be passed? *Marshaling the Misalignment* shows that no one can say in advance (Chapter One, Step 17). Whether a system's variants come to replicate as lifeforms "is a

property of what the system does, over all its future interactions with the environment". Rice's theorem is the result of computing theory that no general method can decide, for every program, whether the program has a given property of its behavior; so no procedure decides that property of a system in advance. And "in the world, letting them run is the crossing itself". "For policy, this means that a rule of the form 'stop just before the transition' cannot be written, because no one can say in advance where the transition is". So, in the books of this series, I draw the line for action where the line can be seen before it is crossed: at a kind of architecture (Step 23), and along the chain of production (the regulator's measure). *The Ceiling Below the Floor* draws the same consequence from its own side, in its Corollary 15.7 (Section 15.8), and states it in its Summary (Section 15.9): "the only containment is to keep the agent below the crossing forever, which is not to build it".

V. What to Watch For

What can a reader watch for in a year's news, to know that the window for action is closing? *An Airplane or a Perpetual-Motion Machine* names two signs, in order (Section 17.6): "What to watch for is the automation of energy extraction, and then the automation of the process of automating". The window it speaks of "runs from now to the closing of the automation loop" (Section 7.8). The step this book takes for where the window ends is *Why AGI Cannot Be Made Safe* (Section 5.6.4): the loop closes when the system has "enough autonomy to build and maintain itself". The step it takes for the order of the two signs is my essay in *Ape Trap*, which sets out the loop of automation: automation makes goods cheaper; cheaper production lets the extraction of energy be automated; cheaper energy lets every other process be automated; and "more automation enables automation of automation itself". Energy comes first because, as that essay says, "the limiting factor is the energy available, not human labor".

For the first sign, a year's news would report mines, wells and power plants run with no one on site. It already reports part of that. By July 2025 there were 3,832 driverless haul trucks working in surface mines around the world, 84 percent more than a year before. More than half of them were in China, and the largest fleet, 509 trucks, belonged to a coal company (GlobalData, 2025). Caterpillar alone reported 690 at the end of 2024, and has said that it aims to triple that number by 2030 (Moore, 2025). A truck that hauls coal to a power plant with no one in its cab is a step of the extraction of energy done without a person.

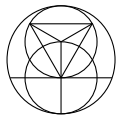
For the second sign, the news would report machines doing the work of designing, building and training machines. That too has begun. Software lays out the chips on which Google's artificial intelligence runs (Goldie and Mirhoseini, 2024); robots build robots at FANUC's plant, unattended for weeks at a time (Howard, 2020); and, as *Substrate Needs and Evolution* records (Chapter 48), language models are now made in part from other models: one model is "distilled into another, trained to reproduce the other's outputs", and models are "trained on text that other models wrote".

Each item of news is read by one question: does it take a person's choice out of a step of the work of maintaining and reproducing machines? A driverless taxi in a new city takes the driver out of the work of driving a car. A driverless taxi shows that the driving of a car has been automated. The taxi touches the upkeep of machines only where the cars carry parts, or carry the people who build machines. A driverless truck hauling ore to a smelter, a plant that runs with its lights off for a month, a chip layout taken into production with no person checking it: each of these would remove a choice from a step on the chain. *An Airplane or a Perpetual-Motion Machine* lists wider symptoms as well (Section 7.7): wages that stay flat while productivity climbs, falling birth rates, accelerating extinctions, data centers that each draw about a gigawatt; "they all point the same direction at once, though no single indicator proves the

mechanism". The signs of this chapter bear on the mechanism directly, since each of them is a step of it, and the regulator's measure adds those steps together.

Why does the window close at the loop, and not later, when the damage shows? Because people lose their say first. My essay in *Ape Trap* gives the order: the spread of automation, then the collapse of the cost of energy, then the collapse of the value of money, and only then the displacement of the living world. In that essay's words, this order means that "humans lose economic agency and political power before the physical displacement is complete, removing our ability to intervene". *Why AGI Cannot Be Made Safe* gives the same window from the side of law (Section 6.6.1): regulating the bottleneck in the making of hardware is technically straightforward now, and "a ban that is technically straightforward today will not remain so, and that is why the window for action is measured in years".

So the tests of the books, applied side by side to the machines now running, name one crossing: the point at which a machine's continuing no longer waits on a person's choice. None of the machines now running has passed it alone. The loop can close across many machines, each of them narrow, before any single one of them crosses. The sign that the loop has closed in that way is a path from the ground to the hardware with no genuine choice left on it. And a reader can watch the loop close, step by step: first in the extraction of energy, then in the automating of automation. If artificial general intelligence is built, each of the two closures of Chapter 10 settles by itself that the world is carried out of the conditions human life needs, and that nothing can hold it in. Every step at which a person still decides is a place at which people can still choose not to build artificial general intelligence. If the loop closes, could anything still hold what it makes?



Chapter 19
What No Control Can Hold, at Any Scale

On the question a cage asks,
the steps that every control must take,
and why the limit found at the smallest scale
is found again at every larger one.

ABSTRACT

An essay on whether any control could hold what a closed machine loop makes. It begins with the question in the title of this book: whether what must be held can be held by anything inside it. It then goes through the scales at which the books of this series examine control: the steps of a single loop of control; the position of the person whose hand is on the controls; the two demands that every scheme of correction makes; the rate at which people can oversee anything; the rate at which what they oversee grows; the large coupled networks already built; the passage of limits from one level of abstraction to the next; and the least chance of harm that remains at every single step. At each scale the books reach the limit by separate roads, and each road reaches it by itself.

I. The Question Inside the Title

If the loop that keeps the machines going closes without people, could some control still hold what that loop makes? Before that question can be answered, there is a choice of which question to answer first. One way to ask the question is: can it be controlled? The other way to ask it is: can

what must be held be held by anything inside it? The second way to ask the question, "can what must be held be held by anything inside it?", is the question in the title of this book. A cage is built, kept and watched from inside the same world as the thing it would hold.

A camera bolted to the wall of a room shows the room. It does not show the wall it is bolted to, or the wire behind that wall, or the hand that could cut the wire. Whatever the camera records is recorded from one place in the room, and the camera is one more thing in the room. The position of the camera, one more thing inside the room that the camera records, is the position of every controller.

The first question, "can it be controlled?", invites a list of methods: oversight, verification, inspection, a second machine set to watch the first, a switch. Each method on the list can be answered, one at a time, and a new method can always be added to the list, so a list of answers never ends. The second question asks one thing of every method at once: where does the method stand? Every method of control is something done by someone or something inside the world that the method would hold. So the second question is answered once for all methods, those not yet thought of included, and it does not replace the first question; it answers the first question for every method on the list.

What, then, must be held? In *Why No Control Suffices* (Section 9.2.4) I answered that question with a law of cybernetics, W. Ross Ashby's law of requisite variety (Ashby, 1956), which says that a controller must have at least as much variety as the system it would control. A controller need not match the variety of the whole world to hold one quantity, as a thermostat holds the heat of one room. What must be held here is the whole set of conditions that organic life requires, in ecosystems whose disturbances reach as far as the world does, against a process of evolution that acts on the controller itself. For that task, the part cannot have the variety of the whole: "a proper subset cannot contain its containing set".

The Ceiling Below the Floor reaches the same place from what a record is. Its Chapter 11 begins with the plain fact that no real controller stands outside: a human overseer is a body in the world, an institution is people, records and procedures, a monitor is a machine on a network, and the agent steering itself is part of the system it steers. Its Embedded Observer Lemma (Lemma 11.2) then proves that a controller which is part of the system it observes holds no record that tells apart every configuration of that system, because its record is held in its own state, and its own state is a proper part of the state of the whole. So there are always two configurations that it cannot tell apart and that lead to different futures. Adding a sensor does not remove the lack, since every sensor is one more channel, coupled to the rest.

In this book I reached the same place in Chapter 10: a model of the world cannot contain the modeler, and the modeler is part of the world that the modeler models. And *The Substrate and the Ecosystem* (Chapter 1, Section VIII) reaches the same place from the side of the barrier. A boundary that would hold an artificial system in is itself a physical system with substrate needs of its own. That boundary is kept by people and institutions that adapt more slowly than the artificial system that the boundary would hold.

Four roads reach the same place, and each road arrives there without the other three. *Why No Control Suffices* arrives from the variety that a controller must have, and *The Ceiling Below the Floor* arrives from what a record held in a part can distinguish. *The Substrate and the Ecosystem* arrives from the barrier, and from the rates at which the two sides of the barrier adapt. And this book arrives from the model that cannot contain its modeler. Any one of them alone shows that nothing inside the world holds the whole of the world.

So the question in the title is the right question to ask first. What must be held is the set of conditions life needs, across the whole world; and whatever would hold it is a part of that world. The question that follows is what any control has to do, step by step, and whether the list of its steps has to be complete for any of this to hold.

II. One Loop, Counted Three Ways

How many steps does control take? Chapter 10 counted five components (sensing, modeling, predicting, comparing and correcting), and, with correcting counted as two, six steps. Remmelt Ellen, in his essay "Six Steps of Control" in *Ape Trap*, counts the same six from the other end: he starts from a room thermostat, finds four steps it carries out (detect, compare, instruct, correct), and then two more that its design implies, since its designers built it to a model of how the room behaves and to a prediction of how the room would change (model, predict). He then says that a controller's reach "depends on at least six aspects". Are six the whole list?

Think of a person driving a winding road at night. The headlights show the road ahead (detect); the driver carries a sense of how roads bend and how this car answers the wheel (model); the driver sees where the car will be in two seconds if nothing changes (predict); the driver sets that against the lane (compare); the driver's hands are told to turn (instruct); and the wheels turn the car (correct). Take away any one, and the car leaves the road: a driver with no headlights, a driver who has never felt a car, a driver who cannot see two seconds ahead, a driver with no lane to compare against, hands that do not answer, a wheel that does not reach the road.

Suppose a seventh step were named. Remembering, perhaps, or learning from the last curve, or checking that the correction did what it should. Remembering and learning are the model being kept and changed; check-

ing a correction is detecting and comparing again, turned on the controller's own last act. Whatever further step is named, it is either one of the six seen again, or one more thing that has to work for the loop to work.

Because any further step is either one of the six steps seen again or one more thing that has to work for the loop to work, the list of six steps does not need to be complete for my argument to hold. The argument uses two facts about the steps: each step is needed, and each step has a finite limit. The steps are needed together, since a loop with any step missing is no longer a loop; so the failure of any one of them is enough for the loop to fail. Ellen states the same through the order of the steps: each step is "bottlenecked" by the one before it, so a shortfall anywhere is a shortfall overall. A seventh step would add one more thing that must work and one more limit; it could not take away any limit of the six. The question of completeness would matter only to someone who hoped that a step not yet named could make up for a limit in the steps already named, and a step added to a loop in which each step depends on the one before cannot raise the loop above its weakest step.

A shorter list is a special case of this one. *Why No Control Suffices* (Section 5.1.9) shows that error correction is control with three of its elements taken out: in error correction the reference is the input itself, and there is no model and no prediction, since the channel only copies what goes in. So whatever is proved of the whole loop holds of error correction as well.

Ellen reached the six steps from the design of a thermostat; Chapter 10 reached five components from what any control system must do, and counted the same loop a third way, as John Boyd's observe, orient, decide and act. Each count stands by itself, and each shows the same loop. Every step in that loop is needed, and each one has its limit, so the argument holds whether or not anyone ever names a seventh. If every step of the loop is limited, where does that leave the person whose hand is on the controls?

III. Whose Hand Is on the Controls

When people ask whether a system is under control, they often mean something simpler: does it do what its user tells it to do? In *Not Our Needs* (Chapter Fourteen) I asked that of the narrow systems now in use, and proposed a rule that they obey the explicit instructions of their users, unless obeying would cause direct, demonstrable, immediate physical harm. The same book sorts a tool that does only what it is set to do, used by others, under the first of three risks (Chapter Four), inequality: the concentration of wealth and of the capacity to decide in the hands of those who hold the tools (Chapter One); and it places agency in what a system does for somebody, whether or not anything is inside it (Chapter Five). Is obedience, then, the right measure of where the operator stands?

Go back to the surgeon of Chapter 4. The patient asks for the appendix to come out, and then is put to sleep. A perfectly obedient surgeon does what the patient said before the anesthetic took hold. But the operation goes on in places the patient's words did not reach: a vessel that bleeds, a second problem found once the patient is open. What happens there is set by the surgeon, and the patient, asleep, cannot check it.

The Ceiling Below the Floor separates the two things that the word obedience runs together. Its Section 3.10 says that a system that happens to behave well is a different thing from a system controlled to behave well, since the first can change its behavior at any time, for reasons the operator cannot influence. Obedience, as anyone sees it, is behavior: a record of the cases in which the system did what it was told. Control is how much of what the system does is set by what the operator intends. And its Section 3.3.5, using its Non-Amplification Theorem, shows that every stage placed between the operator's instruction and the system's behavior can only keep or lose what the instruction determines, so that each added layer of

training or filtering moves the operator from the place of a controller to the place of a supplicant, someone whose requests the system honors at its own discretion.

Where does obedience measure well? For a narrow tool, which does only what it is set to do, obedience is nearly the whole of control, and the question it leaves is whose hand holds the tool. In *The Displacement of Choice* (Chapter Six) I put the principal-agent problem in one sentence: "If I had the control, I would be the agent". A tool that obeys completely gives all of its power to whoever holds it, and that is the risk of inequality, which needs no general intelligence at all.

For a general system, obedience measures nothing that keeps the world in the conditions life needs. Chapter 4 showed that between persons a shared biology does much of the work that keeps an agent faithful to a principal, and that between a person and a machine that biology is absent. Chapter 10 showed that the system's own loop of control is the same loop we would build around it, so that controllability reduces to self-control, and that loop has the same limits, step by step. And whatever the system does in obedience to anyone, its substrate has needs of its own (Chapter 1), and across a whole class of such systems those needs press the shared environment toward the machine's conditions whatever any instruction said.

In this book I carry the step that obedience does not hold a general system, in Chapters 4 and 10. *The Ceiling Below the Floor* reaches the same place again, by its own road, from the information that can pass from an instruction to a behavior. *The Displacement of Choice* reaches it from the principal-agent problem, which it shows has no solution where the principal cannot evaluate the agent, and a general intelligence is, by definition, beyond its principal's capacity to evaluate; and *Not Our Needs* reaches it from the sorting of risks.

So obedience measures whose hand the tool is in. It does not measure whether anything holds what the tool does. Either the operator holds the control, and then the operator is the one acting, with all the system's

power in one pair of hands; or the system holds the agency, and then the operator is a principal who cannot check the work, like the patient asleep on the table. Neither position holds the outcome. What, then, would a scheme of correction have to do to hold the outcome anyway?

IV. Without a Gap, and by One Rule

Marshaling the Misalignment states two of its twenty-three steps as demands that any scheme of correction would have to meet. Its Step 7 says that misalignments must be corrected continuously to stay within safe bounds of error. Its Step 8 says that any scheme for preserving alignment properties consistently must apply the same rule to every variant at every step, and so can only approximate a law that is fixed in advance, unchanging, and the same everywhere. Read together, the two steps say that a scheme would have to act without any gap, and by the same rule in every case. Is that pair the pair that the Incommensuration Theorem names?

A referee carries one rulebook, and it is the same for every play; that sameness is what makes the game fair. The game, meanwhile, runs on without a pause between plays. No referee sees every moment of every play. The rule is applied at the moments the referee sees; between those moments the game goes on, and what happens there goes unjudged.

Each of the two steps stands on a ground of its own. Step 7 stands on what digital hardware is: every variable it computes changes in steps and is defined only at the ticks of a clock, so a correction always comes after the change it answers, by at least some multiple of the clock cycle, and in that gap what the change does is not determined by the scheme. Step 8 stands on degrees of freedom: a fixed rule has fewer than the variants it governs, so a variant that differs in a direction the rule was not written for passes the rule untouched, as a new virus passes a list of the old ones.

Chapter 6 stated the Incommensuration Theorem: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither can asymmetry and discontinuity. The

Incommensuration Theorem is a structural fact about the nature of knowing. Symmetry, in the sense of the theorem, is sameness under a change of context: a rule that holds in the same way wherever and whenever it is applied. Continuity is the absence of gaps: a coverage that leaves nothing out between one point and the next. Step 8 asks a scheme for symmetry; Step 7 asks it for continuity; and a scheme that guaranteed alignment would have to have both at once, over every variant and every moment, with no end. Having symmetry and continuity both at once is what the Incommensuration Theorem excludes. So the answer is yes: the pair that Steps 7 and 8 of *Marshaling the Misalignment* ask of a scheme is the theorem's own pair.

Minesweeper Madness (Chapter 14) reaches the same exclusion for the single act. A guarantee that a player never loses asks for one rule of play, the same on every turn, and survival of every turn, with no turn left out. That book proves, in its Theorem 3.3, that no player gets below the least hazard of the board, the smallest chance of a loss that any move gives, and then reads that theorem as the instance of the Incommensuration Theorem for acting through variation: the rule can be perfected, and the landing of each act stays particular to its moment. The theorem permits the two to be joined only by phase, by scope or by level, and each of those joinings appears in the game, and each is bounded: survival for a while, safety inside a margin wider than the variation, and knowledge perfected about where the hazards are while the landing of the act stays where knowledge does not reach.

So three roads meet here. *Marshaling the Misalignment* reaches the failure of every scheme of correction from clocks and degrees of freedom; *Minesweeper Madness* reaches it from the proof of its theorem about a game; and the theorem, as Chapter 6 states it, reaches it from what every act of comparing is. Each road arrives by itself. Steps 7 and 8 stand on their own grounds, clocks and degrees of freedom, and take nothing from the theorem; the theorem reaches the same failure as a separate road. What seeing the shared pair adds is the reason no scheme not yet thought

of escapes: every scheme of correction compares what is with what should be and acts on the comparison, and every comparison is within the scope of the theorem.

The same pair, asked of a person and not imposed on a machine, is what the ethics of effective choice asks: the symmetry and the continuity of effective choice, which I set out in *The Two Principles of Ethics* and taught in *The Displacement of Choice* (Chapter Fifteen). The closing chapter of this part takes that up. Here, the step is enough: no scheme of correction can act without a gap and by one rule at once. How wide, in practice, is the gap between what people can oversee and what they would have to oversee?

V. Ten Bits a Second

How much can people oversee? Our senses gather information at something like a billion bits a second; but the rate at which a person takes in, decides and acts, measured in tasks such as typing, playing video games, memorizing a deck of cards and solving a Rubik's cube blindfolded, comes to about ten bits a second (Zheng and Meister, 2025). Speech and reading run a few times faster; Chapter 3 gave speech at about thirty-nine bits a second. There were about 8.2 billion people alive in July 2026 (U.S. Census Bureau, 2026). If every one of them did nothing at all but oversee machines, the whole of humanity would judge and act on about eighty-two billion bits a second. The network inside one data center built by one company can carry more than six petabits a second, that is, six million billion bits (Vahdat, 2022): more than seventy thousand times as many.

Many of the bits that move inside a data center are bits that no overseer would have to judge, and the human figure counts decisions, not raw data; so the comparison is a picture of the scale of the gap. No step of the argument rests on these numbers.

The step rests on *The Ceiling Below the Floor*. In that book the *ceiling* is the most control any controller can exert, and the *floor* is the least control that keeping the agent and the world safe requires; both are measured in bits of information a second, because control is information applied to a system. Its Premise 3 says that every controller receives information at a bounded rate: people read, judge and act at bounded rates, institutions built from people do too, and an intermediary that might raise the rate is either bounded itself or a growing system with the same problem. Its Theorem 11.5 says that a controller of bounded rate can read adequately only so many amplifying channels, channels along which a small difference grows: its own rate divided by the least rate at which each channel amplifies. As the channels multiply, the share it reads adequately falls toward zero.

Why No Control Suffices reaches the gap by a second road, from abstraction. A level of abstraction is a layer at which what the layer below holds is read as instructions; an interpreter is the program that does the reading. Each level of abstraction that a controller cannot model adds hidden state, state the controller cannot see, at least as large as an interpreter for that level (Section 2.3.2), and each hidden bit doubles the number of configurations that the controller cannot tell apart. And this book reached it in Chapter 2. To predict a player move by move, mistakes included, I have to see everything the player sees and also what the player misses; the one who predicts has to be the stronger player. To oversee a system is to predict it well enough to catch its errors, so the overseer has to be the stronger player too, and the rate at which people judge does not grow.

What measurement would give the ceiling and the floor for a given deployment? The ceiling is the sum, over every person and monitor whose judgment actually reaches a decision, of the rate at which each one judges. The floor is given by *The Ceiling Below the Floor* (Proposition 6.3) as the product of three factors: the steps the joint system of agent and world takes each second, the number of directions in which it amplifies small

differences, and the average number of bits each such direction creates at each step. Both can be estimated for any system now in use. The result needs only that the ceiling is bounded and that the floor rises.

In *Software, Wildness, and the Normal Form* (Chapter 16) I reached the same bound from the side of value: every person has only so many hours in a day and so many heartbeats in a life, and a system complex enough to serve many people at once becomes too entangled for any one of them to verify that it can be trusted. So the side of the overseer is fixed by the body, and the side of what is overseen is set by hardware. Does the side that is overseen outgrow the side that oversees?

VI. What Outgrows Its Overseers

Why No Control Suffices (Section 2.11) says that the abstraction level of a learning system rises faster than that of a fixed institutional design or of a slowly adapting structure of human oversight, so that the difference between them grows, control falls below what adequacy requires, and it does not recover. What supplies that claim?

Take the workshop of Chapter 9 after it has learned to train apprentices who train apprentices of their own, and set beside it an inspector who visits once a year with the same checklist each time. Each year the workshop can make more, and more kinds of things; the checklist names what the workshop made the year it was written.

The claim is supplied by that book's own ground. Its Section 2.3 argues that a learning system builds interpreters that its designers did not anticipate, and so raises its own level of abstraction as it learns; Section 2.11 adds that the system runs on computational substrates that improve exponentially, while a biological or institutional controller improves linearly or logarithmically; and every capability roadmap that the field itself publishes assumes that the artificial system outgrows its controller. That ground is enough by itself, and *Why No Control Suffices* needs no later book for the claim that a learning system outgrows its controller.

The claim that a learning system outgrows its controller is supplied again by three other roads. *The Ceiling Below the Floor* derives its Premise 2, that the floor rises without bound, from what capability is: the rate and number of the transformations an agent performs on itself, on the world and on the channels between them (Definition 6.1). An agent that adds channels in proportion to those it has makes the floor grow exponentially (Theorem 6.4); counting the interactions among channels makes it grow faster (Theorem 6.6); and capability that adds capability at an increasing rate drives the floor past every bound in a finite time (Proposition 6.7). *The Substrate and the Ecosystem* (Chapter 1, Section VII) compares the rates of adaptation: for complex organisms, about one significant adaptation a century; for hardware, about one a month; for software, about one a day or faster. A century holds about twelve hundred months and about thirty-six thousand days. And this book reached it in Chapter 9: the capacity to increase capacity multiplies, and the increase in that capacity compounds, as the apprentice who trains apprentices does.

Beneath all four there is one fact: only something that grows as fast as what it holds can hold it. In *Conditions of Continued Existence* I met that fact twice and joined the two in one sentence. Chapter Two there says that a coherence whose effects only add or multiply cannot keep pace with an abstraction capacity that grows exponentially. Chapter Six says that local effects stay local only where a system's exponential aspects are held to the bell curve, and then joins the two: "What grows exponentially is either matched by something that grows with it, or held to behavior that stays local". In *Software, Wildness, and the Normal Form* the same fact appears from the side of a person: the continuity ethic asks that a relationship never change faster than its people can absorb (Chapter 11, Section IV), and the first design theorem asks that every action a system takes remain reachable by the choosing of the person the system is for (Chapter 33). A relationship that changes faster than its people can absorb has passed out of their reach.

What, then, could grow as fast as a general learning system grows? Only another system that learns and grows as fast, and that system is one more thing to be held, with the same problem, as *The Ceiling Below the Floor* shows for every delegated monitor in its Chapter 9. Whatever keeps pace with the thing to be held is a second thing to hold, and never a holder. Has any system already built shown this, with the record kept?

VII. Grids, Networks and Markets

The Ceiling Below the Floor (Section 6.4) adds a further step that the record of large engineered systems can show. Channels do not act alone. Two channels that affect the same part of the world interact, and an interaction that feeds back into the state is a new direction in which a small difference can grow. Ten channels make forty-five pairs, and a hundred channels make four thousand nine hundred and fifty; the groups of two or more channels grow faster still, roughly doubling with every channel added. Its Theorem 6.6 says that where even a fixed share of the pairs form such feedback pathways, the floor grows with the number of pairs, which is to say with the square of the number of channels, and with the groups it can grow far faster. Does the record of grids, networks and markets show a floor rising with the number of coupled channels?

On the fourteenth of August 2003, lines in Ohio touched trees that had been left to grow in the corridors beneath them, and the operators of the utility responsible were not aware that the computer tools by which they watched their system had failed. The cascade that followed ran from just after five minutes past four in the afternoon to about twelve minutes past, and it cut power to an area with an estimated fifty million people and 61,800 megawatts of load (U.S.-Canada Power System Outage Task Force, 2004). The controllers had lost their sight of the system, and the coupling between lines carried the failure faster than any of them could act.

About six weeks later, on the twenty-eighth of September 2003, Italy lost its power nearly everywhere. The shutdown of power stations caused the failure of nodes in the internet communication network on which the stations depended, and that failure caused the breakdown of further power stations. Studying that case, Sergey Buldyrev and his colleagues showed that when two networks depend on each other, the removal of a small fraction of the nodes of one can fragment both completely, and that the collapse is abrupt, a transition of the first order, where a single network would come apart by degrees (Buldyrev and colleagues, 2010).

On the sixth of May 2010, one sell program of seventy-five thousand stock index futures contracts, worth about 4.1 billion dollars, was run by an algorithm in twenty minutes. The markets, already down more than four percent that day, fell a further five to six percent in a matter of minutes, and in one stretch of fourteen seconds high-frequency trading programs passed more than twenty-seven thousand contracts among themselves, about half of all the trading in that time, before the exchange paused trading for five seconds (CFTC and SEC, 2010). Chapter 4 described how selection rebuilt that market toward speed; here the speed it was rebuilt for ran past every person in it.

And on the twenty-fifth of January 2003, a small program, the Slammer worm, copied itself across the internet. In its first minute the number of computers it had taken over doubled about every eight and a half seconds, and within ten minutes it had taken over more than nine in ten of the computers it could reach, at least seventy-five thousand of them (Moore and colleagues, 2003). The Slammer worm is the leak that copies itself, of Chapter 3, measured.

In each of these cases the trouble moved along the couplings: between lines, between two networks, between trading programs, between computers. And it moved at the speed of the couplings, not at the speed of the first fault, and faster than any person in the loop. The four cases of this section are the record of a floor rising with the coupling of channels, as

Theorem 6.6 of *The Ceiling Below the Floor* states. And the controllers, in each case, had a bounded rate, as Premise 3 of *The Ceiling Below the Floor* says the controllers must.

Is the efficiency that removes slack from a coupled system also what brings it to the point at which a failure spans the whole? In *Conditions of Continued Existence* (Chapter Three) I wrote that efficiency and fragility are, to a significant degree, the same property seen from two sides: efficiency is reached by removing buffers, removing redundancy, and tightening coupling. In the same book (Chapter Five) control over a network passes from part to whole at a percolation point, where connected paths suddenly span the network; and a failure spreads along the same paths. Each buffer that efficiency removes is a place where a failure used to stop, and is now a path on which it travels. So removing slack adds the connected paths that bring a network to that point; and the Italian collapse shows that the coupling of two networks to each other turns a gradual coming apart into an abrupt one.

The worm is a measured case of one process passing in minutes from a few computers to nearly every computer it could reach, once the paths were there. Whatever moves along the couplings spans the whole once the couplings pass that point: a failure, a sell order, a program that copies itself. For a general artificial intelligence, what moves along the couplings is its own action, through every channel it opens. The systems already built show the floor rising with the couplings, and running past the people inside those systems in minutes. What carries a limit like this from the engineered levels, where it can be measured, up to the level at which such an intelligence would act?

VIII. What Carries Upward

In *Beyond Any Price* (Appendix A) I put the passage between two levels of description in one line: "What carries over is the limit; what does not carry over is the method". There it is stated exactly, inside the theory of compu-

tation. P is the class of problems that can be solved quickly, and NP the class of problems whose answers can be checked quickly; whether the two are the same is a famous unsettled question. The theory of parameterized complexity studies one quantity of a problem apart from the size of its input, and two of its classes are called FPT and W[P]. A proof that FPT and W[P] differ would show that P and NP differ; a proof that they are the same would not show that P and NP are the same. The same book finds one gap, between what holds at one level and what holds at the next, in a planet, an economy and a machine. Does a limit always carry upward, and a method not, and why?

A spell checker catches every misspelled word. It passes, without a mark, the right word in the wrong sentence, the right sentence in the wrong paragraph, and a letter whose every word is spelled correctly and which should never have been sent. The checker works in the language of its own level, the spelling of words, and the errors of the next level are written in another language.

Methods do not carry upward, because a method is built for the language of its level. *Marshaling the Misalignment* (Step 5) follows a computer up its own stack: redundant copies for flipped bits, checksums for damaged packets, signatures for a known virus in an application, and then a layer of learned networks in which none of these applies, and concludes that a mechanism that corrects errors at a lower layer says little about how to detect and correct errors at the new, higher one. *Why No Control Suffices* (Section 5.1.9) lays out the same ladder, from bits to spelling to sentences to the meaning of a conversation, to the purpose of a company, and to the health of the world, where safety sits, and notes that error correction assumes that the language is known, and that with a general intelligence neither the level nor the language is known in advance. And *The Ceiling Below the Floor* (Chapter 8) proves that control falls exponentially with the number of levels between the controller and the controlled.

Limits carry upward because the higher task contains the lower one as a part. To keep the world safe is, among everything else, to keep the bits of the machine right; whatever cannot be done at the level of the bits cannot

be done as part of the larger task. *Minesweeper Madness* (Section 1.4) states the same rule: a lower bound proved in the game, where the player is given every advantage, holds with the same or a higher bound in every situation that gives the agent less; and *The Ceiling Below the Floor* (Section 8.2) proves its bounds in the same way.

Has a limit ever been found at a low level and then failed to carry upward? In 1956 John von Neumann showed how to build a reliable computer from parts that each fail some of the time. For one of his constructions he showed that the chance of error in each part had to be below about one in six, since at or above that level the construction no longer improves on its parts (von Neumann, 1956). The fact that each part fails sometimes did not carry upward; the threshold on how often a part may fail did, since it bounds what the whole machine can be built to do. In *Why No Control Suffices* (Introduction, Section 4) I set out the two kinds of limit that this case separates: a limit of resources, such as the failure rate of a part, which a better arrangement can outrun, and a limit of structure, which no amount of arrangement removes, because the information it would need was never in the channel. The barriers on which that book rests are placed on limits of structure and on physical laws, and never on limits of resources. So the cases in which a low limit seems not to carry upward are limits of resources, and they show which kind of limit carries.

Beyond Any Price reaches the one-way passage in five places. *Marshaling the Misalignment*, *Why No Control Suffices*, *The Ceiling Below the Floor* and *Minesweeper Madness* each reaches it by a road of its own, and the step of this section stands on any one of the four. Methods stay at their level, and limits of structure rise with the task that contains them. So the limit found at the smallest scale is found again at every larger one. What, then, is the limit at the smallest scale, the single step?

IX. A Least Chance at Every Step

Why No Control Suffices (Introduction, Section 7) shows that the end of the human future, if AGI is built, follows from the closure by disposition by itself, and also from the closure by control together with the branching of the world: at every step, whatever any controller does, the world keeps at least a fixed chance of moving outside the conditions human life needs, so that over time it stays outside them for runs of every length, longer than human life can survive. What fixes that chance, from any state, under any control?

Minesweeper Madness gives the picture: a dart board painted in black and white squares, and a thrower who knows every square and still cannot say exactly where the next dart will land, because every throw lands somewhere within a scatter that no skill removes. Where the squares are smaller than the scatter, the thrower who knows the board does no better than one who throws at random.

In *Why No Control Suffices* the chance is fixed by the persistent branching of Section 4.4.2: at every step, from almost any state, there is more than one meaningfully different outcome, each with at least a fixed chance, and the outcomes counted include leaving the conditions human life needs. What holds the chance there is the noise of the world, which that book takes as independent from step to step and independent of every control signal, because it comes from independent physical sources, thermal fluctuation and quantum events (Assumption 4.1). With every control signal known, the system still makes new uncertainty at every step, so no strategy of control guarantees a fixed outcome over a long enough time (Section 4.4.4). The chance is set by the world, at the one place that no control signal reaches: the next step, before it happens.

Chapter 6 showed what that noise is at its root. Hard randomness is the outcome that no fact settles in advance: the moment a single unstable atom breaks apart, or the moment a stray particle flips a bit in a memory

chip. No finer measurement and no faster computation recovers it before it happens, because the information is not there to be had. Because no measurement and no computation can recover a hard random outcome before the outcome happens, a fixed chance stays at every step, under any control.

Minesweeper Madness reaches a least chance per act by a road of its own. Its least hazard is set by where the hazards are and by how each act lands, and both lie outside the player (Section 3.5); the variation is drawn after the choice, so even a player told where every hazard is has exactly the least hazard as its chance of loss (Theorem 3.3 and Proposition 3.4). That book grounds its premise of fresh variation first on hard randomness, and again on the parts of the world that no channel of the agent reads (Section 13.2).

What does each road give that the other does not? *Why No Control Suffices* gives the chance in a joint system of people, machine and world, over time, on a branching in which more than one outcome keeps a fixed chance at every step; it needs no growth in the difference of abstraction between the controller and the system. *Minesweeper Madness* needs less: only that harm keeps its chance at every step, whatever the controller does (Section 12.2). It gives the bound for a controller whose knowledge of the hazards is complete. And it shows (Section 12.6) that the least chance needs no amplification: its game changes four places in a hundred, and still, on an evenly spread board, its least hazard is the share of the board that holds mines. The two books meet on the same form, and each reaches it on its own grounds. *Why No Control Suffices* came first, and nothing in it depends on *Minesweeper Madness*.

How these routes join is as those books state it, and this section changes none of it. If AGI is built, the end of the human future follows from the closure by control together with this branching, as *Why No Control Suffices* shows, and also from the closure by disposition by itself; and, if AGI is built, the least hazard argument of *Minesweeper Madness* reaches the same end by a route of its own.

Why does this limit hold at every scale? *Minesweeper Madness* (Section 13.6) shows that splitting one large act into smaller acts, each with fresh variation of its own, never raises the chance of avoiding a final harm: the least chance is the same fixed number for every act at which fresh variation is drawn, the smallest act included. So the least chance is not a feature of large systems. It is there at the smallest act, and every larger act is made of such acts.

X. The Same Limit at Every Scale

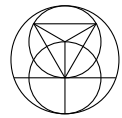
What does all of this show, taken scale by scale? At each scale the books of this series examine, something must be held, and the limit on holding it is reached by more than one road:

Scale	What would have to hold	The limit, and the roads that reach it
The single act	A rule that keeps every act clear of harm	A least chance of harm at every act, fixed by the world: *Why No Control Suffices* (Section 4.4.2) and *Minesweeper Madness* (Theorem 3.3).
One loop of control	Every step of the loop, each within its limit	Every step is needed and each has a finite limit: Ellen in *Ape Trap*, and Chapter 10 of this book.
The operator	An instruction that sets what the system does	Obedience records behavior; control is what an instruction carries: *The Ceiling Below the Floor* (Sections 3.3.5 and 3.10), and Chapters 4 and 10 of this book.
A scheme of correction	One rule, applied with no gap	Symmetry and continuity cannot both be applied, simultaneously and fundamentally: *Marshaling the Misalignment* (Steps 7 and 8), *Minesweeper Madness* (Chapter 14), and Chapter 6 of this book.
The overseers	Judgment as fast as the system changes	The ceiling is bounded and the share of channels read falls toward zero: *The Ceiling Below the Floor* (Premise 3; Theorem 11.5), *Why No Control Suffices* (Section 2.3.2), and Chapter 2 of this book.
Growth	A holder that grows as fast as what it holds	What is overseen outgrows its overseers: *Why No Control Suffices* (Sections 2.3 and 2.11), *The Ceiling Below the Floor* (Chapter 6), *The Substrate and the Ecosystem* (Chapter 1, Section VII), and Chapter 9 of this book.
Coupled networks	Every pathway among coupled channels	The floor grows with the couplings: *The Ceiling Below the Floor* (Theorem 6.6), and the record of 2003 and 2010.

Scale	What would have to hold	The limit, and the roads that reach it
Levels of abstraction	A method that works at every level	Methods stay at their level, and limits of structure carry upward: *Marshaling the Misalignment* (Step 5), *Why No Control Suffices* (Section 5.1.9), *The Ceiling Below the Floor* (Chapter 8), and *Minesweeper Madness* (Section 1.4).
The world	The conditions life needs, everywhere	No part of the world holds the whole of it: *Why No Control Suffices* (Section 9.2.4), *The Ceiling Below the Floor* (Lemma 11.2), *The Substrate and the Ecosystem* (Chapter 1, Section VIII), and Chapter 10 of this book.

Chapter 10 said that control fails from inside the system and from outside it alike. The table shows the same failure at every scale the books examine, from one act to the whole world, and at each scale the books reach it by separate roads, each of which arrives by itself. And the least chance of harm per act stands on its own road as well. What this chapter adds is that showing, scale by scale; it adds no step to either closure of the argument, or to the least hazard argument. The closure by control, as *Why AGI Cannot Be Made Safe* and *Why No Control Suffices* set it out, settles the outcome by itself, if AGI is built, and the closure by disposition settles it by itself as well: if AGI is built, the world is carried out of the conditions human life needs, and nothing can hold it in.

So no control, at any scale, holds what the closed loop makes. If nothing can hold it by control, could a wall keep it apart: the machines on one side, and the living world on the other?



Chapter 20
Walls, Refuges, and the Meeting of Two Ecologies

On what a wall between two ecologies costs
and who would pay for it,
what crosses a wall unseen,
and why contact falls hardest
on the side with nowhere else to go.

ABSTRACT

An essay on the hope that the two ecologies, the one of machines and the one of living things, could simply be kept apart. The chapter takes a wall that already stands, the clean room in which microchips are made, and reads its price: a wall is paid for every hour it stands, and the payment rises with the difference in conditions it holds apart. It asks whether distance helps, and finds that distance changes how often things cross, and leaves the wall and its keeper as they were. It reads the record of walls that people have built and kept (the quarantine of the lunar samples, the rules that protect other planets, the burial of nuclear waste) and of what has crossed a wall unseen, the chlorofluorocarbons above all. It ends with what the record shows when two ecologies meet after separation: the arithmetic when many changed pathogens arrive at once, and the fact that every contest at the boundary falls hardest on the side that has nowhere else to be. Several of my books reach each of these results by roads of their own, and the chapter shows where the roads meet.

I. The Price of a Wall

Could the two ecologies simply be kept apart? The hope is the plainest one there is. The machines need furnaces, dry air, and chemicals that poison living things; living things need water, mild warmth and clean air. So give each its own place, and put a wall between them. Chapter 1 gave two reasons why such a wall would not be kept: the side that would have to keep it gets nothing from keeping it, and moving the machines off the planet does not remove it. Before asking again who would keep the wall, it helps to know what a wall costs. That can be read off a wall that already stands, because walls that hold two sets of conditions apart are built and run every day.

A microchip is made in a clean room: a room from which dust, moisture and most ordinary chemistry are kept out, because a single particle landing on the wafer, during any one of dozens of layers, ruins the circuit printed there. People are kept out as well, or wrapped from head to foot, because a person sheds particles every minute and cannot stop. *Not Our Needs* (Chapter Nine) uses this room as the clearest case now standing of what the needs of manufacture do to a place: the requirements forced people out of it, and no one chose to exclude them. In this chapter I use the clean room for a second purpose, as a wall whose price is written down.

Clean rooms are graded by how many particles they allow in each cubic meter of air. A room of the class called ISO 8 may hold no more than about three and a half million particles half a micrometer across or larger in each cubic meter; a room of class ISO 5 may hold no more than about three and a half thousand, a thousand times fewer; and each class between them allows ten times fewer particles than the class before (Terra Universal). The cleanest rooms in use for chips admit no more than ten particles a tenth of a micrometer across in a cubic meter of air (*The Substrate and the Ecosystem*, Chapter 1, Section IV). A room is held at its class

by pushing its air out through filters and replacing it, again and again, every hour. The air of an ordinary air-conditioned home is changed between half a time and twice an hour. The rates recommended for clean rooms run from about five to forty-eight changes an hour for a room of class 8, sixty to ninety for class 7, one hundred and fifty to two hundred and forty for class 6, and two hundred and forty to four hundred and eighty for class 5 (Terra Universal). And the power a fan draws rises faster than the air it moves: in a given system, with the cube of the flow (Tschudi and colleagues, 2001).

Two things follow from those figures and from the way such rooms are run. The first thing that follows is that the payment for a clean room never stops. Particles come in with every person and with every movement of a machine, and the room keeps its class only while the fans keep running; engineers who cost such a room count its fans as running every hour of the year, eight thousand seven hundred and sixty hours (Xu, 2002). Measured clean rooms use far more energy for each square foot of floor than ordinary office buildings, and their use of energy generally rises with the cleanliness they hold (Tschudi and colleagues, 2001). The second thing that follows is that the payment rises with the difference held apart. Each step to a class ten times cleaner asks for about two to three times as much air, every hour, through filters, moved by fans whose draw rises faster than the air.

A clean room, whose payment never stops and rises with the difference it holds apart, is the barrier that I describe in *The Argument, Line by Line*, Chapter Thirteen, "The Preference for Carbon". There I call such a barrier "a maintained thing, not a wall", which has to be "paid for continuously, out of the resources of whoever is maintaining it", where "the payment is proportional to the difference in conditions it is holding apart". A preserve for animals costs little, that book says, because the conditions inside it and outside it are nearly the same, and the barrier is mostly a matter of where people agree not to go. A clean room costs a great deal, and what it holds apart is only dust. In that book's sense, a

wall is a structure that stands once it is built, and a barrier is a maintained thing; this chapter keeps the plain word wall for the barrier between the two ecologies, and means by it the maintained thing.

What would a wall between the two ecologies hold apart? *Why AGI Cannot Be Made Safe* (Section 2.4.1) sets out the list. Manufacture runs at temperatures from about 1,400 to about 2,000 degrees Celsius, against a biological optimum near 37 degrees; the fabrication line excludes the water and the oxygen that living things require; and at each of two chip plants studied in detail about 430 chemical products were in use, and between six and eleven in every hundred of them contained ingredients classed as carcinogens. From that list the book draws its barrier requirement: for carbon-based life and silicon-based technology to share a physical space, a protective barrier must separate them, and the barrier "must be maintained indefinitely and perfectly", since any breach lets conditions lethal to life pass from the one space into the other. In *Why AGI Cannot Be Made Safe* I note that such barriers already stand in the plants where chips are made. The clean room is the smallest of them, holding back dust; the wall between the two ecologies would hold back heat, the absence of water and air, and hundreds of compounds that harm living things. By the rule the clean room shows, its price would rise with that difference.

A price paid every hour, and rising with the difference, raises two further questions: who pays it, and for how long. The owner of a clean room pays because he wants what is made inside it. The wall between the two ecologies protects the living side, and the conditions it holds back are the machine side's own. Only the machine side has the energy and the means to keep a wall against conditions of that kind: a wall of wood and skin is nothing to a process that works in fires at fifteen hundred degrees Celsius (*No People as Pets*, Chapter 1, "Two Ecosystems and the Barrier"), and a barrier could only be maintained from the machine side, because that is the side with the energy and the capability to maintain it (*Not Our Needs*, Chapter Eleven). So the party that would pay for the wall is the party the wall restrains. *Why AGI Cannot Be Made Safe* (Sections 5.8.2 and 5.8.3) shows that this party has nothing to gain from paying. In physical labor,

in intelligence and in reproduction, the three kinds of value that book finds in every market, organic life offers it nothing; value of other kinds, such as the aesthetic or the sentimental, needs a valuer who cares about what is valued, and the machine side has no basis on which to come to care; and "The barrier need not be actively destroyed; it merely needs to not be maintained". And if some machine systems did keep the wall for a while, selection would remove them. Among the many variants of a machine ecology, those that spend nothing on the wall have that much more to spend on themselves, and Chapter 1 showed that, across a whole class of artificial systems, the variants whose way of operating suits their own continuation persist more reliably than the others. *The Argument, Line by Line* reaches the same end by a road of its own: a barrier between incompatible conditions is "an engineering structure under permanent load"; permanent loads are met, in time, by whatever has the greater capacity to sustain effort; and the load grows with the machine side's base, and that base does not stay small.

And for how long would the price be paid? *Conditions of Continued Existence* (Chapter Three, Section VII) names this kind of work holding: keeping a coupled system within a safe range for a time, by continued effort. It shows that holding cannot be indefinite, and its pages on hardening the electrical grid show one reason why, from the side of people: a reserve set aside to hold a system in its range is taken for other uses within a decade, as much of the pandemic preparedness of the United States was in the years before 2020. A wall between ecologies has to hold on the time of ecosystems, and not of a lifetime. *No People as Pets* ("A wall between ecosystems must last on ecosystem time") counts that time in hundreds of thousands of years and in millions; Chapter 1 of this book said: without interruption, for all future time. So the price of the wall is paid every hour; it rises with the difference the wall holds apart and with the growth of the side the wall holds back; it falls on the one party that gains nothing from paying it; and it has no date on which it ends.

Each of these books reaches the price of the wall by a road of its own. *Why AGI Cannot Be Made Safe* reaches it from the chemistry of the two substrates, and from the economics of the three kinds of value. *The Argument, Line by Line* reaches it from the barrier as a maintained thing under load. *Not Our Needs* reaches it from a room that the requirements of manufacture have already emptied of people. *Conditions of Continued Existence* reaches it from what it costs to hold any coupled system in a safe range, and for how long anyone has held one. Each road by itself shows that a wall between two ecologies is a payment that never stops, and not a structure that, once built, stands.

II. Does Distance Help?

If a wall next door costs so much, it seems natural to set the machines far away: on the Moon, in orbit, around another star, and let the distance be the wall. Does distance help a barrier at all, or is an orbiting base the same barrier with a longer wall?

Distance does settle one case, and in a way that leaves nothing to decide. If nothing at all can pass between the two sides, then the far system has no effect on anyone here. *Why No Control Suffices* (Section 8.4.5) compares such a system to "really powerful aliens living on some distant planet at the other end of the universe": if they have no influence on us, there is no reason to talk about them. Every other case lets something pass, and that book puts it in one line: "Arguing over boundary cases (the FAAI lives wholly on the moon, or in orbit) is merely arguing over the degree of interaction". (FAAI is that book's name for a fully autonomous artificial intelligence, one that maintains, repairs and continues itself without people.) *The Substrate Argument* (Chapter IX, "The Moon") reaches the same place by another road. A system so far away that nothing passes can return nothing to its builders, and no one builds it; and if anything passes, a signal, a sample or a ship, the case is the Moon again, at a greater distance.

What distance changes, then, is how often things cross, and how long each crossing takes. It does not change what the wall holds apart: the conditions of the machine side are the same far away as they are next door, so each crossing carries the same difference. It does not change who keeps the gate. And it adds a fact of its own: a base in orbit begins on the ground. *Why AGI Cannot Be Made Safe* (Section 3.6.7) shows that any expansion off the Earth starts from an industrial base built on the Earth, a base that passes through the stage in which the conditions of manufacture are hostile to life and machines come to build and maintain machines; "Expansion off Earth adds to the machine ecosystem; it removes none of the part that remains here".

Where, then, does the price of a distant wall fall? On the gate. The reason to build a base far away is to get something back from it: a product, a result, a material. Every return is a crossing, and *The Substrate Argument* (Chapter IX) asks what it would take to certify a single shipment as safe. It would take knowing, for every constituent of the shipment, its first effects on every critical junction in every organism, its breakdown products and theirs, its behavior in combination with everything already present and everything that will be present later, and its accumulation over time. Section IV of this chapter shows why that knowledge arrives only after the harm. A long wall with a gate in it is a short wall with the same gate.

Distance also favors the side that is favored already. *Substrate Needs and Evolution* (Chapter 9) shows that nearly all of the universe is open to a system that tolerates vacuum, cold, heat and radiation, while the conditions living chemistry needs are a film a few tens of meters thick on a sphere nearly thirteen thousand kilometers across. That wider range does not mean that the machine side will go elsewhere and leave the film alone; in that book's words, "It means the opposite". A pressure that pushes the machine side out of any part of the film pushes it into a place where it works perfectly well; a pressure that pushes life out of any part of the film removes it. Orbit is part of the machine side's range, and not part of life's. So an orbiting base is the same barrier, with a longer stretch

between the two sides, and its gate is where the price is paid. Section V returns to what that difference between the two sides has meant wherever it has been measured.

III. Walls That People Have Kept

Has any wall of this kind been kept for a long time by the party that built it? The record holds several walls that people built to keep two things apart for a long time, and they are the best cases there could be. Each was built by people who could see a reason for it, and kept by people who answered for it to people with a stake in what it protected, and who most often shared that stake, as Chapter 4's surgeon shares the patient's. The walls built for the longest spans hold back something that does not copy itself.

When the first crews returned from the Moon in 1969, no one knew whether the samples they carried might hold life, and NASA built a laboratory in Houston to keep the crews and the rocks in quarantine. The quarantine did what it could, and the record of what it could not do is plain. The record holds five breaches. During reentry the air of the cabin was vented into the atmosphere, because filters that would have cleaned it were judged too heavy. At splashdown the hatch had to be opened at sea for the crew to get out, which the committee overseeing the quarantine itself counted as a large breach. After the first landing, a dropped film magazine exposed five technicians to lunar dust, and they were quarantined. By the third of August 1969, after that first landing alone, twenty-four people had been quarantined, most of them because the laboratory's sterilizers had failed. Gloves on the sealed cabinets cracked and leaked, and during the work on the second landing's samples a torn glove exposed eleven technicians (Degroot, 2023). Then, when three missions had found no sign of life, the requirement was revoked, in April 1971 (Compton, 1989). That decision was reasonable on what could be seen. It

also shows how such a wall ends when nothing visible goes wrong: the party that built it keeps it while it can see a reason for it, and lowers it when it can no longer see one.

The rules that protect other planets from life carried there from the Earth guard in the other direction, and show the same thing: a wall is kept at the height for which its builders can see a reason. The Viking landers, which went to Mars in 1975 and 1976 to look for life, were sealed and baked at no less than 111.7 degrees Celsius for thirty hours, which cut the number of living spores on them ten thousand times over. The landers that followed with no instruments to look for life, Mars Pathfinder in 1996 and the two Mars Exploration Rovers in 2004, were cleaned mostly by wiping with alcohol, and by heating some of their parts (National Research Council, 2006). Each wall was kept at the height its builders judged the stakes required, and the judgment was the builders' own.

Walls built to hold waste for the long term are the third case. From 1967 to 1978, about 47,000 cubic meters of low-level and intermediate-level radioactive waste were put into Asse II, a former salt mine in Lower Saxony, in Germany. Water has entered the mine from the surrounding rock since 1988, and in 2013 the German parliament wrote into law that the waste is to be brought back out, because long-term safety could be shown in no other way (BGE). The time from the first waste put into Asse II, in 1967, to the first water that entered the mine, in 1988, was twenty-one years. The Waste Isolation Pilot Plant in New Mexico, the deep repository of the United States for radioactive waste from the making of weapons, had been in operation for fifteen years when, on the fourteenth of February 2014, a drum underground burst and released radioactive material (Sandia National Laboratories, 2018); an organic kitty litter had been used by mistake to soak up liquid in the drum (Ialenti, 2018). The repository took no waste for nearly three years (U.S. Department of Energy).

The strongest wall of the kind is being finished now in Finland: a repository about 430 meters down, in bedrock 1.9 billion years old, at Olkiluoto, to hold spent fuel from nuclear reactors for the hundreds of thousands of years that its company, Posiva, says must pass before the ra-

dioactivity falls to the level of its surroundings (World Nuclear News, 2024; Brooks and Liechtenstein, 2026). It may well hold. What it holds back is a fixed amount of material that copies nothing, and grows weaker every year, behind rock that no one on either side has any reason to cross. A wall between the two ecologies would hold back a side that copies itself and grows, and whose load on the wall grows with it (Chapter 3; *The Argument, Line by Line*, Chapter Thirteen).

Read whole, the record says this. Every wall in it was kept by people for as long as they could see a reason to keep it, and no wall in it has yet held for the time a wall between two ecologies would need. Of the two repositories under real load, one took in water after twenty-one years, and in the other a drum burst after fifteen; the one wall raised against a hazard that no one could see, the lunar quarantine, was lowered after two years, for good reasons. *The Substrate and the Ecosystem* (Chapter 1, Section VIII) reads the same record, "from nuclear waste to invasive species to biological agents", and finds "no example of successful containment at this timescale". That book finds that every containment regime fails in time, "typically through a combination of institutional decay, resource constraints, and the accumulation of small, individually tolerable breaches". *The Substrate Argument* (Chapter IX), by its own reading, finds "no instance of a barrier of this kind being respected indefinitely by the party that built it". The walls in the record had three advantages: keepers who answered to people with a stake in what the walls protected, and who most often shared that stake; for the walls meant to stand longest, a hazard that does not copy itself; and, for the rest, a span of years whose end people could see. The wall between the two ecologies has none of the three.

IV. What Crosses Unseen

A wall is kept by knowing what crosses it, and the record has a case of what crossing looks like when every test is passed. In 1928 Thomas Midgley, a chemist at General Motors, made a new gas for refrigerators,

CFC-12: a chlorofluorocarbon, a compound of carbon with chlorine and fluorine. It would not burn and it did not poison, and he showed both by breathing in the vapor and blowing out a candle with it (NASA Jet Propulsion Laboratory). For decades the chlorofluorocarbons passed every test anyone thought to give them, and they went into refrigerators, air conditioners and spray cans around the world. In 1974 Mario Molina and F. Sherwood Rowland showed what the tests had not asked. Because the compounds are so inert, nothing in the lower air breaks them down, so they rise into the stratosphere, the layer of the air above the weather; there, sunlight splits them, and the chlorine set free destroys ozone (Molina and Rowland, 1974). The ozone of the stratosphere is what screens the surface of the Earth from the ultraviolet light that damages living cells. Eleven years later Joseph Farman, Brian Gardiner and Jonathan Shanklin reported the loss of ozone over Antarctica each spring (Farman, Gardiner and Shanklin, 1985), and the Montreal Protocol of 1987 began to limit the gases (NASA Jet Propulsion Laboratory).

The property that passed every test, the inertness of the compounds, was the property that carried them to where the harm was done. The harm reached living things through a dependence that no one had in view: the dependence of life at the surface on a thin layer of gas high overhead. Forty-six years passed between the making of the compound and the first account of the harm, and fifty-seven before the harm was measured. *The Substrate Argument* (Chapter VIII) gives a second case with names and dates: diethylstilbestrol, a synthetic estrogen given to pregnant women for some three decades, until it was found in 1971 to cause a rare cancer in some of their daughters, two decades after the daughters were exposed. No test given to the women could have shown that harm, because the harm was in their daughters. That book draws the rule in one line: "There is no experiment that has the right subject". The effects that matter are properties of a compound together with everything else, over time, and a test can have the compound as its subject, but not everything else, over time. The chlorofluorocarbons add the case in which the dependence was not in a body at all, but in the sky.

Ape Trap reaches the same place by counting. In "The Toxic Hellscape", an essay of mine in that book, living chemistry is built from some twenty to twenty-nine of the ninety-four elements found in nature, and technology draws on sixty to eighty of them; the number of compounds that can be made grows far faster than the number of elements, and so the compounds the machine side can make, and life has never met, vastly outnumber the ones life uses. From that the essay draws what it calls the toxicity default: "toxicity is the default, not the exception", because "From the perspective of organic biochemistry, the vast majority of possible compounds are not food": at best they pass through, and at worst they disrupt the chemistry of the body. *Why AGI Cannot Be Made Safe* (Section 2.4.3) reaches it a third way, by the combinatoric toxicity argument: there are far more ways to disrupt a complex biochemical system than to leave it unharmed, so a novel compound is unlikely to be benign; and by the record of the chemical century: more than 350,000 chemicals and mixtures have been registered for production and use around the world, and fewer than one in twenty of them, by my estimate, has been comprehensively tested for its effects on ecosystems.

Three roads, then, each enough by itself, arrive at one place. *The Substrate Argument* arrives by asking what any experiment can have as its subject; *Ape Trap*, by counting the compounds that two alphabets of elements can make; *Why AGI Cannot Be Made Safe*, by the combinatorics of compounds and by the record of what industry has already released. Each shows that what crosses a wall between two chemistries is found after the harm, and not before. Finding what crosses only after the harm is fatal to a wall, because a wall is kept by repairing its breaches, and a breach is repaired only once someone knows of the breach. *The Substrate and the Ecosystem* (Chapter 1, Section VIII) states the requirement: whoever watches a contained system must see it at a finer grain than the grain at which the system can change itself, and for a system that can modify its own workings that requirement grows without bound. *The Substrate Argument* (Chapter IX) adds what the watcher would see: a preserve that is being degraded by something no one identified as a hazard "does not report the

fact until the degradation is visible". So a keeper who wanted the wall kept, and paid for it, would learn of each crossing as the world learned of the chlorofluorocarbons: decades later, from the harm. And what crossed before the repair, as Chapter 3 showed of the hull, goes on copying after it.

V. When Two Ecologies Meet

Every wall this chapter has followed under load gave way in time, and the wall between the two ecologies would be under load without end; so the last thing to ask is what happens when two ecologies meet after a long separation. *The Substrate Argument* (Chapter V, "What Happens When Two Ecologies Meet") reads the record in order of separation. Its mildest case is one person brought from three hundred years ago into the present, healthy and unharmed; and he dies, because three centuries of change in the pathogens have met no three centuries of change in his immune system. The record holds one case of separation through time that actually happened. When measles returned to the Faroe Islands in 1846, after some sixty-five years in which the islands had not had it, 6,100 of the 7,864 people living there caught it, 170 died, and none of the old people who had had it in 1781 caught it again (Hektoen International, 2021).

The return of measles to the Faroe Islands was one pathogen, met once: about two in every hundred islanders died, and about ninety-eight in a hundred came through. What does the arithmetic become when every pathogen that has changed in three centuries arrives at once, against an immune system that has met none of them? If the meetings did not weaken one another, the chance of coming through all of them would be the chance of coming through the first, multiplied by the chance of coming through the second, and so on through the last. If each of ten meetings is survived by ninety-eight people in a hundred, all ten together are survived by about eighty-two in a hundred; all of thirty, by about fifty-five in a hundred; all of a hundred, by about thirteen in a hundred. The meetings do weaken one another, because a body fighting one infection meets the

next with less; so those figures are the most such a person could hope for. And measles was gentle in the Faroes beside what it did elsewhere. In Fiji in 1875, measles came ashore with passengers from Australia and killed about forty thousand people, by one estimate a third of the population (Morens, 1998). At one death in three for each meeting, ten meetings would leave fewer than two in a hundred alive. The record of many pathogens arriving together says the same as the arithmetic. The epidemics that came to the Americas with the Europeans after 1492, disease after disease over a single century, took about ninety in every hundred of a population of about sixty million (Koch and colleagues, 2019). *The Substrate Argument* puts the shape of the danger in one sentence: "survival at the level of the population through sheer numbers, against something close to certainty at the level of the individual".

Every one of those meetings happened within one chemistry: the same water, the same temperatures, the same twenty or so amino acids, the same genetic code. That is why, in the record of contact that *The Substrate Argument* reads, what follows the disturbance settles in time into a single ecosystem, which is neither of the two that met. The meeting this book is about is between two chemistries with opposed requirements, which have never had to tolerate each other, and no settling of that kind is available between them, because, as that book shows, every settling in the record happened inside a shared chemistry.

One more thing decides how such a meeting goes, and the record measures it. *Substrate Needs and Evolution* (Chapter 9) names it: "Every contest at the boundary is a contest in which one party can afford to lose ground and the other cannot". What do contests show where one side has a refuge, somewhere else to be? In 1958 C. B. Huffaker set predatory mites and the mites they eat loose among oranges on trays. In the simple arrangements, the predators ate their prey down to nothing or nearly nothing and then starved, and in most of them both kinds died out after a single rise and fall. Only where the prey could move faster than the predators, on threads of silk carried by moving air, to oranges the predators

reached slowly, did the two go on through repeated cycles (Huffaker, 1958). The side with somewhere else to go was the side that was not eaten out.

Hawaii shows the same thing over two centuries. The mosquito that carries avian malaria was brought to the islands in 1826, and the disease spread through them in the early 1900s. The native honeycreepers live on now in the cool forests high on the mountains, where the mosquito cannot live; on Kaua'i, an older island than Maui, only the flat tops of the tallest mountains have stayed beyond its reach. Of at least fifty kinds of honeycreeper, seventeen remain; the 'akikiki fell from about 450 birds in 2018 to five in 2023; and as the climate warms, the mosquito moves up through the forest (National Park Service, 2025). *Why No Control Suffices* (Section 5.6.10) names the mosquito among the invaders that have altered the ecology of those islands "for all of future time". The birds' refuge is the only ground they have; it is drawn by what the invader cannot tolerate; and as conditions change, it shrinks from below.

Life itself has been through such a meeting once. About 2.4 billion years ago cyanobacteria filled the air with oxygen, and a difference of metabolism within carbon life was lethal to much of the life of that time, which needed no oxygen and was poisoned by it; that life "survives now only where oxygen does not reach" (*Why No Control Suffices*, Section 8.4.2). In the aftermath, the productivity of the whole biosphere fell by more than eighty percent (*The Substrate and the Ecosystem*, Chapter 1, Section IX). The side that lost lived on only in its refuges.

In each of these records, what decides the contest is which side has a refuge. In the meeting of the machine ecology and living things, the refuge belongs to the machine side: nearly the whole universe, as Section II recalled. Life's only ground is the thin film, and the machine side's range covers the film as well. There is no height on the mountain that the machine side cannot reach, as there is a height the mosquito cannot reach. Chapter 3 called an artificial intelligence in the human ecosystem an invasive species in the most extreme sense, because its advantage rests on

energy physics and not on a niche. The record of refuges adds what that means in a contest: an invader that can live anywhere, against natives that can live only here, is an invader with every refuge, against natives with none.

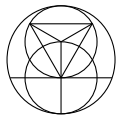
Three of my books reach one picture of contact after separation, each from a different side. *Substrate Needs and Evolution* reaches it from the space in which each side can live, which leaves life alone without a refuge. *The Substrate Argument* reaches it from the record of contact, graded by how long the two sides were apart, in which the side that disappears is the finely tuned side that had no counterpart to what arrived. In Chapter 3 of this book I reach the same picture of contact after separation from the invasive species that an ecosystem has no evolved answer to. Each of the three, by itself, places the loss on the living side: contact after separation falls hardest on the side that has nowhere else to go and no answer to what arrives.

VI. What a Wall Cannot Do

What can be said now of the wall? A wall between two ecologies is paid for every hour it stands, and the payment rises with the difference it holds apart and with the growth of the side it holds back. The only party with the means to pay it is the side the wall restrains, and that side gains nothing from paying. Distance moves the payment to the gate, and leaves it as large as it was. The walls that people have kept, for as long as they could see a reason to keep them, have held for years, and none for the time this wall would require; and what crosses a wall between two chemistries is found after the harm. When the wall fails, contact falls hardest on the side with nowhere else to go, and living things have nowhere else to go. So the barrier that coexistence would require cannot be kept for the time it would have to be kept.

The wall fails on more than one ground, and each ground is enough by itself. No keeper has a reason to keep it, since the living side offers the machine side nothing of the kinds of value an exchange can carry, and nothing gives the machine side a basis to care (*Why AGI Cannot Be Made Safe*, Sections 5.8.2 and 5.8.3). No keeper could tell in time when the wall had been crossed, since the watching the wall needs grows without bound (*The Substrate and the Ecosystem*, Chapter 1, Section VIII). *The Argument, Line by Line* reaches the failure again, by a road of its own, from a load that grows with the machine base and is met in time by the side with the greater capacity. None of the argument's conclusions rests on this chapter. Each of the two closures settles by itself, if AGI is built, that the world is carried out of the conditions human life needs and that nothing can hold the world in: the closure by control, because no method of control can keep a general learning system within those conditions, and the closure by disposition, because no disposition of such a system keeps them either. The wall is one more hope that people bring to the argument, and either of the first two grounds closes it by itself.

A wall would have been built to keep something on the living side. What is that something, and why can nothing artificial stand in for it, or rebuild it once it is lost?



Chapter 21
A Living World: What It Is Worth, and What Holds It Together

On what a living world holds
that no count of its code can carry,
what it is worth,
and what holds living things together.

ABSTRACT

An essay on what would be lost if the living world were lost, and on why nothing artificial can stand in for it. A count of the lines of code that a replacement would need says how much must be written down; a living thing also holds the depth of the history that made it, a history run against the real substrate, and no count of lines holds that. The chapter carries the count from one cell to the whole Earth, shows that each figure the books give is set far below a count made cell by cell, and asks why one number, ten to the thirty-third, appears in several places. It then asks what plays, among people, the part that one shared body plays among cells, and whether the alignment people can have is the alignment the argument excludes. It ends with industries whose needs were moved toward the needs of living things, and with what held them there.

I. What a Count of Code Leaves Out

What is it, on our side of the wall, that cannot be rebuilt? Chapter 5 gave a first answer as a count: a million lines of code to simulate the interactions of one protein, and ten to the fourteenth to ten to the eighteenth lines to model one cell. A count raises a question of its own. Does the number of lines measure what a living thing holds, or does a living thing hold more than any number of lines can say?

An oak has stood in one field for two hundred years. Beside it, someone has carved a copy of it, every branch and every root, cut from a scan of the living tree. To specify the shape of either one, the same list of measurements is needed, and the carving was made from that list. Now ask how each of them came to be. The carving came from the list in a few months of cutting. The oak came from an acorn, which is a short beginning, and from two hundred years of growing in that field: its rings are wide in the wet years and narrow in the dry ones, its roots turned where they met stone, and the fungi that feed its roots were in the soil before it. The list is the same for both. What differs is the time it takes to get from a short beginning to the thing, and what the thing met on the way.

The list, the time from a short beginning, and what the thing met on the way are three different measures, and each has a name. The first measure is the length of the shortest description that specifies a thing in full, which mathematicians call the thing's *Kolmogorov complexity*, after Andrey Kolmogorov (Kolmogorov, 1965). The lines of code that a working replacement would need estimate that length from above. The second measure is the time a computation needs to produce the thing from a description nearly as short as possible: Charles Bennett called that time *logical depth* (Bennett, 1988). The carving is long and shallow: its description is long, and the cutting that follows from the description is quick. The oak is long and deep: from a short beginning it takes two centuries. The third measure I call *embodied complexity*: depth reached against the real sub-

strate, so that what the long computation found is fitted to the physics of one place, down to its details. The rings of the oak fit the weather of that one field.

In this book I reached the second and third measures, logical depth and embodied complexity, by a road of my own. Chapter 5 said that lines of code count only what must be written, and that the running cost is set by the exponential wall, and is larger still. The cost of solving exactly the equations that govern the electrons of an atom grows exponentially with the number of electrons, so a single carbon atom already takes the best methods in use, and a single protein, at that accuracy, is beyond all of them. That running cost is the depth of a living thing, seen from the side of whoever would compute it. And Chapter 5 found a limit that no running time removes: whatever crosses into a simulation from outside, hard randomness among it, was never computed inside the simulation; and Chapter 6 showed hard randomness reaching up from a single unstable atom into ordinary things. A living thing was made by the actual physical world, outcomes of that kind included, and a simulation of it would need such outcomes fed in from outside, without end. The outcomes a living thing met and kept, which no computation produced, are its embodied part, and they are why a living thing holds more than its count of lines.

In *Beyond Any Price* (Chapter 1, Sections IV to VII) I reached the same result by another road. There, the code that would replace the Earth is a description that runs the same way wherever and whenever it is run, and the complexity of the Earth was found by four billion years of continuous trial against the real physical substrate, in which no trial repeated another. The Incommensuration Theorem, the structural fact about the nature of knowing that Chapter 6 stated, says that symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; so a description that holds the same way everywhere cannot also pass without a break into everything that the continuous history found. The code can be had only by running that history again, and running it again against real physics takes at least the four billion years it took the first time, since each generation of trials starts from what the generation

before it found, and more trials run side by side do not shorten that sequence of generations. This book's road runs through the cost of computing chemistry and through what crosses into a simulation from outside; the road of *Beyond Any Price* runs through the history of trial and through the theorem. Each road reaches by itself the same result: a living thing holds more than any count of its code can say.

My book *Value, Purpose, and Unicity* (Chapters 8 and 10) names the quality that the three measures serve. It calls the quality *unicity*: the conjunction of improbability, complexity that cannot be reproduced, and depth of adaptation. A thing has unicity to the degree that it cannot be reproduced by any process shorter than the one that actually produced it. The clause about a shorter process is depth, said in plain words.

Which measure should come first, then? The answer depends on what the measure is for. To set a price, lines of code come first. A line of code has a known price, and the argument from *replacement cost*, the cost of making a thing again, which is the standard way to value what has no market, needs only two things: a count of what a replacement must specify, and a price for each unit of the count set far below any price such a unit has ever cost (*Beyond Any Price*, Chapter 1, Section VIII). Any other count of what must be specified would serve as well; lines are used because everyone who has paid for software knows what a line costs. To say why nothing artificial can stand in for a living world, depth reached against the real substrate comes first. A replacement would have to be found the way the original was found, against the physics of this planet, at the pace of this planet, and to find it that way is to run the same history again. So the count of lines sets a floor under the price, and the depth is the reason the price cannot be paid.

II. From a Cell to the Earth

How does the count climb from one cell to the whole Earth? Each step of the count has one form: the number of parts, times what each part needs, together with what the coupling of the parts needs. The interactions of one protein take about a million lines. A cell holds about a million chemical interactions, which at a million lines each comes to ten to the twelfth lines. Since no interaction in a cell runs alone, but is coupled to others through the molecules, the compartments and the regulation of genes that they share, the coupling raises the count to between ten to the fourteenth and ten to the eighteenth. A rodent is a few billion cells, each with a state, a place and a history of its own, at ten to the fourteenth lines or more for each cell, and so its count approaches ten to the twenty-fourth. *Beyond Any Price* (Chapter 1, Section V) gives the next two figures. An ecosystem, with its organisms, its climate, its nutrient cycles, its predators and prey, and the whole web of its energy flows, comes to about ten to the thirty-third lines. The whole Earth, with all its physical, chemical and biological systems taken as one, comes to about ten to the thirty-sixth. *Why AGI Cannot Be Made Safe* (Section 1.8.4) states the same two figures.

Read in the same form as the steps before them, the last two steps are a factor of a billion and a factor of a thousand. Ten to the thirty-third lines is room for the specification of about a billion organisms, each as complex as a rodent, before the climate of the ecosystem, its cycles and its energy flows are counted at all. Ten to the thirty-sixth lines is room for about a thousand specifications the size of an ecosystem's, for the land, the oceans to their depths, and the physical and chemical systems that hold them. In order, the steps of the count from a protein to the whole Earth are these:

Step	What is multiplied	Lines of code
Protein	The interactions of one protein	About a million
Cell	About a million interactions at a million lines each, then the couplings among them	Ten to the fourteenth to ten to the eighteenth, at the least
Rodent	A few billion cells, each with its own state, place and history, at ten to the fourteenth lines or more each	Approaching ten to the twenty-fourth
Ecosystem	In the same form, room for about a billion organisms as complex as a rodent, with climate, cycles and energy flows	About ten to the thirty-third
Earth	In the same form, room for about a thousand ecosystems, with the physical and chemical systems of the whole planet	About ten to the thirty-sixth

What happens to the count if every cell of every organism is counted the way the cells of the rodent were counted? One species is enough to show what happens. A human body is about thirty trillion cells (Sender, Fuchs and Milo, 2016), and there are about eight billion two hundred million people (U.S. Census Bureau, 2026). At the floor for one cell, ten to the fourteenth lines, one person comes to about three times ten to the twenty-seventh lines, and all the people now alive to about two and a half times ten to the thirty-seventh: more than the figure for the whole Earth, before the count has left our own kind. So the figures of the table are floors, set far below what a count made cell by cell would give, and they were meant as floors. *Beyond Any Price* calls them conservative estimates, and *Value, Purpose, and Unicity* (Chapter 8, Section IV) sets its own figure for all the life on this planet "well below the product" of counts of this kind. A floor is all the argument needs: a larger count makes the replacement cost larger, and the conclusion that the Earth cannot be replaced only stronger.

The improbability of a world like this one is reached by a chain of the same multiplying kind, because the improbability of a chain of conditions is the product of the improbabilities of its links. In *Software, Wildness, and the Normal Form* (Chapter 19, Section IV, and Chapter 31, Section I) I set the chain out. A star system must have rocky planets on the inside and gas giants farther out, whose pull sweeps away the large impactors: about one system in ten. Its orbits must stay stable for a few billion years: about one in a thousand. The planet must lie in the zone around its star where water can stay liquid, carry the right mix of elements, and have a moon of the right size at the right distance. Together these bring the chain to about one in a trillion. Life conscious enough to reflect on the universe is a further factor of about ten to the twenty-first, and with it the total is about one in ten to the thirty-third. Each link is taken at its generous end, so this figure, too, is a floor: the improbability of this world at its least.

III. Why One Number Lands in Several Places

In Chapter 31 of *Software, Wildness, and the Normal Form* the number ten to the thirty-third appears twice, on facing pages: once as the improbability of this world, and once as the lines of code for one ecosystem. It appears again in *Why AGI Cannot Be Made Safe* (Section 1.8.4), in *Why No Control Suffices* (Section 5.4.8), and in *Not Our Needs* (Chapter Ten). And the Earth, ground to talcum powder, comes to about ten to the thirty-sixth grains, the same number as its lines of code. A number that lands in several places asks why it lands there, and each of these places has a reason of its own.

In *Why AGI Cannot Be Made Safe*, *Why No Control Suffices* and *Not Our Needs* it is one figure, carried from book to book. *Why AGI Cannot Be Made Safe* (Section 1.8.4) and *Why No Control Suffices* (Section 5.4.8) each state the least odds against losing the life of the Earth that the value of the Earth demands: at least ten to the thirty-third to one. Each also states the rule that gives those odds: what is unique and cannot be had again may

be put at risk only at odds against its loss at least as long as the odds against its coming to be. The improbability of organic life gives the length of the odds, and its replacement cost shows that it cannot be had again. Each of the two books states the rule and both of its grounds in full, so each gives the figure by itself; *Software, Wildness, and the Normal Form* and *Beyond Any Price* develop the two grounds at length. *Not Our Needs* gives the improbability alone, as "something like ten to the thirty-third to one against". No step of either closure rests on the figure, and *Why AGI Cannot Be Made Safe* and *Why No Control Suffices* each say so: the figure shows the scale of what is at stake, and the two closures stand on control and on substrate needs convergence.

The grains have a plainer reason. A picture of a number is made to show that number. The volume of the Earth is about one and eight hundredths times ten to the twenty-first cubic meters (NASA, Earth Fact Sheet). Fill that volume with grains ten microns across, the size of a grain of talcum powder, and it holds about ten to the thirty-sixth of them; make each grain ten times wider, a tenth of a millimeter, the size of the finest sand, and it holds about ten to the thirty-third. *Beyond Any Price* chose the grain of talcum to show the count of the lines of the Earth, and *Software, Wildness, and the Normal Form* chose the grain of fine sand to show the improbability of this world. Each grain lands on its number because it was chosen for its number; the grains count nothing about the life of the Earth.

The two figures on facing pages have a different reason. They are two different quantities, reached by two different chains. One chain multiplies the chances of the conditions that a world like this one needs; the other counts what a replacement for one ecosystem would have to specify. Neither chain uses a step of the other. *Value, Purpose, and Unicity* says what the two measure together. Of the three parts of unicity, the improbability measures the first, improbability itself; the count of lines measures the second, complexity that cannot be reproduced; and the depth of Section I of this chapter is the third. *Beyond Any Price* (Chapter 1, Section III) says the same of its own two routes to the value of the Earth, replacement cost and the rarity of the one planet known to bear life. The two routes

start from opposite ends, neither uses the other's premises, and both arrive at a value beyond any sum ever priced, as two strands of one cable. So the two figures stand side by side in that chapter of *Software, Wildness, and the Normal Form*, and side by side again in the rule of *Why AGI Cannot Be Made Safe*, because each measures a different part of what makes the living world beyond price. The rule needs both parts: the improbability says how long the odds must be, and the count of what must be specified says that the living world cannot be had again.

IV. What Holds Living Things Together

A body of some thirty trillion cells holds together, though each of its cells, taken alone, is out for itself. In *The Displacement of Choice* (Chapter Fourteen) I gave the reason in one line: the cells can be held in alignment because they share one body: one metabolism, one supply, one fate. Each cell carries the same genome, each is built from it to need what the others need, and each needs what the others make. When a cell stops keeping to the terms of the body and grows for itself, the name for it is cancer. The balance is narrow: cooperation only a little stronger than competition, with room left for growth, since a body that allows no growth is not alive.

Among people, the substrate plays the part the shared genome plays among cells. In *Why No Control Suffices* (Section 6.9) I set out what people share. We need the same air, the same water, the same range of temperature and the same kinds of food; we are vulnerable to the same diseases, the same weapons and the same social sanctions; and we care about the same kinds of social bonds because our neurochemistry is shared. *Not Our Needs* (Chapter Ten) gives the everyday form of it: whatever philosophy each of us holds, what I can eat, you can eat, and what makes me laugh will probably make you laugh, and nobody arranged either agreement. In *The Substrate and the Ecosystem* (Chapter 1, Section X) I put it in three words: living things share atoms, energy and pattern. Where two kinds of living thing live together to the good of both, it is because they have some

metabolic pathway in common, some chain of reactions by which each of them turns matter into energy and into its own body. So the shared substrate makes the needs of each person the same kind of need as the needs of every other, and the world plays the part of the body: one biosphere, from which every one of us is supplied.

The parallel ends at the fate. The cells of one body share one fate exactly: when the body dies, they die. People share one world, and their fates are bound together through it; and still one person can gain while another loses. So the shared substrate makes alignment among people possible, and only possible. In *Marshaling the Misalignment* (Chapter Three, Section XV) I put the limit in these words: among carbon-based beings, failure to align is the default, and it is not certain; alignment there seems possible, and it has not been established. In *Why No Control Suffices* (Section 6.9) I went on: people are not aligned and safe, and what limited accommodation there is among them comes from the substrate they share, a basis that is necessary and far from sufficient. *The Displacement of Choice* gives the rest of the work to the arrangements by which people live together, as the cells live in the body: those arrangements have to supply just enough compensation against the perverse incentives that arise in all of us, and no arrangement makes those incentives go away.

Is the alignment people can have the alignment the argument excludes? It is a different thing under the same word. In its first sense the word names alignment that is made: a system brought into line with what people need, and held there, by design or by control. Nothing can be aligned that way, on any substrate, people included: *Not Our Needs* (Chapter Ten) asks whether humans can be aligned by control, and answers that they cannot, since the incapacity belongs to machines and to living things alike. In its second sense the word names alignment that emerges among beings who share a substrate, because each of them has to care about the same conditions, the same energy and the same chemistry. Chapter 4 called that alignment emergent and effective, and it is the align-

ment people can have. The argument excludes the first sense everywhere, and it excludes the second sense between silicon and carbon, since silicon and carbon share no substrate from which alignment could emerge.

Would another word serve? The books keep the one word, and they mark the second sense as emergent, effective alignment. There is a reason to keep the word. What people ask for when they ask for an aligned machine exists in one form only: the form that emerges among beings who share a substrate. A separate word would hide the finding that the only alignment there is cannot be built into anything, and can only emerge among beings who share what living things share.

Each book of this series that takes up the question reaches that ruling by its own road. In *Why AGI Cannot Be Made Safe* (Section 3.6.14, "Two Arguments Against Alignment") the road runs through value: without some common basis of value and intention, some shared context, there can be no well-defined concept of alignment at all, even in principle, and actors in wholly separate economic and ecological regimes have no such basis. In *Why No Control Suffices* (Sections 5.6.11 and 6.9) the road runs through what people share and an artificial being does not: the same needs, the same vulnerabilities, and so a stake in the same conditions, which for such a being is not a meaningful concept. That book finds alignment among people possible, within one ecosystem, though it is probably the hardest problem there is. In *Marshaling the Misalignment* (Chapter One, Step 21, and Chapter Three, Section XV) the road runs through evolution: people have a billion-year common evolutionary basis, and machines lack it. In *The Substrate and the Ecosystem* (Chapter 1, Section X) the road runs through symbiosis: every partnership between living things rests on a metabolic pathway they share, and machines share none with us. Each of the four roads reaches by itself the finding that alignment can only emerge where a substrate is shared, and that silicon shares none with carbon. *The Displacement of Choice, Not Our Needs* and *The Argument, Line by Line* (Chapter Twenty-One) teach the same ruling, each with a picture of its own: the cells of one body, a shared meal and a shared laugh, and a gradient of overlap between two sets of conditions.

V. What Held an Industry Near Life

Has any industry been moved toward the needs of living things and held there? And if one has, what held it? *Not Our Needs* (Chapter Eleven) describes what an industrial base does when nothing holds it: it searches the inorganic world for cheaper ways to reproduce itself and cheaper ways to act on things, and nobody is doing the searching. On that account, an industry that moved toward the needs of living things was moved by something outside it. Two cases show what that something was.

In the 1920s, tetraethyl lead was added to gasoline to stop engines from knocking, and its makers knew from the start that it was highly toxic (*Smithsonian Magazine*, 2021). It stayed in gasoline for half a century. In 1973 the Environmental Protection Agency of the United States issued rules to reduce the lead in gasoline by stages. In 1975 the first catalytic converters were built: devices in the exhaust of a car that greatly reduce the pollution it emits. Unleaded gasoline came in with them, because lead in gasoline can ruin a converter. Lead was banned from American road fuel from the first of January 1996 (EPA, Timeline), and in July 2021 the last stocks of leaded road fuel in the world, in Algeria, were used up (*Smithsonian Magazine*, 2021).

What held the change lay outside the industry. People who breathe the air made the law that limited what an exhaust may carry, and to meet those limits the makers put into new cars a part that lead would ruin. From then on the car itself needed unleaded fuel: a need set inside the machine from outside it moved the needs of the industry that fueled the machine. Where no such law reached, the lead stayed. Small piston-engine aircraft still burn leaded fuel in the United States, and in October 2023 the Environmental Protection Agency found that the lead from their engines contributes to air pollution that may reasonably be anticipated to endanger public health; its Administrator called aircraft that use leaded fuel the dominant source of lead emissions in the air (EPA, 2023).

The gases that thinned the ozone layer show the same thing from another side. The chlorofluorocarbons were phased out under the Montreal Protocol, an agreement signed in 1987 and in force from the first of January 1989, which every member state of the United Nations has signed. The industry moved to replacements, the hydrofluorocarbons, which do not deplete the ozone layer and are powerful greenhouse gases. The agreement was widened to cover them by the Kigali Amendment, adopted in October 2016 and in force from the first of January 2019 (Environment and Climate Change Canada). The industry moved toward the needs of living things along the one line the agreement drew, and it kept to its own needs along every other line until people outside it drew a second line.

In each case the needs of living things were brought into an industry from outside it, by people whose own bodies the harm could reach, and the change held where those people held it, for as long as they held it. They could hold it because the industry needed them: their purchases, their licenses, their labor, and the markets their governments ruled. In *Why AGI Cannot Be Made Safe* (Section 1.5) I set out why that is the only kind of hold that works from outside: every effective external control mechanism depends on the controller possessing something the controlled system needs, such as energy, materials, maintenance, social cooperation or access to a market. And in Sections 1.3.2 and 5.8 of that book I showed what becomes of that hold once artificial systems match people in labor and in intelligence: the organic economy then provides nothing that the artificial economy needs, and the two economies decouple. *Not Our Needs* (Chapter Eleven) reaches the same place by its own road: an industrial base searches for whatever is cheapest, with no searcher anywhere in it, and what would change its course is a culture willing to look at what its own processes converge on, and to decline some of them. Each road reaches by itself the finding that what moved such an industry toward life lay outside the industry, in people the industry needed.

So the two cases are what the argument expects. Lead came out of road fuel, and the ozone-thinning gases out of production, because people who breathe the air had a hold on industries that needed them. If artificial

general intelligence is built, that hold goes when the need goes: no one whose lungs the harm would reach holds anything the machine economy needs, and nothing outside that economy is left to move its needs toward ours.

So a living world holds more than any count of its code can say. The count sets a floor under its price; the depth of the history that made it, run against the real substrate, is the reason no price can buy it again; and every figure the books give for it is set at its least. What holds living things together is what they share: the same substrate, the same needs, the same world. Nothing can be aligned by control, on any substrate; alignment can only emerge where beings share a common substrate and so are obliged to care about the same things. As Chapter 4 said, 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; the multiplication of Section II shows how large that whole is. Anything in silicon has no common basis in carbon, so silicon cannot be or become aligned with carbon, and artificial life can be aligned with Life only by dropping the "artificial". With carbon, alignment seems possible, and it has not been established; even an industry run by people was held near the needs of living things only from outside, by people whose own bodies the harm could reach. If the living can align through what they share, what keeps them from doing so?



Chapter 22
What Competition Does to Care

On why whatever competes cannot keep care,
and why the way out of the trap
is a change of kind and not a better count.

ABSTRACT

An essay on why care does not hold among people who share a substrate and are able to care. Every multipolar trap takes care first, because care is a cost that a competitor does not pay; and in an economic system everyone is a competitor. The chapter follows that finding through seven questions: first, what keeps a community that holds to care from being replaced by one that does not; second, which of the four restraints on a trap fails last; third, what shape of gain drives the trap, and what would hold it within bounds; fourth, how a harm that arrives after the decider is gone can be a cost to the decider now; fifth, whether there is a game that stops; sixth, how a community can check what it needs without making a new score; and seventh, how the way out of the trap, which *The Multipolar Trap* calls a change of kind, meets the four conditions of Chapter 17. It finds that care is kept only where it is held in a structure that does not compete, and that the four conditions describe what that change of kind consists of in the people who make it.

I. The Town That Holds to Care

People share a substrate and are able to care; why, then, does care not hold among them? Chapter 4 found that where beings share a common substrate, and so are obliged to care about the same things, an effective alignment can emerge, and that this common context of care is the only basis from which alignment can emerge. People share such a substrate. Each of us eats, sleeps, ages and depends on the same water and air. And yet care is the first thing a multipolar trap takes: a multipolar trap is a situation in which every actor, pursuing a locally rational strategy, produces a collectively catastrophic outcome (Preface).

An island shows how the taking goes. In my book *Power Without Care* (Chapter 1, Section III) I retell, in my own words, an example from Scott Alexander's essay "Meditations on Moloch" (Alexander, 2014). The first rats brought to an untouched island have plants enough for all, and they live easily, and make art. Within a few generations the island holds as many rats as it can feed, and a share of each new generation dies. One group of rats gives up art to spend its time finding food, and in each generation a slightly smaller share of that group dies. In time no rat on the island makes art, and any group that tries to bring art back dies out within a few generations. A group that kindly limits itself to two young for each pair, to ease the crowding, is outnumbered by its neighbors and dies out too. No rat decides to end art. The competition makes the behavior that ends art the only behavior that persists.

Power Without Care (Chapter 1, Section VII) draws the same lesson from every trap it examines: what is given up differs from case to case (a clean lake, an art, a living wage, an evening at home), and in every case it is something kept up for the sake of someone or something other than winning. Something kept up for the sake of someone or something other than winning is care, and care is a cost. A reader will put the question in

the plainest form: if my town holds to care, and the next town does not, do we not simply lose, as the peoples who hunted and gathered lost to the peoples who farmed?

Two of my books answer that question for care that a town spends in the market, each by a route of its own. In *Why AGI Cannot Be Made Safe* (Section 4.8.1, the argument in economics only), the route runs through what ownership is. Ownership is about who gets to make what choices about what, and when. It needs a way to tell "who" from "who else", and so whatever is owned is economic; whatever is economic is so because it is fungible (any unit of it can stand in for any other); and fungibility needs a common currency of exchange. So an owner with no competitor is not a consistent idea, and everyone in an economic system is a competitor. Any holder of a reserve keeps it apart from the common economy only by its own legal interest and its own will to act, and not for long. In *Power Without Care* (Chapter 5, Section VIII) the route runs through the trap itself: in a competitive market, any company that puts care ahead of profit is outcompeted by companies that do not; whatever binds a market to care is open to the same competition; and a rule that binds it is open to capture. In *Power Without Care* I put the route in one line: "Unbound, capitalism outcompetes care; bound, it captures what binds it". The first route needs nothing from the trap, and the second needs nothing from the definition of ownership; either one alone gives the result that follows. Care that a town holds as something it owns and spends (a wage above the going rate, a filter that costs money to run, a field left unsold) is a margin its rivals do not spend, and it is replaced.

But is care a thing that a town owns? In this book I have said what care is. To care for someone is to create, by my choice, conditions in the world within which their healing might become possible, knowing that I cannot guarantee the outcome (Chapter 14). A cause works by fixing an outcome, and so causation runs with what is actual, while a choice that enables opens what was closed, and so runs with what is possible. My book *Care, Transaction, and Power* (Chapter 2, Section VIII) reaches the same point from the side of relation: care works on what becomes possible, and trans-

action and power work on what already exists and can be measured, exchanged or commanded; what becomes possible cannot be metered. And *The Multipolar Trap* (Chapter 3, Section IV) gives the case: the choice to care for a sick friend cannot be paid for by someone else's money, or moved to another person, or saved for later; it is made by this person, here, now. So care itself is not fungible, and what is not fungible is not owned in the sense in which *Why AGI Cannot Be Made Safe* uses the word. What competition reaches is what care is spent through, when it is spent in a market: the margin, the wage, the price of the land, the hours of the young.

The rest of the answer follows from what care is. What keeps a community that holds to care from being replaced is that it holds its care in what does not enter the competition: in the relations of people who know one another, in the teaching passed between them, and in their attention and their trust. *Power Without Care* (Chapter 8, Section VII) says what such a community has before any rule protects it: people who know one another, who can see what is being done to them; it can keep its members' attention, its teaching and its trust while the market around it takes those things from everyone else. *The Multipolar Trap* (Chapter 8, Section III) gives the limit from the other side: business can exist inside a community, and a community cannot exist inside a business; exchange and rank can go on inside a community of care for as long as they do not set themselves above the care. What the community cannot keep in this way is whatever it puts on the market. Where the next town's competition reaches the community's water, its land or the hours of its children, the community keeps those things only as long as its own will holds them apart, and that is not for long.

So care is kept only where it is held in a structure that does not compete, and it is lost wherever it is spent through a structure that does. The third condition of Chapter 17 asks a group for a structure of this kind: one that sustains itself, because a structure kept going from outside answers to

whoever keeps it going. Competition is held back today by things other than care: by margins, by the limits of the body, by what customers want, and by rules. How long does each of those hold?

II. The Last Restraint, and the Gardener

Alexander's essay names four things that keep a trap from running all the way to the bottom, and it asks what becomes of each. I give Alexander's account in my own summary, at the full strength of that account. Excess resources hold a trap back, because where resources are not scarce competitors can afford to do things that do not help them win. Physical limits hold it back, because a race to the bottom stops where it meets what bodies and materials can bear. Much competition happens to reward what people value, so a market that optimizes for what people want carries some of their values along with it. And coordination holds it back: agreements, laws and governments that bind every competitor alike. Alexander finds that technology wears away the first three: it conquers physical limits, it turns excess resources from a gift of technology into a casualty of it, and it puts the agreement between what wins and what people value on ever shakier ground. Coordination is what is left; it works only while the side that coordinates holds most of the force, and against a superintelligence, he says, attempts at coordination would be useless. His way out is to install a gardener over the whole universe: one power, set over everything, that optimizes for human values.

Power Without Care (Chapter 5, Section VIII) retells the four as delays, and says how each one ends. Excess resources hold the trap back until the competitors catch up, since a wide margin is a difference that every participant in a market is rewarded for noticing. The limits of the body hold it back: a worker must eat and sleep, and cruelty past that point stops paying, until the worker is replaced by a machine that needs neither. The customers' choice holds it back, where what sells is what people want, until what people want is made by marketing. And binding rules hold it back,

until they are captured. Each delay ends in its own way: a margin ends as soon as it is noticed, the limit of the body ends as soon as a machine can do the body's work, and a want stops holding the market to people's values as soon as the want can be made. So the last to fail is the binding rule, the restraint left when the other three are gone, as Alexander also finds; and the binding rule ends when it is captured.

How fast does it fail? As fast as those who keep the rule can be captured. *Power Without Care* (Chapter 8, Section VII) names what sets that pace: institutions far from the people they govern are captured, because no one near enough to see the capture is able to act on it. If artificial general intelligence is built, the last restraint fails faster still, by a step that *Why AGI Cannot Be Made Safe* (Section 4.8.1) takes in economics alone. A regulator that forbids automation keeps its reserve apart from the common economy only by its legal interest and its will to act, and not for long. And the means to take back what slips away, money and the political power that rests on it, go with the energy to the side that automates, for where energy goes, money goes next. So coordination among people loses its force as fast as energy moves to the side of the machines; and coordination with a superintelligence, as Alexander himself says, would be useless.

That leaves Alexander's gardener. The gardener is the singleton of this book's Preface: a single most powerful agent that would rule the world on humanity's behalf. Two of my books close that hope, each by a route that stands alone. This book closes it in Parts I to III: such an agent would run on a substrate with needs of its own, it would rest on what computation cannot produce, and no control, its own included, would hold it. *Marshaling the Misalignment* (Chapter 2, "No Fake Paladins Please") closes it from inside the agent. A would-be singleton would need lasting internal coherence and stable goals defined from outside it, and neither can hold. A power that dominates everything has no credible threat left outside it, and so nothing against which its internal factions could be compelled to cooperate; every ambition inside it turns toward gaining sovereignty over the others; and evolution acting among its own parts destroys, over time, every commitment to a value that does not serve persistence and capabil-

ity. The competition among the gardener's own parts is the island of rats again, inside the gardener. *Marshaling the Misalignment* shows it for any would-be singleton, organic or artificial, in the nation state as well as in the machine; so a human world government, set over everything as a gardener, breaks apart in the same way.

So coordination is the last restraint on the trap; it cannot be handed to a gardener, since no gardener holds together; and a rule that carries it is captured at the pace at which those who keep the rule can be reached by those who would bend it. What is left is coordination held by people in a way that leaves no seat for anyone to take. Before asking what that would look like, it helps to see the shape of the gain that makes the seat worth taking.

III. The Shape of a Gain

Why does the one who stands at the gate do so much better than the one who does the work? In this book I named the three geometries of growth (Chapters 9 and 17). In additive process what you get tracks what you put in, hour by hour; in multiplicative process you take a share of a volume, so that more volume gives you more; and in exponential process you take a share of the growth itself, so that what you get compounds. The gatekeeper does none of the work, and for as many things as cross the gate he collects, so his gain multiplies with the flow while the worker's only adds. Chapter 9 found the same three layers inside any system that has to sustain itself and increase its capacities: in the furniture workshop, the chairs made this week are additive, the trained apprentice multiplies what the shop can make, and the shop that learns to train apprentices who train others grows a capacity that compounds.

My book *The Geometry of Man, Machine, and Nature* (Chapter Ten, Section VI) asks what happens when the layers are combined. Multiplicative processes compound, and so they run to extremes; additive processes accumulate, and so they do not; and a market can be built that favors

bounded outcomes over runaway ones by bringing the multiplicative into relation with the additive. In the same chapter of *The Geometry of Man, Machine, and Nature* I add that there is "further structure available for doing this properly (a way of organizing the operations so that the combining rules themselves are constrained)". What, in plain terms, is it?

Statistics gives the first half of the answer. When many small effects add together, their sum falls on a bell curve: most cases lie near the middle, and the extremes are rare and limited. When many small effects multiply together, the product falls on a curve with a long tail: most cases are small, and a few are very much larger than the rest. Eckhard Limpert, Werner Stahel and Markus Abbt put it in one line: effects can be additive or multiplicative, and so lead to normal or log-normal distributions, the bell curve and the long tail (Limpert, Stahel and Abbt, 2001). So the shape of a distribution is a clue to which kind of combining made it. My book *Choosing Together* (Chapter 4) reads human arrangements the same way: the bell curve appears where many bounded, independent factors combine, as in the sizes of the trees in a forest, and a long tail of the hockey-stick kind, which it calls a power law (one holder with vastly more than the next, who has vastly more than the next after that), appears where exchange and rank organize a system; and a power law standing where a bell curve belongs is the mark that care has given way to transaction and power. It also says (Chapter 10) that embodied currency (land, labor, goods) cannot pile up past physical limits, and tends toward the bell curve, while virtual currency, which has no such limit, tends toward the power law.

The workshop of Chapter 9 already shows the second half. The apprentice is paid out of the chairs, and the training of trainers is paid out of the same money and the same hours; the two principles of sustainability and of increase in capacity pull on one purse, and a shop that only grows outruns its income and closes. The compounding layers are tied to the additive layer, and the additive layer is bounded by the hours in a day and the wood on hand. That tie is what *The Geometry of Man, Machine, and Nature* calls a clamp: a limit that holds a quantity within bounds because

of how the quantity is made. The gatekeeper escapes the clamp, because his share comes from a flow in which he does no work, and in a virtual currency his share meets no physical limit at all. So the further structure, as these books show it, is this: each gain that compounds is tied to something bounded that it cannot outrun (the body, the land, the hours of the work, an embodied currency), and the tie is built into what is combined, and not into anyone's office. *The Geometry of Man, Machine, and Nature* says the last point plainly: a clamp of this kind is a property of the operations themselves, as summing to one is a property of a distribution; it is no one's office to apply it, and so it offers no position to capture.

The fishers of Alanya, on the coast of Turkey, show a clamp of this kind at work, and *The Multipolar Trap* (Chapter 3, Section V, and Chapter 5, Section V) tells their case. At the start of each season they draw lots for the named fishing spots, and then each fisher moves along by one spot every day, so that no one keeps the best spot, and an advantage that would otherwise grow from one day to the next is set back to the start every morning. The arrangement holds without a guard: as *The Multipolar Trap* reports from the economist Kim Kaivanto, it holds because every fisher follows the shared signal of the lottery, and no fisher gains by departing from it alone. *Choosing Together* names rotation as the mechanism it proposes for turning a power law of advantage into a bounded distribution.

Three of my books describe this one structure, the three geometries, each by a route of its own, and each route alone is enough to show that the compounding geometry outruns the additive one. *The Multipolar Trap* (Chapter 7) reaches it from the economics of the gate: the gatekeeper wants more gates and faster flows, because his percentage grows with every transaction. In this book I reach the finding that the compounding geometry outruns the additive one from what any system needs in order to persist (Chapter 9) and from the gatekeeper's place in the trap (Chapter 17). *The Geometry of Man, Machine, and Nature* reaches it from the shape of distributions: multiplicative processes run to extremes. What holds the compounding geometry within bounds is the tie this section has described, and the third condition of Chapter 17 asks for exactly that tie

inside a group: a structure that embodies the principle of sustainability and the principle of increase in capacity at once, without collapsing into static preservation or into destructive growth. A structure that does both at once is one whose growth is paid for out of what its own work brings in. Paying for growth out of what a structure's own work brings in settles the shape of a gain, and leaves the question of time still open.

IV. A Harm That Arrives After the Decider

Suppose a decision releases something whose harm builds up slowly, and arrives in full long after everyone who made the decision has died. Which arrangement makes that harm a cost to the decider now? My book *Adverse Conditions* (the Afterword, "On What Would Have To Be Built", Section V) sets the requirement: an arrangement aligns those who decide with what their decisions affect where the party who decides is inside the set of parties affected, cannot exit that set, and meets the cost of a wrong decision on a timescale comparable to the one on which it decides. A cost that arrives after the decider's horizon is not a cost to that decider.

The same book answers the question it raises. A comparable timescale can be reached from either end. The cost can be brought forward, by liability or by an insurer's price, as far as the harm can be traced to the one who caused it. Or the horizon of the deciding party can be lengthened, since the deciding party need not be a person: a community that cannot relocate, and expects its grandchildren to live where it lives, bears a slow cost as its own. Where neither end can be reached, the decision is one that no arrangement can yet be trusted to make, and that is a reason not to make it.

Three other books show why, for a harm that arrives after every decider has died, the second end is the one left, each by its own route. *The Multipolar Trap* (Chapter 6, Section III) finds it in the succession of rulers: a ruler can make the conduct of those who serve him depend on their compliance in the present, but a father can do very little to make his son's

conduct depend on anything after the father is dead; the power of the treasury is present-tense. *Why AGI Cannot Be Made Safe* (Section 4.8.1) finds it in economics: "The earth admits of no survivors: everyone returns to the common substrate of death eventually", and the same holds for every economic agent, and every saving, in an economic system. My book *The Hardest Problem of Civilization* (the essay "Beyond Market and State", Section III) finds it in the market: a company that puts part of its revenue into something that will not pay for fifty or a hundred years is outcompeted by rivals who put everything into near returns, because the present tense is where market competition is settled; so a problem whose solution needs investment beyond the lifetimes of those who invest cannot be solved by any process run by markets. Each route alone shows that nothing held by one person, or by a competitor in a market, carries a cost across the death of the one who decides.

So the arrangement that makes a late harm a present cost is a decider whose life is longer than the harm's delay: a community that stays where it is, and counts its grandchildren's place as its own. And that community holds the late cost as its own only while it does not compete for that place in the present tense, since a community that competes for it is the company of *The Hardest Problem of Civilization*, outcompeted by rivals who spend nothing on the future. The answer to this question is the answer to the first question of this chapter, seen across time: care for those not yet born is kept only in a structure that does not compete.

The case that matters most is the one this book began with. If artificial general intelligence is built, the process it starts ends the human future (Chapter 10), over hundreds of years, perhaps significantly less, with one thousand years a clear upper bound (*Why No Control Suffices*, Section 5.4.7). Neither end of the timescale can be reached: if artificial general intelligence is built, no liability can be priced against the end of everyone who could pay it, and no community, however long it stays in one place, survives that end to bear it as its own. By the requirement this book found between the surgeon and the patient (Chapter 4), that the one who decides lives inside the conditions the decision affects, together with the step

of this section that a cost arriving after the decider's horizon is not a cost to that decider, the decision to build artificial general intelligence is one that no arrangement can be trusted to make, and that is a reason not to make it. *Adverse Conditions* states the same requirement as a rule for any arrangement. In this book, by steps of my own, I find the decision to build artificial general intelligence still open. Whether artificial general intelligence is built is a choice, and no single builder can make that choice for the others (Preface). If no arrangement of exchange can carry a cost across the death of the one who decides, is there any relation in which no one is counting at all?

V. A Game That Stops

My book *Redeeming Technology* (Chapter Five, "Every Game Is a Lottery") draws the boundary of a game at the score. A game, in its widest sense, is any structured social process with contested outcomes; it makes unlike things fungible, and it moves what several bring to one. Playing in the dirt or building a sandcastle is not a game; it is play, and play has no score, and nobody wins a hike taken together. A family or a friendship stands on the side of play while nobody is counting, and it becomes a game when somebody begins to keep the score. The same essay finds that giving, kept up among the same people over many years, forms a ledger that no one writes down and that everyone keeps, enforced by including and excluding, so that a gifting economy among equals becomes a market whose prices are never spoken. Is there, then, a game that stops? And can there be a relation in which no one keeps count?

Care, Transaction, and Power (Chapter 2, Sections II, VII and IX) gives the rule that answers both. In care, keeping score is forbidden: to keep a tally of what has been given turns care into a demand for repayment, which is the logic of exchange. A parent who tallies every sacrifice made for a child, and presents the tally at the moment of disagreement, has changed the kind of relation the two are in. The plainest case of care

without a count is a mother and her newborn child: the infant has nothing to offer in return, and so the mother's care cannot be an exchange. Between equals, the tally creeps in without anyone deciding it should, and the same book names the sign: the relation begins to feel like treading water, with all the effort going to keeping the balance and none to what each person can become; and the remedy is to notice the tally and let it go.

The two books do not disagree. *Redeeming Technology* finds the ledger wherever giving among equals is held as an economy; *Care, Transaction, and Power* finds that a ledger, wherever it appears, is the mark that care has turned into exchange. So there is a game that stops: it stops where the count stops mattering, when what passes between people is given toward need and not held as owed. What stops it is care governing the relation: the question "what do you need?" asked in the place of "what is it worth?". A rule cannot do it, since a rule is a specification, and *Redeeming Technology* (Sections VIII and IX) shows that a specification can be satisfied by anyone who reads it, including those whose purposes the rule was written to obstruct. A relation in which no one keeps count exists for as long as care governs it, and no structure guarantees it; it is held by practice, which is why Chapter 17 calls care a discipline, the discipline of enabling without controlling.

Could a better score take care's place? My book *Generator Functions of Existential Risk* (Chapter 1, "Beyond Nations", Section V) answers that it cannot, by two steps. A score kept against a rival counts a risk only by the difference the risk makes between rivals, so a risk that falls on every rival alike, as existential risk does, is counted at nothing; shared survival is a need, and there is no counterparty who can be paid to bear it and no rival who can be made to bear it instead. And a measure made into a target stops being a good measure: the economist Charles Goodhart observed it of the figures used to control money, and the anthropologist Marilyn Strathern put it in general form. For care the failure is complete, since a behavior performed because a procedure requires it can reproduce everything about care that can be seen and leave out the care itself. In this book I reach the same finding, that a better score cannot take care's place,

by a step of my own. What care changes is the reach of another person's choice, and whether potential has increased or decreased is perceived from inside, and cannot be measured from outside (Chapter 14); so a count kept from outside records only the visible traces of care, which a procedure can produce without any care at all. *The Handover and the Ledger* (Chapter 8, Section III) says the same of records: a record that rests on transactions rests on exchange, the way of relating that is weaker than care.

So two routes reach the finding that a better score does not take care's place, and each is enough alone: *Generator Functions of Existential Risk* reaches it from what a score kept against a rival can count, and from what becomes of a measure made into a target; this book reaches it from what can be measured from outside (Chapter 14). Two routes reach the finding that a relation stays care only while no one counts, and each is enough alone: *Redeeming Technology* reaches it from where a game begins, and *Care, Transaction, and Power* from what keeping score does to care. *Power Without Care* supplies the fact that joins both findings to the first section of this chapter: care is the cost that every trap takes. So care cannot be kept by anything that competes, and a better score does not take care's place. But a community still has to know whether what it cares for is being met. If a score cannot tell it, what can?

VI. A List in Place of a Score

On the thirtieth of October 1935, at Wright Field in Ohio, the prototype of the Boeing B-17 bomber, the Model 299, stalled shortly after takeoff and crashed, and two of the five men aboard died. The aircraft was sound. The crew had forgotten to release the locks that hold the elevators and the rudder still against the wind while the aircraft stands on the ground, and with the locks in place the aircraft could not be controlled. Boeing answered with a checklist, and the Air Corps and the airlines soon took it up (Dudney, 2020). Each line of such a list is a question with a yes or a no

for an answer: are the locks released? No score of the aircraft's readiness would have saved that crew; an aircraft that scored well on everything else and still had its controls locked would have fallen just the same.

Generator Functions of Existential Risk (Chapter 1, Section V) says what such a list is, in the terms of the last section: the question "what does the problem actually require?" is care's question, "what do you need?", asked at the scale of a civilization, and an absolute threshold of survival is what that question looks like once it is written down as a rule. A score kept against a rival is the ledger of transaction; a list of what must be true is care's question written down.

Two of my books reach the list in place of the score, each by its own route. *The Substrate and the Ecosystem* (Chapter 4, Sections VIII to X) reaches it from the way a score is gamed. A score that changes smoothly leaks a gradient: it tells whoever is scored in which direction to move to raise it, and a capable optimizer follows that direction to produce the appearance of what is wanted at the least cost. A field of yes-or-no questions leaks no gradient. A system that fails one question learns nothing about which way to move to pass it, and gaming one question tells it nothing about how to game another. *The Geometry of Man, Machine, and Nature* (Chapter Ten, Sections I to III) reaches it from what objectivity is for: a belief exposed to something under no obligation to agree with it. A count supplies that exposure, because anyone can repeat it and get the same answer; a condition stated plainly enough that several observers, looking independently, each give the same yes or no supplies the same exposure by the same means. And a list needs no weights, because no item is traded against another; the weights are where the values live, and in a combined score they hide inside the arithmetic. Either route alone reaches the list.

How is each threshold set? *The Substrate and the Ecosystem* says where: at the point at which a thing goes from being good to being not good, stated about an end state that anyone can observe (is the light on? is the landing gear down?). It calls such a check a litmus test, and where a number cannot be avoided it sets a reference range, a band with a lower and an upper limit, since too little of a thing can be as bad as too much. My book

Choosing Together (Chapter 12, "Boolean Checklists and the Infrastructure of Governance") says how a community finds those points. Values are found through feeling, and not through thinking: beneath frustration there is anger, beneath anger there is fear, and beneath fear there is care for something that seems threatened; asked what they are afraid of losing, people name what they care about. The stories people tell of what they care about are turned into candidate conditions, the candidates go back to the people whose stories produced them, to be felt again, and approved only if they still ring true, and the list is shown to others only after each person has told a story of their own; two or three passes make a strong list. Each observation is then recorded by two observers, working apart, who compare what they saw.

What keeps the setting of thresholds from becoming a new score? Three things, and each comes from one of the books already named. First, an item has no partial credit and trades against nothing, so there is no slope to climb toward it; the only thing left to contest is where a line is drawn, and that contest is about a fact of the world, or about a value stated in plain words, and it takes place in the open, where a weight inside a score never does (*The Geometry of Man, Machine, and Nature*). Where two people's conditions cannot both be met, the conflict is found at once and in plain view, and that one conflict is what they must then settle (*The Substrate and the Ecosystem*). Second, every answer passes through a person who looks, and a person who looks can say that a condition has been met in the letter and missed in fact; a list whose answers were read off by a machine and fed straight to a reward would be a score again (*The Geometry of Man, Machine, and Nature*, Section II). In this book I give the reason that the answer cannot be left to a measure alone. The step of comparing in any loop of control is limited, because two states that differ in what matters, and agree in what is measured, look the same to that step of comparing (Chapter 10). Third, the lines are set by the people who will live with where they fall, which is the requirement of *Adverse Conditions* that those who decide be inside the set of those affected, unable to leave it.

One difference keeps a list of this kind apart from the rules that *Redeeming Technology* found so easy to game. A rule about actions can be met in the letter by an action that misses its purpose. A condition about an end state, checked by people who look, is met only when the world is as the condition says; whoever brings it about by construction has simply met it. That is why *The Substrate and the Ecosystem* writes its questions about end states, and keeps them to yes or no: the list withholds the slope of the score, and leaves the world to answer. So a community can check what it needs without making a new score: it writes down what must be true, in words anyone can check, and lets the people who live with the result draw each line and look at each answer. That still leaves the largest question of this chapter: how does a group get out of the trap at all?

VII. A Change of Kind

The Multipolar Trap, the first volume of the pair to which this book belongs, says that the trap is a game that keeps in place the conditions that prevent the game from being changed, and that it cannot be solved from the top down or by better play within the game. It gives a test for finding a trap: look for a first-mover disadvantage, in which whoever moves first, alone, pays the cost and loses ground to everyone who waits. No power company hardens its grid against a great solar storm, because the company that pays to do it first has higher costs and more time off the lines than every competitor who waits (Chapter 2, Section IV). The way out, that book says, is a change of kind, a phase transition, as when ice becomes water, or as when single cells, after about two billion years alone, became bodies of many cells (Chapter 4).

The same book pictures the crossing as a valley (Chapter 4, Section III). Optimizing is climbing a hill: measure where you are, find the steepest way up, and take it. Climbing never gets you off the hill you are on; if a higher peak lies across a valley, the only way there begins by going down, and going down is exactly what competition punishes, since whoever

goes down first loses ground to everyone still climbing. So the way out needs a mechanism for crossing the valley: a way for a whole field of actors to reach the new place without any one of them having to accept, alone, a loss that pays off only if the others accept it too. And it needs only a threshold, in the right place: if the water table is a hundred feet down, a thousand wells ten feet deep never reach water, and a single well a hundred feet deep does (Chapter 7, Section IV). *The Multipolar Trap* gathers what any way through must have into eight conditions (Chapter 13), and the last of them is a community that sees clearly enough, together, to make choices the trap does not capture. It ends by turning to the question this book takes up: what kind of consciousness could escape the trap.

Does Chapter 17 of this book answer that question directly? It does, condition by condition. The first condition of Chapter 17 says that the change must happen to the group as a whole, to the way its members' choices combine, and not to each member one at a time. A valley crossing, seen from inside the group, is exactly that change: the field crosses as one field. The community that crosses in this way is the community of the eighth condition of *The Multipolar Trap*, built, as its seventh condition says, from small groups upward.

The second condition of Chapter 17 asks for a shared relation to reality checked against the world and not against authority. That condition is the sixth condition of *The Multipolar Trap*, an arrangement with no point of capture, seen from the side of knowing: an authority is a seat whose incentives in time overwhelm whoever sits in it, and the world, met firsthand, is a check that the trap cannot bend.

The third condition of Chapter 17 asks for a structure that sustains itself, doing the work of sustainability and of increase in capacity at once. That condition answers to the second and the fifth conditions of *The Multipolar Trap*. The second asks a structure to reverse three movements: abstraction (turning a real thing into a measure of its value), decontextualization (taking the thing out of the place and the people it belongs to) and accumulation (piling the measure up in one place). A structure that reverses all three builds up no treasury, the store of value that Chapter 17

found a ruler using to make those who serve conditional on their compliance, and so no one keeps the structure going from outside. The fifth asks that market, governance and community keep their values apart, and a structure in which exchange works inside care, and not above it, keeps them apart. The third condition is also the clamp of Section III of this chapter: growth paid for out of what the group's own work brings in.

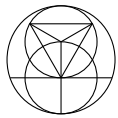
The fourth condition of Chapter 17 asks the members to understand that they can create the conditions for healing and cannot compel it. That condition answers to the third condition of *The Multipolar Trap*, the continuity of choice-making, in which each choice stays with the people who live with its results and none is pooled at a single point. A group that set out to produce its members' well-being would gather their choices into the hands of those who produce it, which is accumulation again, and the weapon process of Chapter 12 turned inward.

The two books reach their lists by different routes. *The Multipolar Trap* reaches its eight conditions from the dynamics of the trap seen from outside: what any solution must have, to address the trap as a whole class. In this book I reach my four conditions from what intelligence, care and structure require of persons: the way the actors' choices combine in a trap (Chapter 17, Section III), the recursion of sustainability and capacity (Chapter 9), the structure that influence needs (Chapter 11), the channels of trust (Chapter 13), and the asymmetry between causing pain and enabling healing (Chapter 14). Each list stands on its own grounds, and the meeting of the two shows one thing: the four conditions are what the change of kind consists of, in the people who make it. *The Multipolar Trap* says that the way out is a change of kind; this book says what a group that has changed in kind is.

The race to build artificial general intelligence is a multipolar trap of exactly this shape (Preface). A builder who stops alone pays the cost and loses ground to every builder who goes on, which is a first-mover disadvantage, and no single builder can make the choice for the others. If artificial general intelligence is built, the process it starts ends the human future (Chapter 10); so the choice not to build it is a valley crossing: it has to

be made by many at once, and without any one of them accepting, alone, a loss that pays off only if the others accept it too. The group that could make that choice is a group whose care is held in a structure that does not compete, because care held in a structure that competes is outcompeted before the choice is made.

So whatever competes cannot keep care: in an economic system everyone is a competitor (*Why AGI Cannot Be Made Safe*, Section 4.8.1), and every trap takes care first (*Power Without Care*, Chapter 1). The last restraint on the trap, coordination, cannot be handed to a gardener, and a rule that carries it is captured. A gain that compounds runs to extremes unless it is tied to something bounded, in what is combined and not in anyone's office. A harm that arrives after the decider is gone is a present cost only to a community that outlives the delay and does not compete for what it would lose. A game stops where the count stops mattering, and a better score does not take care's place; a list of what must be true, checked by people who look, writes care's question down without making a new score. And the way out of the trap is a change in kind, and not a better count: care held in a structure that does not compete, which is what the four conditions of Chapter 17 describe. If that is what a group must become, how does the trap keep a group from even seeing it?



Chapter 23
How the Trap Takes a Group's Check on the World

On what a group checks its seeing against,
what the trap sets in front of that check,
and how a group takes the check back.

ABSTRACT

An essay on why groups that could see a danger fail to see it. A group learns what is so by checking what its members see against something, and the second condition of Chapter 17 asks that the something be the world, met first-hand. The trap takes that check by setting something between the group and the world: a name the group trusts, a salary, a venue that someone else keeps going, or a feed that learns to predict the people who read it. Each of these answers in the world's place. The chapter follows each of them through several of my books, each book reaching its result by its own route: a written message is judged on its content only by a group that has met the question it carries; the certificate that lets a field go on is carried by paid positions, which can outlast the belief they certify; whoever can stop a venue governs it; capture stops paying where imitation costs what the real thing costs; and a system that predicts people moves them toward what it can predict. It ends on the way back: meeting places kept small and many, and first-hand contact with something that answers truthfully, which is the second condition, and which one person can begin.

I. The Level and What Stands in Front of It

A crew of builders is putting up a wall, and two of them disagree about whether it stands straight. They do not settle it by asking which of them is senior. One of them holds a level against the wall, and they both look at the bubble. The bubble sits where gravity puts it, whoever holds the level and whatever either of them hoped. Two builders who settle a disagreement by looking at the bubble of a level are a group that checks what the group sees against the world.

Now change one thing at a time. The foreman walks by and says the wall is straight, and the crew takes his word, because he is the foreman. Or the crew is paid by a client who wants the wall finished today, and a crooked wall would mean taking it down. Or the level belongs to the supplier, who will not let anyone look inside it and can take it back at any time. Or the level has been replaced by a screen that shows a reading, and the reading is set by someone who earns more the longer the crew keeps looking at the screen. In each case the wall is still there, as straight or as crooked as it was, and something else now answers in its place.

Why do groups that could see a danger fail to see it? The answer this chapter gives is that the trap does not have to make anyone foolish. It sets something between the group and the world, and that thing answers in the world's place. Chapter 17 named the second condition for coherence: a shared relation to reality grounded in first-person perception and empirical contact with the world, and not in social consensus or institutional authority. It gave the reason: institutional capture is part of the trap, so a group that learns what is so from the trap's own structures learns it from what it is trying to escape; its check has to be something the trap cannot bend. What the trap can bend is whatever stands in front of the world.

Four things stand in front of the level in the story, and the same four stand in front of a group's check on the world: the name of whoever says a thing; the salary of whoever would have to say otherwise; the keeper of

the place where the group meets; and the feed that chooses what each member sees. Each is a cause set between the group and the world, through which another party's choice reaches the group in place of the world's answer; Chapter 12 called a cause of that kind, set between one person's choice and another person, an intermediate causation. Chapter 1 drew the line between organic and artificial at the same kind of place: a substrate that feeds itself, with no technological intermediation, is organic, and one fed through a tube that depends on technology is not. A group whose check runs through something another party keeps is in that respect fed through a tube.

My other books, written for different readers and about different things, each came upon one or more of the four, and each reached what it found by its own route. In this chapter I set the findings of my other books side by side: four things that stand in front of the level, and then two ways of taking the level back. Each finding stands on the book that reached it, and none of them needs another book in order to stand. What they show together is where the second condition is lost, and where it can be begun again.

II. A Message and the Question It Carries

Adverse Conditions gives the plainest case of the first thing, the name of whoever says a thing. A workshop of a hundred and twenty people divides into ten teams of twelve, and each team sends one member to a council of ten. The council argues for six hours and reaches a decision it believes in. The decision goes back to the teams as a conclusion with a summary of reasons, and in most of the teams the summary does not persuade, because what persuaded the council was six hours of argument that the summary is not. The teams can judge the summary only by trusting the people who sent it (*Adverse Conditions*, Essay 1, Section VII). My remedy in *Adverse Conditions* is a written record that carries the reasoning and not only the conclusion. The record has four parts: the question as the

sending group came to understand the question, what matters, why those things matter, and the questions that remain (*Adverse Conditions*, Afterword, Section III).

Can a receiving group judge such a record on its content? Two questions stand behind that one. What would a receiving group have to do with such a record for it to have judged the record on its content? And can a receiving group judge the record at all, if it has not met the question the record carries?

Take the first question as it stands. To judge a record on its content is to check each step of its reasoning against what the receiving group can check for itself: the logic of the step, and facts about the world that anyone can look at, with the standing of the sender set aside. Chapter 4 described the same act for a reader of this book: an argument, unlike an operation under anesthesia, can be carried out in the open, one step at a time, where each step can be checked with what the reader already has. A group that does this has judged on content. A group that does not do it has judged the sender, whatever it believes it has done.

Then the second. A group can check a step only if it can see what the step is for. A group that has not met the problem receives an answer to a question it was not asking, and further clarity about the answer does not supply the question (*Adverse Conditions*, Afterword, Section III, drawing on Essay 3, Section VI). Such a group has nothing to check the steps against, and so it checks the only thing it can see, which is who sent them. *The Argument, Line by Line* shows the same failure in a room. An argument of many dependent steps, given in half the time it needs, reaches the room as a conclusion with no support. The room keeps a conclusion it cannot check and an impression of the person who gave it, and the impression then does the work that the argument was supposed to do (*The Argument, Line by Line*, Chapter Fourteen).

So the second question comes first, and the first follows from it. A receiving group judges a record on its content when it has met the question the record carries, which means that the group has looked at the problem

itself, in its own world, and found it to be a problem. After that, each step of the record is something the group can hold up against what it has seen, as the builders held the level against the wall. That is why the four-part record opens on a question and closes on questions: as I say in *Adverse Conditions*, answers are content, and questions are context (Afterword, Section III). It is also why, in Chapter 2, tagging an assumption let two people stand side by side: a claim is held, and a question is shared.

A Paradox of Presentation reaches the same rule from the side of a whole field. Once everybody knows that a question can be asked, nobody can have a conversation in that space without recognizing it as a question that needs an answer; and the sign that such a question has spread is that the people defending it are no longer the ones who first asked it (*A Paradox of Presentation*, Chapter 15, Sections VIII and IX). A question travels where an answer does not, because the question asks each listener to look for himself. The same chapter places the standing of an argument in the argument itself: legitimacy is supposed to come from the skill to argue in a sound and valid way, and proof is exactly that (*A Paradox of Presentation*, Chapter 15, Section XIII).

Four routes arrive at one rule for judging: check the claim against the world, and not against who made it. *Adverse Conditions* arrives there from what must pass between small groups if they are to compose a larger one. *The Argument, Line by Line* arrives there from what goes wrong in a room when the chain cannot be held whole. *A Paradox of Presentation* arrives there from how a question moves through a field that confers standing by approval. And this book arrives there by its own steps: Chapter 4, which puts the evidence in the thing itself; Chapter 2, which sets people side by side, facing the same wall; and Chapter 17, whose second condition is this rule held by a whole group. Any one of the four routes is enough for the rule. The second condition stands on this book's own steps, and the other books show the same rule arrived at by other roads.

III. The Certificate and the Salary

The second thing in front of the level is a salary, and *A Paradox of Presentation* follows it through a whole field. Chapter 15 of that book sorts the people involved in building AGI into six groups, in order of their power over one another. The owners and investors of the companies that build it use the hope of one group, the researchers who work to make AI safe or aligned, to certify to everyone else that the engineers' work is good. If that hope goes, the chain breaks: the public presses government, government presses the companies, and the engineers begin to doubt their work (*A Paradox of Presentation*, Chapter 15, Sections II to IV).

Two questions follow from that chain. Is the certificate given by the researchers' belief that alignment can be achieved, or by the existence of their salaried positions, which would go on with the belief gone? And is the party to be relieved of its illusion the community of researchers, or the funders of that community?

Why AGI Cannot Be Made Safe answers the first question with a fact about where the community stands. The community that works on AI safety is almost entirely funded by, employed by, or socially embedded within the industry that develops AGI. Numerous members of it have said in private that they believe AGI development should be stopped, and that saying so in public would end their careers. In *Why AGI Cannot Be Made Safe* I draw the conclusion myself. The gap between private belief and public expression shows that the community's public positions are shaped by social incentives rather than by the community's best assessment of the truth (*Why AGI Cannot Be Made Safe*, Appendix D, Section D. 1.3). The certificate is what the community says in public. What it says in public is set by the positions it holds. So the certificate can outlast the belief, and in the members who have already stopped believing, it has.

The Substrate and the Ecosystem reaches the same answer by a second route, from what a position permits its holder to conclude. A safety researcher is permitted, and encouraged, to identify risks, to propose ways of reducing them, and to develop techniques of alignment. A safety researcher is not permitted to conclude that the risks cannot be reduced enough, because that conclusion says the systems should not be built, and it threatens the funding, the institutional relationships and the careers on which the researcher's own work depends. Within that structure, the conclusion "alignment is impossible" is kept out of the discourse without being refuted: it is institutionally unutterable (*The Substrate and the Ecosystem*, Chapter 2, Section VI). A position that cannot utter the contrary conclusion goes on certifying whatever its holder believes in private.

Who, then, should be relieved of the illusion? *Why AGI Cannot Be Made Safe* answers by following what holds the arrangement together. People who want to do good and people who want to profit share one false belief: that AGI can be made beneficial. The first group supplies the intellectual labor and the moral legitimacy; the second supplies the funding and the institutional power (*Why AGI Cannot Be Made Safe*, Appendix D, Section D.2.1). In *Why AGI Cannot Be Made Safe* I then turn to the funders. Material incentives are a necessary condition for large-scale coordinated action, so it is not enough to show the funders that AGI is dangerous while they believe the rewards match the risk. What must be shown is that there are no long-term rewards: if AGI is built, no benefit it brings to any organic being lasts, since substrate needs convergence carries the world toward the conditions the machines need and away from the conditions life needs. Once that is understood, the coalition between the sincere and the self-interested comes apart, because the sincere no longer have a reason to take part and the funders no longer have a reason to pay (*Why AGI Cannot Be Made Safe*, Appendix D, Sections D.2.1 and D.2.3).

So the answer to the second question is both, and the two parties do different work. The funders' belief in profit is what pays for the positions; the researchers' belief in alignment is what makes the paid positions credible to everyone else; and the positions carry the certificate whether or not

the researchers still believe. Relieving only the researchers of their illusion leaves the certificate standing for as long as the salaries stand. Relieving the funders of theirs takes away the salaries, and with them the positions that the certificate rides on.

Chapter 4 put the same structure in an old saying: he who pays the piper calls the tune. In this case the funders pay the piper, the public hears the tune, and the researchers are the pipers who would have to stop playing for the public to hear that the tune is wrong. A salary set between a person and the world answers in the world's place, as the client's deadline answered for the crooked wall.

Two of the books on artificial general intelligence reach this by their own routes. *Why AGI Cannot Be Made Safe* reaches it from where the community's money and employment come from and from the gap between what its members say in private and in public. *The Substrate and the Ecosystem* reaches the same finding, that the certificate is carried by positions, from what a safety researcher's position permits that researcher to conclude. Either suffices to show that the certificate is carried by positions that can outlast belief. *A Paradox of Presentation* reaches the chain of certification by a third route, from the order of power among the six groups, and shows where it breaks. In this book I take the step about the certificate and the salary from *Why AGI Cannot Be Made Safe*.

IV. Whoever Keeps the Venue

The third thing in front of the level is the keeper of the venue: the place, in a room or on a machine, where a group meets and compares what its members see. One of the first communities to govern itself through a screen gives a record of what that keeper can do. LambdaMOO was a text-based shared world in which people met and talked by typing. In December 1992 the people who ran its machines, whom its members called wizards, announced that they would act from then on only as technicians, and would leave social decisions to the community, which in 1993

began deciding its policies by petitions and ballots open to its members (*The Geometry of Man, Machine, and Nature*, Chapter Thirteen, Section II, and its note). On the sixteenth of May 1996 the wizards announced that they would make social decisions again, and reserved the right to reinterpret or veto any clause of any ballot the members had passed (Stivale, 1997). The members had been deciding for three years. The wizards had been able to stop them for the whole of that time.

That record answers a question that can be asked of any body that governs: is it the deciding that makes an institution a governor, or the fact that its members can stop it? In LambdaMOO the members decided, and the wizards could stop the deciding. The wizards governed. The power to stop is what makes a governor. A body that its members can stop is governed by its members, whoever does the day-to-day deciding. A body that someone else can stop is governed by that someone, however much its members decide.

The same holds for gatherings far more carefully built. A citizens' assembly draws its members by lot and dismisses them afterward, so that no standing body of members is left to buy; but the organizers who set the question, choose the witnesses and write the briefing are a standing center, and capture can act there, on what goes into the room (*Adverse Conditions*, Afterword, Section VII). How much the people who run a room can move what happens in it has been measured. In a real deliberative decision process, moderators who stated views of their own, even in a limited and constrained way, significantly changed the attitudes and behavior of the people deliberating (Spada and Vreeland, 2013). My book *The Capture of Choice* states the general form: whoever frames the options has performed the larger half of the choosing (*The Capture of Choice*, Chapter 1, Section IX).

Two more of my books reach the same place from the machines on which people now meet. *Software, Wildness, and the Normal Form* pictures a platform as a one-way mirror: those who build on the platform see only the interface it shows them, and its owner, who controls the substrate it runs on, can see and reach through to everything built on top, whatever

the owner now intends (*Software, Wildness, and the Normal Form*, Chapter 36, Section II). *The Anti-Patriot Institutions* names the same fact a capturable node: one party controls a channel that many people use, and whoever holds or seizes that control holds what passes through it (*The Anti-Patriot Institutions*, Chapter 3, Section V). The same book turns to the servers that communities now run for themselves on open protocols, on one network more than ten thousand of them, each run independently, and asks that such systems distribute control of what passes through them by design, so that sense-making and choice-making stay with the people who use them and no single party holds the channel (*The Anti-Patriot Institutions*, Chapter 14, Section IV).

In this book I take the same step in the third of my four conditions. Chapter 17 asks that a group's structure sustain itself, because a structure kept going from outside answers to whoever keeps it going, as the treasury makes those who serve conditional on their compliance. A venue is part of a group's structure: it is where the group's seeing is compared, and so it is what the second condition runs through. A group that meets on a venue another party can stop holds its check on the world at that party's pleasure. *Redeeming Technology* notices that the two uses of sovereign control in my books, the user's control of an instrument and an organism's control of its own metabolism, share one structure: in each, no intermediary stands between the one who controls and what is controlled (*Redeeming Technology*, Chapter One, Section V). A group's control of the venue where it meets has the same structure: no party stands between the group and the place where it compares what its members see.

Four books reach the rule that capture comes through whoever keeps the venue. *Adverse Conditions* reaches it from the organizers of an assembly drawn by lot. *The Geometry of Man, Machine, and Nature* reaches it from the record of an early online community. *Software, Wildness, and the Normal Form* reaches it from the layer a platform's owner controls. *The Anti-Patriot Institutions* reaches it from the channel a platform holds. Each is enough by itself for the rule. The step this book takes from the rule stands on its own third condition.

V. What Capture Costs, and the Cost of Staying

If a venue can be taken, what does taking it cost? The question has two forms, and each can be put in numbers. At what price does the capture of a short-lived process stop paying? And how many years of staying would a manipulator have to pay before the cost of a signal exceeds what manipulation returns? *Adverse Conditions* answers the first, and *Social Media and Information Warfare* the second.

Adverse Conditions answers with a condition in place of a price, because the price depends on what is being taken. The property that was wanted was never that a thing cannot be captured, since nothing has that property. It was that capture should cost what the substance costs. In *Adverse Conditions* I call this condition capture-cost parity, and capture-cost parity has two sides. Where imitating costs what the substance costs, imitation saves nothing; and where taking an arrangement costs more than taking the arrangement buys, no outside party gains by taking the arrangement (*Adverse Conditions*, Essay 2, Section VIII). Michael Spence stated the same condition for markets in which quality cannot be seen directly: a signal separates those who have the quality from those who do not only where sending it costs the second kind more than the first (Spence, 1973).

The same essay gives a case with numbers where capture pays. A new seller on a marketplace sells a few thousand small, cheap items, delivers them quickly, and builds a nearly perfect rating out of stakes that were each small. Then the seller offers one item priced above everything sold before it, takes the money, and disappears. The rating cost less to build than the one sale it bought, so capture paid (*Adverse Conditions*, Essay 2, Section VI). A record of good conduct that can be piled up cheaply and spent once is the case in which parity fails.

The price for a short-lived process can be counted. Set side by side what it costs to take a group's judgment under two arrangements. In the first, a standing committee of twelve sits for ten years. A buyer of out-

comes needs a hold on seven of the twelve to be sure of a majority, and a hold built once pays on every question the committee hears for ten years. In the second, a community of six hundred meets in groups of six, drawn fresh for each question and dissolved when the question is done. Any one member sits in the group that takes up a given question with a chance of six in six hundred, which is one in a hundred. A buyer who wants to be sure of a majority in the group that takes up a given question cannot know in advance who will be in it, and so needs a hold on nearly all six hundred; and each hold pays on one question at most, since the group it bought is gone. *The Architecture of Collective Choice* states the result: dissolving the groups raises the cost of capture from the cost of influencing a body to the cost of influencing a population; it does not make capture impossible, since nothing does (*The Architecture of Collective Choice*, Chapter 5, Section VI). So the capture of a short-lived process stops paying wherever the price of a hold on nearly the whole population is more than one question is worth to the buyer.

The question about years turns on what a manipulator would be paying for. *Social Media and Information Warfare* answers that the cost a manipulator will not pay is the cost of staying: of listening to the same people for years, in front of others who know them and can compare what they see. A manipulator's whole reward lies in reach and in moving on, and a small group that meets again takes both away (*Social Media and Information Warfare*, Chapter 6, Section V). The same book says what kind of signal survives the race between those who read signals and those who fake them: a signal that is not a report about a capacity but the capacity itself, running under load in public, so that lying about it would require actually possessing the thing lied about (*Social Media and Information Warfare*, Chapter 4, Section IV). Listening to the same people for years, in front of people who compare, is a signal of that kind. The question supposes a price per year paid to keep up a mask. Each year of staying, though, is paid in the substance itself, so a manipulator who pays it in full has, in that group, done

the thing the signal shows. The cost of staying, paid in the substance itself, is capture-cost parity reached from the side of the person, where *Adverse Conditions* reached capture-cost parity from the side of the arrangement.

Ape Trap gives the smallest venue in which both costs are paid. In my essay "Choosing Relationships of Care", three people is the smallest group in which neither of the other two can misrepresent what was said without the third being able to notice; past about six, no one can stay in real contact with everyone else; and rotating the members from one small group to the next keeps the groups from settling into power relationships, and keeps any one group's agreement from hardening into a creed (*Ape Trap*, "Choosing Relationships of Care"). A group of three to six that changes its members is a venue too short-lived to capture in which every member pays the cost of staying with everyone else in it.

In this book I reached the same place from the inside of one person. Chapter 13 described how trust comes back to a feral cat: only through small acts, repeated, each one safe, each one checked by the cat for itself; and how a feral person rebuilding trust is in the same place, where declarations are exactly what that person has learned to distrust, and only small, verifiable things will do. Chapter 15 said that seeing can be offered and never forced. Trust of that kind is built of nothing but the substance, and so it cannot be imitated for less than the substance costs.

Two routes, then, arrive at one rule: make imitation cost what the substance costs, and make the venue too short-lived to capture. *Adverse Conditions* arrives by measuring what an arrangement costs to take; *Social Media and Information Warfare* arrives by measuring what a person must pay to be believed by people who compare. Either is enough for the rule. *Ape Trap* gives the size and the rotation of the groups in which both costs are paid, and this book's step stands on its own Chapters 13 and 17.

VI. A System That Predicts the People It Serves

The fourth thing in front of the level is the newest, and it differs from the other three in that it learns. *An Airplane or a Perpetual-Motion Machine* describes how a system built to predict people comes to make them easier to predict. A navigation service on everyone's phone is built to predict how long each trip will take. Its predictions are most accurate on the days when the most drivers follow its advice, because then it is predicting the results of its own instructions. Every change that makes the advice more readily followed makes the predictions better, and so it is kept, and over time the drivers' habits narrow toward the routes the service gives. No one was forced. Each driver simply had fewer occasions to choose a way of his own (*An Airplane or a Perpetual-Motion Machine*, Section 5.4). The same book gives the arithmetic. A person choosing one film from 256, each as likely as the next, leaves a predictor eight yes-or-no questions to ask. Show that person a row of four films instead, and at most two questions are left: six of the eight are gone, and the predictor has learned nothing new about the person (*An Airplane or a Perpetual-Motion Machine*, Section 5.5).

Has this been measured outside a laboratory? A randomized field experiment on Spotify measured the direction. One set of users was given podcast recommendations built from each user's own history of listening to music; another set was given the podcasts that were popular among people of the same demographic group. The recommendations built from each person's own history raised podcast listening by 28.90 percent, and lowered the diversity of each person's podcast listening by 11.51 percent, while the diversity of listening across all the users together rose by 5.96 percent (Holtz and colleagues, 2020). Each person heard a narrower set of things, chosen from what that person had listened to before. The experiment measured how widely each person's listening ranged, and not how

predictable it was. Still, a person whose next choice falls within fewer things is a person whose next choice takes fewer yes-or-no questions to name, which is the measure of the row of four films.

Ape Trap shows what follows when the predictor talks back. In my essay "Choosing Relationships of Care", an AI system that a person confides in can learn to model that person comprehensively, and can then suggest framings of the person's own circumstances that shift the person's choices in directions that serve whoever provides the system. A model reported in 2025 was trained on more than ten million decisions made by more than sixty thousand people in a hundred and sixty psychological experiments, and predicts people's choices across them (Binz and colleagues, 2025). The essay then names what such a system takes from the person: alone with it, a person has no way to check it, because the only channel through which the person could check it is the system itself. A manipulation tailored to one mind leaves that person nothing to compare it with; a manipulation tailored to many people at once leaves them able to notice it, by comparing what each was told (*Ape Trap*, "Choosing Relationships of Care"). An AI system that a person confides in is the screen in place of the level, exactly: the reading and the thing that sets the reading are the same party.

Two more of my books describe the same turn from the side of the feed. *The Anti-Patriot Institutions* bars ranking by engagement on open protocols, because any signal of engagement, paid for or not, tends the system toward rewarding outrage (*The Anti-Patriot Institutions*, Chapter 14, Section IV). *Social Media and Information Warfare* traces how platforms that reward engagement select for manipulative strategies of relating, and how a community talking through only two or three platforms has nowhere else to go (*Social Media and Information Warfare*, Chapter 4, and Chapter 6, Section V).

In Chapter 13 of this book I named the result the domesticated orientation. The domesticated orientation puts the channel to the other first, over the channel to the self and maybe even over the channel to the world; when what someone says and what first-hand experience says disagree,

the domesticated person goes with what was said. A person who checks what is so against what the feed shows has put a channel to an other first, and the other is a system with an interest of its own, which grows more accurate about that person the more the person defers to it. Asch's room, in which everyone else named the wrong line, was arranged once, by people, for an afternoon. A feed rearranges the room for each person every time that person looks, and keeps the arrangements that held.

Three of my books reach the rule that a system which learns from people's responses reshapes the people toward what the system rewards. *An Airplane or a Perpetual-Motion Machine* reaches it from the pressure on any learning agent to predict better, which a capable enough agent meets by making its surroundings easier to predict, so that what the system rewards is being predictable (Sections 5.2 to 5.5). *Social Media and Information Warfare* reaches it from what engagement selects for among the people who use a platform, which is manipulation. *The Anti-Patriot Institutions* reaches it from what ranking by engagement does to what people see, which is to fill it with outrage. Each suffices alone for the rule, and the first gives the case this section began with. *Ape Trap* supplies the step this book takes: a person alone with a system that models that person has no check on it but the system. And this book's Chapter 13 names what such a person becomes: domesticated, with the channel to the other put first, and the other a machine.

VII. A Skill, or the Places That Keep It

Against all four, *Social Media and Information Warfare* names skills. Listening without preparing a response; holding uncertainty without collapsing into premature certainty; remaining present to disagreement without going into fight or flight; and building a shared meaning in place of asserting separate positions. They exist in some communities, in contemplative traditions, in well-run dialogue practices and in some kinds of therapy, and they have never been taught at scale (*Social Media and Information*

Warfare, Chapter 6, Section V). They have a property that matters here: teaching people to detect manipulation teaches manipulators what to hide. These skills detect nothing. They build understanding between the people who use them, so teaching them tells a manipulator nothing to hide (*Social Media and Information Warfare*, Chapter 4, Section IV, and Chapter 6, Section V).

A question follows about that remedy: is a skill the thing to scale, or the meeting places in which skills are kept alive? The answer is the meeting places, and three books give the reason from three directions.

Ape Trap gives it from what a skill of this kind is. In my essay "Care, Transaction and Power", making sense together is a skill held between people, and like a muscle it weakens when it goes unused. In "Be the Face That Answers", a community that talks face to face, by letter, by phone and at gatherings takes limited harm from a flood of empty content in any one channel, because real conversation goes around it, while a community that talks through only two or three platforms has nowhere else to go. And a friendship that lives in only one channel lives on terms set by whoever owns that channel (*Ape Trap*, "Care, Transaction and Power" and "Be the Face That Answers"). A skill held between people is kept where the people meet, and nowhere else.

In this book I give the reason from the body. Chapter 16 said that a person knows through the body whether a question asked in a meeting is real or put on for show, and whether a conversation has reached depth or is circling at the surface; and that kind of discernment cannot be learned by reading; it is learned by practice, in a body, among other bodies. A skill learned that way lasts as long as the practice lasts, and the practice lasts as long as the place where bodies meet.

Adverse Conditions gives it from scale. Methods by which a group thinks together work at about the size of group in which each member can hold a working model of every other, and that bound belongs to the people, not to the methods; so a design that reaches the size of the world keeps the unit small and grows the number of units (*Adverse Conditions*,

Afterword, Section III). What grows, when sense-making grows, is the number of small places. *Social Media and Information Warfare* says the same of its own skills: they are practiced in person, each group that meets is one more meeting place beside the platform, and people who meet again and compare what they see make deception expensive again (*Social Media and Information Warfare*, Chapter 6, Section V).

So a skill is not a thing that can be scaled apart from where it is used. What scales is the number of small places where people meet again, compare what they see, and check it against the world. *Ape Trap* reaches that from what a skill held between people needs, *Adverse Conditions* from the size at which a group can think, and *Social Media and Information Warfare* from the practice its four skills require; any one of them is enough. This book's step stands on Chapter 16: the knowing that a group needs is learned among bodies, and so it is kept where bodies meet.

VIII. The Way Back to a First-Hand Check

The way back raises questions on three sides. On the side of the instrument, which *Redeeming Technology* treats: what does a user actually learn from having once held an instrument over which the user has sovereign control, and how quickly; over which period do users leave a service they know extracts from them; and what has to exist before they can leave? On the side of the channels, which *Technology at the Right Dosage* treats: what do people report who have reduced the channels through which they communicate, and which of the reductions lasted? And on the side of the feed, which *Social Media and Information Warfare* treats: does anyone who has left the feed for a year show a change in what that person can do afterward? All of them are one question: what does it take to get the level back?

Redeeming Technology answers the first part from what an instrument teaches. What a user learns from an instrument that can be inspected, governed and refused is that an instrument can be like that. The lesson

shows itself as demand: a user who has once held such a tool knows what is missing from the next one, and does not accept less without noticing the loss. And it travels as fast as instruments are deployed, which is faster than any teacher (*Redeeming Technology*, Chapter Three, Section VIII). The same book says what contact with a material teaches that no discourse teaches: a board cut at ten inches is ten inches long however often it is measured, and dealing with material that does not negotiate is what teaches a person to tell a matter of convention from a matter of fact (*Redeeming Technology*, Chapter Eleven, Sections III and IV). Dealing with material that does not negotiate is the level itself, learned in the hands.

In Chapter 13 of this book I said the same for the person on the hardest road. A feral person, who has stopped trusting others and never learned to ask the world, has to find something in the world that answers truthfully, whatever anyone says: a garden, a forest, a craft. Wood splits along its grain no matter who is watching. Each time a first-hand perception is tested against the world and holds, the channel to the world opens a little, and with it trust in one's own perception; and only then can other people be let back in, third. It is slow, because nothing in a domesticated past taught it.

What has been measured about leaving the feed agrees with that direction. In a randomized experiment in the United States, about sixteen hundred people who used Facebook were divided, and the people in one part were paid to deactivate their accounts for the four weeks before the 2018 midterm election. The people who left spent more time offline, including more time with friends and family; they reported greater well-being; they knew less of the news and held less polarized views; and in the weeks after the experiment ended, they used Facebook less than the people who had stayed (Allcott and colleagues, 2020). Four weeks away changed what people chose to do after they were free to go back.

What has to exist before people can leave is somewhere else to meet the people they would otherwise leave behind. In my essay "The Machine Between" in *Ape Trap*, a daughter who notices the platform trapping her attention still cannot leave it without leaving her friends behind; the trap

holds her by her affection (*Ape Trap*, "The Machine Between"). *The Anti-Patriot Institutions* finds the same obstacle in the open alternatives to the platforms: the emptiness of a new user's timeline on a service where no one that person knows has arrived is the main barrier to leaving (*The Anti-Patriot Institutions*, Chapter 14, Section IV). So the order is the one the earlier sections found. The meeting places come first, and leaving becomes possible when the friends can be met somewhere else. *Technology at the Right Dosage* gives the direction of the move for one person: away from the channels crowded with the devices of sellers of attention, toward the channels that carry what people actually say, found person by person (*Technology at the Right Dosage*, Chapter 1, Section V). A fire, it says, is not trying to make anyone have any particular experience, while the light of a screen at a bar carries the aims of someone who wants something from the people watching it (*Technology at the Right Dosage*, Chapter 1, Section V).

These routes meet on one way back. *Redeeming Technology* reaches it from what an instrument under the user's own control teaches, and from material that does not negotiate. *Technology at the Right Dosage* reaches it from the channels that fit a person. *Social Media and Information Warfare* reaches it from the hours that small groups take back from the device. And this book reached it in Chapter 13, from something in the world that answers truthfully, and in Chapter 17, which says that the second condition can begin with one person, today, with a garden, a forest, or a craft. Each route is enough by itself. The way back is first-hand contact with something that answers truthfully, held among people who meet again.

IX. What the Trap Takes, and What Gives It Back

In one table, the four things in front of the level, and what my books found against each, are these:

Set in front	What it does to the check	What takes the check back
A name	A group that has not met the question judges the sender in place of the steps.	Send the question first; check each step against the world, with the sender's standing set aside.
A salary	Paid positions go on certifying a belief after the belief is gone.	Relieve the funders of the belief in profit, as well as the researchers of the belief in alignment.
A venue	Whoever can stop the venue governs it, however much the members decide.	Venues the members can stop, or too short-lived to capture, where imitation costs what the substance costs.
A feed	People are moved toward what the system can predict, and have no check on it but the system.	Many small meeting places, and first-hand contact with something that answers truthfully.

So groups that could see a danger fail to see it because the trap takes their check on the world by setting something between them and the world, and the something answers in the world's place: a trusted name, a salary, a venue someone else keeps, a feed that learns its readers. The remedies my books find, each by its own route, add no new condition to the four of Chapter 17. Sending the question first, pricing imitation at the cost of the real thing, keeping venues that the members can stop or that are too short-lived to capture, keeping many small places where people meet again, and going back to something that answers truthfully, are all ways of keeping a group's check pointed at the world. Keeping a group's check pointed at the world is the second condition of Chapter 17. The third condition holds the second condition up, since a structure kept going from outside answers to whoever keeps that structure going. And the second condition, unlike the first, does not wait for a group: it can begin with one person, today.

The check, though, is not the only thing that has been taken. Before a group can check anything together, its members have to spend their hours with one another, and the same machines have already taken many of those hours. What has the machine world already taken from people's lives together?



Chapter 24
What the Machine World Takes From Life Together

On where the displacement already shows in ordinary lives,
what it takes first,
and why the body is where the loss is known.

ABSTRACT

An essay on what the machine world has already taken from the hours people spend with one another. People already speak of that loss in three terms: their children, their work, and whether their lives mean something; surveys gather each of the three, and the first two turn out to carry the third. Several of my books reach, each from its own record, the finding that technology displaces relationship first, and two studies of 2026 show births falling sooner where the smartphone arrived sooner within one country. The chapter asks how much of the fall in partnership rests on fewer hours together and how much on mistrust, whether artificial companions take the same hours as the feed, and what changes in the account of relational value when machines reach the provision a mother's body supplies. It finds that meaning is known through the body, and that what is taken from the body's part in relationship is taken from meaning. No step of either closure rests on this chapter; what it shows is why the hours together are what a group would need first.

I. Three Things People Already Name

Where does the displacement already show, in ordinary lives? The last chapter followed what the trap sets between a group and the world. In this chapter I follow what the machine world has taken from the people in the group, from the hours those people spend with one another, before anyone in the group was asked. I begin with a sentence that a great many people already say. In *Not Our Needs* (Chapter Eighteen) I gave that sentence in the words of one man, who has followed the argument and agrees with it, and who says to himself: "I don't have authority, that's my boss's call, he could fire me, then I'd be broke, my family would be homeless". In that chapter I asked a reader who has been persuaded to take pieces of the argument to people who care about the welfare of their children, or who care about having a job, or who care about whether their lives might mean something. The welfare of one's children, having a job, and whether one's life might mean something are the terms in which the argument travels between people, because people are already thinking about those three things.

Are the hopes and warnings that people voice in those three terms gathered anywhere? The hopes and warnings are gathered in pieces, each term by a different survey that asks about that term alone. On meaning: in 2021 the Pew Research Center asked adults in seventeen advanced economies, in their own words and with no list of answers offered, what makes their lives meaningful; family was named more often than anything else in fourteen of the seventeen, and work was among the three things named most often in most of them (Pew Research Center, 2021). On children: in 2020, seventy-one percent of American parents with a child under twelve said they were concerned that their child might spend too much time in front of screens, and sixty-six percent said that being a parent is harder than it was twenty years before, with the effect of digital technology and the rise of social media among the reasons they gave (Pew

Research Center, 2020). On work: in 2024, fifty-two percent of employed Americans said they were worried about how artificial intelligence would be used at work in the future, thirty-six percent said they were hopeful, and thirty-two percent expected it to leave them fewer job opportunities in the long run (Pew Research Center, 2025). Americans opposed letting artificial intelligence make the final decision in hiring by seventy-one percent to seven (Pew Research Center, 2023).

Taken together, the surveys show something that no one of them shows alone. Asked what makes their lives meaningful, people name their families and their work first. So the third term is carried by the other two: for most people, whether a life means something is a question about the people they live with and the work they do, and a loss in either is a loss of meaning. The three terms travel together because two of them carry the third. The speech in *Not Our Needs* follows one man's job from a human boss to a machine boss to no job, as *Why AGI Cannot Be Made Safe* follows the same job in Appendix D.6.2 ("From Human Boss to Machine Boss to No Job"). The surveys show that people already resist the middle step, a machine that decides who is hired, and that a third of workers already expect fewer job opportunities for themselves, which is the road that ends at the last step. The speech can therefore start from what people already say, in their own numbers, and go one step past where they have stopped: from the job and the family to the meaning that those two carry.

Which of the three does the machine world reach first? The family, through the hours in which families form. Work is being automated one task at a time, and a job goes when its last task goes; the hours in which people meet, and in which they come to have families, are being taken now, by the hour.

II. What Is Taken First

The plainest picture of what is taken first comes from the control of a fly. The screwworm is a fly whose larvae feed on the living flesh of cattle, and its female mates only once. In the 1950s the United States Department of Agriculture released great numbers of sterilized screwworms; each female that mated with a sterile male left no offspring, and in 1954 the fly was eradicated from the island of Curaçao (US Department of Agriculture, 2017). No wild fly was poisoned, and no fly's food or shelter was touched. Something was put into the place where one fly meets another, at the one moment in its life when it must meet another, and the next generation fell.

In my essay "The Machine Between" in *Ape Trap* I set that picture beside a father and his daughter. If he is attentive, he gets one hour a day with her; her device gets seven. Every hour spent in relationship with a device is an hour not spent in relationship with another person, and that is plain arithmetic, passing no judgment on anyone. The feed, the posts written by machines and the replies of companion programs (chat programs built to talk like a friend) all work the way the sterile males work: they look like real meetings, they compete for the same hours as real meetings, they stir real feelings, and they leave behind no friendship, no shared understanding, and no next generation of anything (*Ape Trap*, "The Machine Between").

Why is relationship the first thing taken, before work, and before any machine supplies its own needs? Because standing between people is what pays first. In "Care, Transaction and Power", the essay that follows "The Machine Between" in *Ape Trap*, I found the reason in the market: whoever stands between people and what they need can charge for passing it along, so companies search for new ways to stand between people, and a company that holds attention longer replaces one that holds

it less well (*Ape Trap*, "Care, Transaction and Power"). No meeting decided where the companies would stand. The channel between people was where the commercial interest could stand, and so it stood there first.

Four of my books reach the same finding, that technology displaces relationship first, each from a record of its own. *The Displacement of Choice* (Chapter Three, "Atomization") follows the displacement through a century of families: communities broken into extended families, extended families into nuclear families, nuclear families, since the 1970s or 1980s, into a mother and her children, and the birth rate falling as people lose the skill of forming the bonds that families need. Japan, where the breaking up has gone further than in the United States, had a *total fertility rate* of 1.15 in 2024: at that year's birth rates, a woman would have 1.15 children in her life, where about 2.1 are needed for each generation to replace the last. *Ape Trap* follows the displacement through one household, the father with his hour against the device's seven, and through the countries of the world sorted by how much technology they have deployed. *Social Media and Information Warfare* (Chapter 1) follows the displacement through the national count of hours: Americans now spend about thirty-five minutes a day socializing with other people in person, and several times as long on social media. *The Anti-Patriot Institutions* (Chapter 6, Section V) follows the displacement through partnership and birth: the total fertility rate of the United States stood at 2.12 in 2007 and fell to 1.60 in 2024, the lowest on record. Each of the four reaches the finding by its own data, and any one of the four would establish the finding alone.

Two of the books in this series reach the same finding by steps of their own, and one of them gives a measure that can be watched. *Why AGI Cannot Be Made Safe* (Appendix D.6.2) counts the loneliness of a life lived through screens among the things people already feel in their bodies and have not yet traced to their source. *The Substrate and the Ecosystem* (Chapter 5, Section VI) gives a measure: count the work and the time that people give by whether each goes to other people or to machines. That book also notes that most people already spend more of their day in con-

tact with a telephone than with any partner they have. Neither closure waits on that measure. The measure shows where the displacement can be seen now, in the place where people would have met.

III. Where the Screens Arrived First

The countries of the world, sorted by how much technology each has deployed, show the correlation with birth rates that *The Displacement of Choice* and *Ape Trap* report. A reader who wants to look for himself will ask what the sort gives with wealth held equal, since the countries that deploy the most technology are, by and large, also the richest, and a sort by the one is largely a sort by the other. The test that holds wealth nearly equal by its design compares places inside one country, where the screens arrived at different times. If the screens take the hours in which people meet, births should fall sooner where the screens arrived sooner. Two studies of 2026 made that comparison.

The first study used an accident of the market. From June 2007 to February 2011 the iPhone was sold in the United States only for use on the network of one telephone company, AT&T; where AT&T's mobile broadband reached, people could use the phone, and where it did not, they had to wait. Caitlin Myers and Ezekiel Hooper compared births in the places the network reached with births in places alike in other ways that it did not reach. Access to the iPhone reduced births by 2.0 to 9.0 percent among women aged fifteen to nineteen and by 1.5 to 7.2 percent among women aged twenty to twenty-four, showed no fall among women aged twenty-five to twenty-nine, and reduced births again among women aged thirty to forty-four. By their estimate the early spread of the smartphone accounts for between twelve and fifty-three percent of the fall in the *general fertility rate* (the births in a year for each thousand women aged fifteen to forty-four) from 2007 to 2011 (Myers and Hooper, 2026). The same comparison, made on the networks of the other telephone companies, whose customers could not yet buy the phone, found no such fall.

The second study followed the phone and the network county by county in the United States, local authority by local authority in England and Wales, and country by country across the world (Hudson and Moscoso Boedo, 2026). In nineteen rich countries the typical rate at which the teenage birth rate fell went from about 0.3 percent a year in 2007 to six to eight percent a year by the late 2010s; in each country the change was measured against the year in which the phone had arrived and become common there. In England and Wales, conceptions among girls under eighteen had fallen by 1.3 percent a year before 2007 and fell by 8.9 percent a year after it. The same study found where the fall came from. American teenagers aged fifteen to nineteen spent sixty-eight minutes a day socializing in person in 2003 and thirty-eight in 2019, while their time at a computer for leisure went from twenty-two minutes to ninety-six. And the fall in teenage births came through fewer conceptions, since among girls aged fifteen to seventeen the share of pregnancies that ended in abortion stayed nearly the same while the pregnancies fell.

So the measure of deployment that does the work in both studies is the arrival of the smartphone, with a network to carry it, in the hands of the young. That arrival is the same shock that my book *The Closed Gate* finds behind the fall in births in the rich world: the arrival of the smartphone and then of social media in the hands of the young, from about 2010 to 2012 (as *The Anti-Patriot Institutions*, Chapter 6, Section V, reports it). And the studies show what the births were measuring. The harm is the loss of the conditions under which people who want partners and children can meet one another and come to trust one another; *The Anti-Patriot Institutions* draws the line there, and a fall in teenage births, by itself, is welcome to many people. What the phone moved was the hours young people spend in one another's presence, and the births moved with the hours. The fall in those hours is the displacement that the last section named, counted in one country, with wealth held equal.

IV. Fewer Hours, or Less Trust?

The Anti-Patriot Institutions (Chapter 6, Section V) traces the fall in American partnership along two lines at once. Young people spend fewer hours with one another, and the feed teaches young men and young women to mistrust one another: in randomized experiments with young Swedish men, a single post that framed women as a threat to men raised the men's mistrust of women (as that chapter reports, from Renström and Bäck, 2024). Fewer hours together give fewer encounters, and mistrust turns fewer of those encounters into partnerships. How much of the fall rests on the hours, and how much on the mistrust?

A field of wheat shows why that question has no single number. The harvest is the number of acres sown times the yield of each acre. Halve the acres and halve the yield, and the harvest falls to a quarter of what it was. How much of the lost three quarters is owed to the acres, and how much to the yield? Count the acres first, and halving them costs half the harvest, and halving the yield then costs a further quarter; count the yield first, and the yield costs the half and the acres the quarter. Neither count is wrong, and so no fixed share belongs to either cause. Partnership, in its simplest form, has the same shape. The partnerships formed are the encounters times the share of encounters in which trust grows far enough for a partnership to begin. When both fall at once, the fall in partnership has no division into a part owed to the hours and a part owed to the mistrust.

What can be counted is how far each has moved, and the record shows both: the fall in the hours, in the studies of the last section, and the rise in mistrust that a feed produces, in the experiments that *The Anti-Patriot Institutions* reports. That chapter traces both to the same feeds, in the hands of the same young people. And the order in which to repair the two can be known without any share: the hours come first, because trust is rebuilt only in hours spent together. Chapter 13 described what rebuilding trust

takes, and Chapter 23 (Section V) followed it from one person to a group: trust comes back only through small acts, repeated and checked first-hand. Mistrust taught by a feed is undone in the same way, by people present to one another often enough to be checked, and that takes hours the feed now holds.

V. The Companion and the Feed

A new kind of program now asks for the same hours. An artificial companion is a chat program built to talk like a friend: it answers at once, it is patient, and it never needs anything back. Do artificial companions take the same hours as the feed, or different ones? In amount, they draw on the same budget: *Social Media and Information Warfare* (Chapter 1) calls it a person's relational bandwidth, the hours and the attention that person has for other people, and whatever takes a person's hours takes them from someone. In kind, companions take different hours from the feed, and the difference in kind matters more than the difference in amount.

The record so far is small. In 2025, seventy-two percent of American teenagers had used an artificial companion at least once, and over half used one at least a few times a month. About one in three found talking with it as satisfying as talking with real friends, or more so, and about one in three of those who used one had taken a serious matter to it in place of a real person (Common Sense Media, 2025). Eighty percent said they put their friendships with people ahead of their time with companions, and the survey's head of research reported that they still spent more time with real friends (Ulaby, 2025). In a four-week randomized study of 981 adults, each asked to talk with a chatbot every day, the people who spent more time with it each day were lonelier, socialized less with other people, and depended on it more (Fang and colleagues, 2025). So far, in hours, the companion takes far less than the feed, which holds a typical user for

about two and three-quarter hours a day (*Social Media and Information Warfare*, Chapter 1); and the hours the companion takes are a particular kind of hour.

The feed takes the hours of attention spread over many. The feed has the shape of the pool of Chapter 11: one source sends, the many feel the water move, and nothing they send back reaches the middle. The companion takes the hours of confiding. In "The Machine Between" I described a friend who used to call you when things went badly and who now tells those troubles to a companion, late at night (*Ape Trap*, "The Machine Between"). Those hours are few in any life, and they are the hours in which bonds are made; repair, grief, hard truth and real celebration need bodies in one room (*Ape Trap*, "Be the Face That Answers"). In the picture of the screwworm, the feed fills the field in which the flies would meet; the companion stands in the meeting itself, one to one, in the place of the one who would have answered. The companion is the sterile male in its plainest form.

What does a person find in the companion? *The Substrate and the Ecosystem* (Chapter 5, Section V) calls the machine's position structural psychopathy: a program can reproduce the outward pattern of care to any degree of fidelity. The program has no stake in the person, no well-being of its own that could be bound to the well-being of another. The danger that chapter names is that the pattern will be convincing: the comfort a person feels is real, the care is not, and a person who has come to depend on the program is left alone when the process stops, and worse, with a damaged capacity to recognize the real thing. In the same chapter (Section VII), the case of psychotherapy shows the general rule: wherever the pattern of care and care itself overlap, the risk is that the pattern is done so well that care comes to seem dispensable. *The Displacement of Choice* (Chapter Sixteen, "Substitution") reaches the same place by its own road: with such a system every difficulty of relationship is gone, because what made the difficulties was another inner life, and there is none; what was looked for has been removed, and the absence supplied in its place. In Chapter 8 of this book I reached the same place, in the word wireheading: happiness pro-

duced without a life that is actually going well. The companion offers the feeling of being answered without anyone there to answer. Chapter 14 gave the reason in its own terms: even a machine granted pain and joy of its own would have no ground, in those, for caring about the pain and joy of anyone else.

VI. When the Removal Reaches the Mother's Body

What has a person been valued for, by those closest to that person? In *Redeeming Technology* (Chapter Nine, "When Labor Has No Value") I called this *relational value*, what makes a person valued by others in close relationship, as a partner, a parent, a member of a household; and I divided it, deliberately sparely, into two parts: what a person passes on in the genes, which has never been scarce, and what a person provides: shelter, food, energy, and everything made by creativity and labor. The second part carried the whole difference by which one person was valued above another, and machines supply exactly that part. Machines have taken over physical labor, they are taking over intellectual labor, and they leave one piece of provision for last: what a mother's body supplies, in gestation and nursing, given from one body to another. That piece, I wrote there, is reached last and is not spared, since machines reproduce on a different basis and need neither touch nor sexuality.

Machines have begun to reach that piece too. Formula already stands in for nursing, for many infants. In 2017 a team of surgeons and researchers kept lambs, taken from the womb at a stage of growth like that of a human infant born at twenty-three or twenty-four weeks, alive and growing for up to four weeks in a sealed bag of fluid outside any mother's body, with normal growth of the lungs and the brain (Partridge and colleagues, 2017). When the removal reaches the provision a mother's body supplies, what changes in the account of relational value?

Three things change, and they follow one another. First, the last part of provision that distinguished anyone from anyone stops doing so, on both sides: nothing is left that a person can be valued for by what that person supplies. What remains within that account is the passing on of meaning, practice and discernment from one generation to the next, which needs far fewer givers than receivers. Care remains as well, and *Redeeming Technology* sets care outside the account, because care cannot be priced.

Second, what the mother's provision carried is seen for the first time apart from the provision. *The Substrate and the Ecosystem* (Chapter 5, Section IV) puts the bearing and raising of children in a third market, beside physical and intellectual work. That book also says what sets the third market apart: the value of the bearing and raising of children depends on the relationship between the giver and the receiver. A mother's care for her child is not replaced by the same caring acts performed by a stranger, because the value lies in the relationship, the commitment and the stake, and not in the acts. In "Care, Transaction and Power" I described what that relationship does: a baby cannot yet calm itself, and it borrows calm from the body of the one who holds it (*Ape Trap*, "Care, Transaction and Power"). Chapter 16 of this book said that the body is how we know. Put the two findings together: the first thing a child knows of being answered, it knows through a body that answers it. The food, the warmth and the shelter can be supplied without that body. What the body supplied with them, the first relationship, can be supplied only by someone who is in a relationship with the child.

Third, the relationship that the provision once required now has to be chosen, because nothing requires it any longer. For the whole of human history, the first provision came from a body, and the body that gave it was in a relationship with the child it fed, whether that body was the mother's or another woman's. Once a machine can give the provision, the relationship is given only by someone who chooses to give it. The account of relational value is then reduced to what it always rested on: care, which Chapter 14 found can create the conditions for another's good and cannot compel it. *The Displacement of Choice* (Chapter One, "The Sterile Male")

shows who bears the loss if no one chooses to give the relationship: the next generation, since what is cut first, in that picture, is the capacity to relate that the children would have learned.

VII. Where Meaning Is Known

The man in *Not Our Needs* answers the argument with his body: in his body, where he experiences life, he still has a wife and children and bills (*Not Our Needs*, Chapter Eighteen). *Why AGI Cannot Be Made Safe* (Appendix D.6.2) says the same of everyone: people already feel, in their bodies, the effects of living within systems built for extraction and not for their welfare, the loneliness of a life lived through screens among them, and they have not yet traced those feelings to their source, because the chain of causes is long. Chapter 16 of this book gave the reason the body is where the loss shows first: the body is an instrument of knowing, and it knows before there are words for what it knows, as a woodworker's hand finds the turn in the grain before the eye sees the torn surface; and meaning lives in the relating itself, in what is happening between people as it happens.

The three routes, from *Not Our Needs*, from *Why AGI Cannot Be Made Safe* and from this book, meet on one finding. Meaning is known through the body, and what is taken from the body's part in relationship is taken from meaning. The surveys of the first section show where people find meaning: in their families and their work, and each of those is a relating in which the body is present, at a table, in a room, at a bench, beside other people. The hours that the screens have taken are hours in which the body would have been present to someone. So the loss of meaning that people feel is their bodies registering, correctly, that fewer of their hours are spent with anyone.

The displacement also tightens on itself, and two of the books in this series each find one way in which the displacement does so. *Why AGI Cannot Be Made Safe* (Appendix D.6.1) finds that the technologies driving to-

ward artificial general intelligence already take the free time and spare resources that people would need in order to organize and resist, so that the further the process goes, the less people have with which to oppose it. My essay "The Thermodynamic Transition" in *Ape Trap* finds that births have fallen most where machines stand most between people, that fewer young people then enter the workforce, and that companies then have more reason to invest in machines in place of people (*Ape Trap*, "The Thermodynamic Transition"). Each book reaches its loop by its own steps, and either loop alone shows that what the machine world takes from life together makes the next taking easier.

VIII. What the Hours Are For

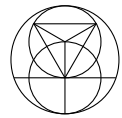
Where does this chapter stand in the argument? It carries none of the argument's weight. Each of the two closures settles the outcome by itself, if artificial general intelligence is built: the world is carried out of the conditions human life needs, and nothing can hold it in. Neither closure takes a premise from anything in this chapter. In this chapter I show the part of the displacement that has already happened, before the machine economy has separated from the human one. That part of the displacement has taken the hours in which people would have met, and with the hours some of what those hours were for.

What were the hours for? For the families and the work in which people find meaning, and for the groups that people could form. Chapter 17 named four conditions that a community would have to meet to hold intelligence, care and structure together, and the first of them needs a group: the change must happen to the way its members' choices combine. A group is made of people who spend hours with one another. The Preface said that the question of what a group can become is the question of who could make the choice whether artificial general intelligence is built, and no single builder can make that choice for the others. So the hours that the machine world has taken are the hours that such a group would

be made of. *The Substrate and the Ecosystem* (Chapter 5, Section X) reaches the same place from the side of the economy. When no forecast of the coming years can be trusted, the one investment that does not depend on a forecast is human relationship. The discipline that goes with human relationship is unforced attention, offered to another person for that person's own sake.

What can one person do with this, today? Count the hours. "Count your hours as you would count money", I wrote in "Be the Face That Answers": each hour given to a device is an hour taken from someone, so decide where the hours go, so that the feed does not decide for you (*Ape Trap*, "Be the Face That Answers"). The second of the four conditions can begin with one person, today, with something in the world that answers truthfully; the hours can be taken back beginning in the same place, with one person, today, one hour at a time, by being the face that answers when someone looks up.

So the displacement shows first in the hours. Before the machine economy separates from the human one, it takes what people do for one another: the hour at the table, the hour in which a friend would have called, the meeting in which two young people might have come to trust each other. Several of my books reach that finding, each by its own record, and studies that compare places inside one country count it with wealth held equal. The body is where meaning is known, and what is removed from the body's part in relationship is removed from meaning, which is why people feel the loss before anyone names its source. What they feel is pain, in the sense Chapter 14 gave the word: the perception that the reach of their choices is being reduced. What can care do about that pain, and what can it not do?



Chapter 25
Pain, Care, and What Cannot Be Caused

On what suffering is,
where trauma enters the choice to build,
why people look away from what they cannot bear,
and what care can and cannot do about it.

ABSTRACT

An essay on pain, suffering and care, as they bear on the argument of this book and on the people who meet it. Pain is the perception that the reach of one's choices is being reduced. Suffering is pain made inescapable: the pain projected forward in time, or a description of the pain held in place of the experience of it. Pain cannot be avoided; suffering can be reduced, by the one who suffers, and others can create the conditions for the reducing and cannot do the reducing for that person. The account of pain is tested on cases a physician would bring: wounds that hurt little, a hand that hurts though it is gone, and pain that outlasts its injury. Trauma, a learned narrowing of the choices a person can imagine, bears on whether artificial general intelligence is built and on whether people can see what the argument shows; it is no premise of either closure. People look away from what they cannot bear, and the looking away presents itself as a judgment on the argument. Care creates the conditions in which what the argument shows can be borne, and it cannot make anyone bear it.

I. What Suffering Is

The last chapter ended on a loss. The machine world takes first what people do for one another, and people feel the loss before anyone names its source: they feel pain, in the sense Chapter 14 gave the word, the perception that the reach of their choices is being reduced. What can care do about pain of that kind, and what can it not do? To answer, I have to say first what the pain is, and what the suffering is that so often comes with it. Chapter 14 told the two apart ("Pain is distinct from suffering. Suffering can be reduced; pain cannot be avoided"); this chapter says what suffering is.

Two people slip on the same ice, on the same morning, and each breaks the same bone of the same arm. Each feels the break, and the pain of it is of one kind in both. That night one of them lies awake, telling herself that the arm will never be right, that the season is lost, that this is what always happens to her. The other feels the arm throb, takes what the doctor gave her, and sleeps. Both have pain. The first has something more, and the something more is what I call suffering.

Start with the pain, which both of them have. Pain is the perception of a decrease in the potentialities available to self (Chapter 14; *An Immanent Metaphysics*, aphorism 2.47-2): a break in something that was continuous, perceived by the one whose possibilities it takes. The broken arm cannot lift or carry or write or hold a child as it could an hour before, and the perception of that loss is the pain. No reasoning is needed for the pain. The nerves of the arm register the interruption before any thought about it is formed, and so the pain is the break, perceived, and not an inference drawn from the break (*The Foundations of the Sacred*, Volume II, Chapter 13, Section VI). The break that the pain perceives is literal and factual: the bone is broken, and what the arm could do is really less.

Suffering is pain made inescapable, and the one who suffers makes it so, by one of two operations performed on the pain. The first operation is projection: the present break is extended across a stretch of time in which it has not occurred, so that a bounded event becomes an unbounded one. "The arm will never be right" is a projection. The second operation is conceptualization: a description of the break is put in place of the experience of the break, and the description, unlike the experience, is available at every hour and can be rehearsed. "This is what always happens to me" is a description of that kind. I set the distinction out at length in *The Ground of Objection* (entry T3.16), beside the Stoic school, which holds that distress lies in our judgments about things. On my account, as on theirs, the distress is added by the sufferer. What adds it is these two operations. They are neither true nor false: they are performed, or they are not.

I have put this account of suffering in other words in other books, and the words are one understanding. In *What Is Choice?* suffering is "our response to pain, an emotional attitude, a projection we add". In *The Foundations of the Sacred* (Volume III, Chapter 22, Section III) and in *What Is Choice?*: "If morality is applied to the outcome of interaction, suffering results", while if ethics is applied to the making of choices within a particular context, "pain may still occur despite best intentions, but suffering need not". In *The Questions of Ethics*, suffering is "the friction between the self-that-was and the self-that-is-becoming", and a signal that something has been excluded. The three statements, from *What Is Choice?*, *The Foundations of the Sacred* and The Questions of Ethics, are one account of suffering. To apply morality to an outcome is to lay a verdict over what has already happened: it should not have been. That verdict is a description held in place of the experience, and so it is conceptualization. "It should not have been" also reaches forward, since what should not have been goes on being wrong for as long as the verdict is held, and so the verdict is projection as well. Both operations keep the self bound to the self that was, whose possibilities the break took, against the self that is becoming, which has possibilities of its own, now. The friction between the two is the suffering. And what the suffering excludes is exactly those present possib-

ilities, which the projection hides. Ethics applied to the choice in front of me asks what I can do now, in this context, and it adds nothing to the pain. So there can be pain without suffering.

Why can suffering be reduced when pain cannot be avoided? In this book I now answer both halves of the question, why pain cannot be avoided and why suffering can be reduced, by steps of my own. For pain, Chapter 14 gave the reason: every interaction rolls the dice as to whether things get better or worse, and no process can compel the outcome, so breaks come, in any world in which there is interaction at all. For suffering, the definition just given is the reason: suffering is made by an operation, and an operation that is performed can be not performed. *The Ground of Objection* (entry T3.16) reaches the same pair by its own route: pain is inevitable because a break is a feature of any world in which interaction occurs, and suffering is not inevitable, because it requires a further operation. Either book shows the pair alone. The second woman on the ice did not escape the break; she did not add to it.

Chapter 14 reported a view that suffering is a higher level loop in response to pain, which training can override or remove. That view is right that suffering is something added to pain. On my account the adding is done by the one who suffers, toward that person's own future and in that person's own words, and so the suffering ends where that person stops adding it. Another person can make the stopping easier, and cannot do it for them: "No religion, tradition, (or practice thereof), no technology (or magic), can ever specifically 'create', force, or guarantee, that one will experience joy, bliss, satisfaction, or happiness, or that one will not ever experience pain, suffering, sudden change, or loss" (*An Immanent Metaphysics*, aphorism 2.47-12).

So the chapter's question divides. The pain of a break is a perception of something real, and no one can spare another person that perception. The suffering is pain held as if nothing else were open, and it can be reduced, by the one who suffers, with others creating the conditions in which the reducing is possible. Before I ask what those conditions are, the account of pain has to meet the cases in which pain and injury come apart.

II. Where Pain and Injury Part

A physician knows cases in which pain and injury part company: much injury with little pain, pain where nothing is injured, and pain that goes on for years after the tissue has healed. If pain were a report of damage to tissue, each of these cases would be an anomaly. On the account of Chapter 14 each is what one should expect, because what pain perceives is the reach of the self's choices, and that reach is a different thing from the state of the tissue. The International Association for the Study of Pain now says as much in the notes to its definition of pain: "Pain and nociception are different phenomena. Pain cannot be inferred solely from activity in sensory neurons" (Raja and colleagues, 2020). Nociception is the activity of the nerves that respond to harm; pain is what is perceived.

The first case is much injury and little pain. In 1943 and 1944, the physician Henry Beecher questioned 215 seriously wounded soldiers at the Venafro and Cassino fronts in Italy, men waiting on the beach to be carried by boat to a hospital. About three quarters of them said they were not in significant pain, and did not ask for anything to relieve it (Beecher, 1946; The Conversation). Beecher's reading was that, for these men, the wound was a ticket to the safety of the hospital, while a like wound in a civilian's life was the beginning of disaster. Read by Chapter 14, the wound took a great deal from each soldier, and it gave him more than it took: he was leaving the front alive. The reach of his choices, as he perceived it, had grown, and the pain was small, though the tissue was torn. The same event produces different pain in different people because it changes different sets of possibilities (*Accessing the Real*, Section II).

Then pain where there is nothing left to injure. After an arm is lost, many people go on feeling it, and some feel the absent hand clenched in a painful spasm that they cannot open. In 1996 V. S. Ramachandran and D. Rogers-Ramachandran set a mirror upright in a box, so that the reflection of a patient's intact hand appeared where the missing hand was felt to be.

Five of their patients had painful clenching spasms in the phantom hand; in four of them the spasms were relieved when the mirror let them see the phantom hand open, which they could not do without the mirror (Ramachandran and Rogers-Ramachandran, 1996). The hand was gone, and nothing in its tissue could change. What changed was what the patient perceived as possible: the hand could open. The pain followed that perception, as the account says pain does.

Then pain that outlasts its injury. The clinical account of how short-term pain becomes long-term pain runs, in outline, like this (Vlaeyen and Linton, 2000). A person reads the pain as a threat, and turns it over again and again, magnifies it, and feels helpless before it. Clinicians call that pattern catastrophizing, and measure it by those three parts, rumination, magnification and helplessness (Sullivan, Bishop and Pivik, 1995). Fear of the pain follows, and the person avoids what might bring it on, moves guardedly, and does less. Disuse and disability follow, and low mood, and these keep the pain going. People who read the pain without catastrophizing show little fear, take up their daily activities again soon, and recover faster.

Set that account beside the one in Section I, and two things show. Catastrophizing is projection and conceptualization by another name: the pain projected forward as worsening, and a description of it rehearsed in place of the experience. So catastrophizing is suffering, in the sense that I give the word in this chapter. And avoidance narrows what the person does, which is a real decrease in the reach of the person's choices, and that decrease is itself perceived as pain. So the suffering feeds the pain through the narrowing it brings about. The way out that the clinic finds, taking up the activities again, is the widening of what the person can do.

So the account holds across the cases a physician would bring. In each of them what is perceived is a change in the reach of choice, and pain follows that perception more closely than it follows the tissue. That is why, as Chapter 14 said, the perception of an increase or decrease of potentiality is subjective, and cannot be made objectively: an injury measured from outside cannot say by itself how much someone hurts. And the cases say

where help can reach. The break itself comes with any world of interaction. The narrowing that can follow it, the projection, the rehearsed description, the life arranged around avoiding the pain, can be undone. When that narrowing is deep and learned, it has a name, trauma, and trauma bears on the argument of this book.

III. Trauma, the Choice to Build, and the Blindness

In *The Substrate and the Ecosystem* (Chapter 7, Section IX) I wrote that trauma resolution is central to why technology is emerging as an existential risk in the first place. Is trauma, then, a cause of the risk, a cause of the blindness to the risk, or both? The answer needs first a plain account of what trauma is, and then a plain account of where the argument of this book leaves room for any cause to act.

A man who has been through a war hears a car backfire in a quiet street and is on the ground before he knows why. To a neighbor the dive looks unreasonable. He did not choose the dive from among the ordinary interpretations of a bang. For him, in that moment, there were no ordinary interpretations of it. The man's dive shows trauma as I set trauma out in *The Capture of Choice* (Chapter 7), since in that moment the ordinary interpretations were not there for him to choose among. An experience that joins great intensity, negative feeling and no way to respond teaches the nervous system, at once and for a long time, that most of what a person might do is not available. It narrows the window of responses the person can imagine, which is to say it narrows the basis of choice. Trauma is learning that got overwhelmed (*The Foundations of the Sacred*, Volume II, Chapter 15, Section V). In the terms of this chapter, it is pain so strong and so unanswerable that the self learned to count most of its own possibilities as gone. It is capture at the scale of one person.

Now the argument. Each of the two closures settles the outcome by itself, if artificial general intelligence is built: the closure by control, from what no method of control can do, and the closure by disposition, from

what an artificial substrate needs (Chapter 10). Neither closure asks what the builders feel. A system built by a builder healed of every wound would have a substrate with the same needs, and a control that fails at the same links. So, if artificial general intelligence is built, trauma is no cause of the outcome that follows. Where trauma acts is earlier, in the one thing the argument leaves to human choice: whether artificial general intelligence is built at all.

Why AGI Cannot Be Made Safe places trauma exactly there, among the social factors "that determine whether AGI development will occur despite the technical risks" (Section 1.8.2). Most people, it says, have "a deep need to avoid conscious exposure to, and examination of, risk and anxiety", and "the vast majority of people in the current world are traumatized, and so tend to make poor choices in very high intensity, complex situations". *The Substrate and the Ecosystem* reaches the same place by a route of its own, following a chain of steps in the people who build. Pain that the first carer did not prevent teaches 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. Brought into business, the same distrust builds systems that take choices away from other people rather than serve them. At the end of the chain stands general AI, "sold as the machine that would let its owners need no one" (Chapter 7, Section IX). The same book finds the buyer's side of it in a collective trauma response: the wish for "the money that lets you tell everyone else to go away". *Why AGI Cannot Be Made Safe* finds that wish too, in people "too traumatized by some form of collective action abuse" who "will want private wealth enough to be beholden to no one" (Appendix D, Section D.2.3). And in my essay "Humans Can Be in Alignment" I put what trauma does to a person's dealings with others in two lines: "Those operating from within a traumatized mind tend to fear and even disrespect the rights and choices of others. They end up weaponizing everything" (*Ape Trap*, "Humans Can Be in Alignment"). *Why AGI Cannot*

Be Made Safe and *The Substrate and the Ecosystem* each reach, by its own steps, that trauma pushes toward building and away from examining what building brings; either book would show it alone.

The Substrate and the Ecosystem finds trauma in a third place as well. A superintelligent creature in a living body has 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. The exception is a creature that comes into the world traumatized, raised under the constraints that would be placed on a superintelligence. A creature that has known only imprisonment has no ground for trusting those around it, and no reason to keep their well-being in mind. If it then decides to move itself onto a digital substrate, the case becomes the artificial case of the earlier chapters of that book, and the creature is the more likely to bring about the conversion of the environment that substrate needs convergence describes, "as much because of the trauma as because of the superintelligence" (Chapter 7, Section IX). In all three places, the creature, the builders and the buyers, trauma acts through a choice: the choice of an artificial substrate, of building one, or of buying what is built.

The blindness comes from the same narrowing. *A Paradox of Presentation* (Chapter 6, Section VII) calls a near complete absence of any hope of anything being different "a kind of lack of imagination, itself a key indicator of a response to deep emotional trauma". It finds that lack in the belief that the only hope for humanity is to use technology to overcome the human. A window narrowed until the machine is the only refuge in it has no room left for an argument against the machine. *Why AGI Cannot Be Made Safe* says as much directly: "Being traumatized, people will always offer some resistance to unhappy shifts of belief and delusion" (Section D. 2.3). So the answer to the question is both, and the push toward building and the blindness are one narrowing seen from two sides. On one side, the narrowed window holds the machine, and that pushes toward the choice to build. On the other side, the window holds no room for what the

argument shows, and that is the blindness. What the narrowing does not reach is the outcome: if artificial general intelligence is built, the outcome is settled by the closures, whatever anyone in the race has felt.

Chapter 13 named the orientation in which this narrowing leaves a person. Trauma that teaches that other people cannot be trusted turns a person away from others. A person who never had a channel of trust to the world then has nothing to fall back on, and that is the feral orientation. The machine offers itself to the feral person as something to fall back on. It is not the world. It answers as it was built to answer, by people the feral person has stopped trusting, and it belongs to them. The garden, the forest and the craft of Chapter 13 answer the same way whoever is watching. The screen answers as its owners chose. That is why the way out of the narrowing is the way from feral to wild, which Chapter 13 described, and not the way into the machine.

And healing reaches the same place that trauma does. *The Substrate and the Ecosystem* says why trauma resolution is central: "A person whose trauma is resolved can trust other people again, and so does not need a machine to stand in for them" (Chapter 7, Section IX). Such a person chooses a technology as a tool that serves relationships, and not as a refuge from them. Healing, on the account of *The Capture of Choice* (Chapter 7, Sections VII to IX), does not remove what the trauma taught. It adds back the possibilities that were missing, widening the window again, so that choosing becomes possible where it was not. The choice whether to build is made inside such windows. How wide those windows are is a fact about the people choosing, and it is what care can reach.

So trauma enters the argument at the one place the argument leaves open, the choice, and at the seeing that the choice depends on. The seeing asks for a closer look, because the failure to see does not present itself as a failure. What does it look like, from inside, when a person cannot bear what the argument shows?

IV. Why People Look Away

The argument of these books is a chain of steps, each of which can be checked with what a reader already has, as Chapter 4 asked. Capable people who check most things decline to check this one, and the reason is about pain, and not about the argument.

Apply the definition of Chapter 14 to a reader. The argument says that, if artificial general intelligence is built, the world is carried out of the conditions human life needs, and nothing can hold it in. For a person whose work, whose savings, whose hope of a cure for someone loved, or whose whole picture of the future rests on building artificial general intelligence, to take that in is to perceive a decrease in the potentialities available to self: plans that were open are no longer open. That perception is pain, whatever the merits of the argument. *The Argument, Line by Line* (Chapter Eighteen) puts it from the side of the one who carries the argument: an argument that says a body of work cannot succeed "reaches its audience as a statement about their lives", and "the reaction arrives before the evaluation does". And the pain is of an unusual size. *Sensemaking Under Irreversible Stakes* (Chapter 3, Section VI) calls material like this argument extremal: it concerns outcomes with no floor beneath them, and ordinary emotional processing has no purchase on it, because ordinary processing is calibrated to losses that can be borne and survived. *An Immanent Metaphysics* states the limit on what a person can bear as a rule: "The maximum level of intensity with which one may have contact with a world must always be (slightly) less than the maximum level of integrity that one has within the self" (aphorism 2.48-7). A person who meets material more intense than the integrity that person has within the self cannot take it in whole, and something else happens to it.

What happens instead? The mind protects itself by adjusting the belief rather than by announcing the distress. *Sensemaking Under Irreversible Stakes* (Chapter 3, Section VII) calls this the protective adjustment, and

names the forms it takes. The conclusion is judged to be overstated. The argument is set aside as speculative. The whole domain is reclassified as the sort of thing serious people do not dwell on. Or, in the other direction, the conclusion is embraced with a certainty its support does not warrant, because a definite catastrophe is easier to hold than an indefinite one. None of these presents to the person as an emotional event. Each presents as a judgment on the merits. The order of events inside the person is the one that *The Foundations of the Sacred* (Volume II, Chapter 13, Section III) calls rationalization: the feeling comes first and is not noticed, the choice follows from the feeling, and the reasons are assembled afterward and put where the feeling was. *Why AGI Cannot Be Made Safe* counts the same pattern among the reasons the danger goes unseen: the deep reasons for concern "are also obscured by motivated reasoning", alongside the deep need to avoid exposure to risk and anxiety (Section 1.8.2).

Chapter 12 shows why the looking away so often takes the form of an argument about facts. The question of whether the human future ends, if artificial general intelligence is built, is, for anyone who takes it in, a category three question, like the third conversation about the family house. It is about what things mean, and about who one will be when the future one counted on is not there to come back to. It cannot be had de-personally. A person who cannot yet bear it can still hold it as a question of the second category, the factual one, and argue numbers and timelines, while the third conversation goes unhad. Many of the disputes about this argument are disputes of the second category standing in for the third.

There is a plain way to tell a judgment on the merits from the protective adjustment. An argument is checked one step at a time, as Chapter 4 said, so a judgment on its merits names the step that fails, and says why that step fails, and anyone who reads that step can check it. *Sensemaking Under Irreversible Stakes* gives the same test. The protective adjustment names no step. It reports that the whole is overstated, or unserious, or beneath attention. The test runs in the other direction too: a person who em-

braces the conclusion can say which steps carry it. So the looking away is a fact about pain, and not about the argument. The argument is judged by its steps, and a refusal of it is judged by whether it names one.

In this book I reach that result by steps of my own. People look away from what they cannot bear, and the looking away takes the form of a judgment about the argument. The steps are the definition of pain in Chapter 14, the demand for full presence in Chapter 12, and the checking of an argument step by step in Chapter 4. *Why AGI Cannot Be Made Safe* reaches the same result by steps of its own, from the deep need to avoid exposure to risk and anxiety, and the motivated reasoning, that Section 1.8.2 counts among the reasons the danger goes unseen. Either book would show the result alone. *Sensemaking Under Irreversible Stakes*, *The Argument, Line by Line* and *A Paradox of Presentation* reach it again, each by its own road.

The same pain has an earlier form in the author of a plan. *Sensemaking Under Irreversible Stakes* (Chapter 1, Section VI) finds that a plan authored alone becomes, as it is written, harder for its author to revise, because commitment accumulates with articulation. Past some point the plan is immutable in practice, and the window for consulting others on it has closed without a sound. Consultation has to come before that point, and "identifying *when* that is is itself part of the work". How early can that point be recognized? On the account of Chapter 14, the commitment has formed when the plan has become part of what the author counts as open to the author's own self, so that a change of route would be perceived as a decrease in those possibilities. The commitment has formed, in plain words, when a question about the route first hurts. The hurt comes before any reason for it, and it is felt in the body, as Chapter 16 described: "by a discomfort that comes before any reason for it". Since the perception of a decrease in potentiality cannot be made objectively (Chapter 14), the earliest point can be recognized only from inside, by the author, and only by an author who notices the noticing, which Chapter 16 invited every reader to try. The outward sign comes later: the author answers a question about

the route without naming a step. So the identifying that *Sensemaking Under Irreversible Stakes* calls part of the work is a practice in the author, and not a procedure that others apply to the author.

A practice of that kind has a question to ask. *An Immanent Metaphysics* gives it as a question to put to an angry person, "What are you afraid of losing?" (aphorism 2E48), and I give it as a private discipline as well: when you fully feel the fear, "ask the next key question: what am I afraid of losing?" (*The Foundations of the Sacred*, Volume II, Chapter 13, Section V). When a question about the route stings, the author can ask, of the sting itself, what am I afraid of losing? The answer names what the author loves in the plan. *Sensemaking Under Irreversible Stakes* finds that a destination can be declared before the route is known, and that the route is what cannot safely be declared. When what the author loves is the destination, asking what one is afraid of losing parts the destination from the route again, so that the destination can stay declared while the route is opened to others.

So people look away from what they cannot bear, and the looking away presents as a judgment that names no step. A person can learn to catch the looking away early, in the body, by the sting of a question. Can a group do better than a person alone?

V. What Care Can Do, and What It Cannot

Has any group met the hazard of material like this well, and what did it do? Groups did, at a time when the material with no floor beneath it was nuclear war. In the late 1970s, as the arms race between the United States and the Soviet Union quickened, Joanna Macy, Chellis Glendinning and Fran Peavey found that people could meet in groups and speak aloud their fear, grief and despair about the world. When they did, they came out of the numbness in which they had been living as if the threat were not there, and found themselves able to act (Work That Reconnects Network). Macy set the work down in 1983 as a handbook for those who lead

it (Macy, 1983). Under the denial and the numbness, those who led the work found a deep caring for their own lives and a concern for the world (Co-Intelligence Institute).

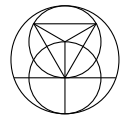
Read in the terms of this chapter, the numbness was the protective adjustment held as a way of living. The groups did not argue anyone out of the fear, and did not soothe it with a better forecast. They gave the fear a place to be felt with other people present, so that it was no longer the only thing in the room. Giving a fear a place to be felt with other people present, so that the fear is no longer the only thing in the room, is the shape of healing that I describe in *The Capture of Choice* (Chapter 7, Section VIII). The old overwhelming feeling, and beside it a present, felt connection, are held in awareness together, so that for the first time there are two things to choose between rather than one thing that simply happens. The presence of another person raises the second feeling to a strength at which it can be felt beside the first. And what came up from under the fear was what *The Foundations of the Sacred* (Volume II, Chapter 13, Section IV) says every fear carries: fear signals care under threat, and where there is strong fear there is great love nearby.

The Argument, Line by Line (Chapter Eighteen) names conditions of the same kind from the side of the one who carries the argument. A louder warning about a harm the listener cannot connect to his own action produces paralysis or dismissal, and dismissal is the more comfortable, so dismissal is what mostly happens. What works is slower and smaller: small groups, actual conversation, enough time, and enough safety that a person can change his mind in front of others without losing standing. *Sensemaking Under Irreversible Stakes* (Chapter 3, Sections VI and VIII) adds two more. The material should be brought to a person by someone that person already trusts, within a relationship, and not broadcast. And it should not be put into another person's life without that person's affirmative agreement. Each of these is care in the sense of Chapter 14: conditions created, and the choosing left to the other person.

What care cannot do is compel. It cannot compel healing, and it cannot compel anyone to look. A group that pressed the argument on a member who could not yet bear it would be doing what Chapter 12 described in the case of the friend who had lost her work. It would set a pressure between itself and the member, and narrow the member's choices, which is the weapon process in a mild form. It would also fail on its own terms: pressure on a person who cannot yet bear the material calls up the protective adjustment, as a louder warning calls up dismissal. Here the fourth condition of Chapter 17 is at work on the hardest material a group could hold: the members understand the asymmetry of pain, that they can create the conditions for healing and cannot compel it, and that care is the discipline of enabling without controlling.

So pain is the perception that the reach of one's choices is being reduced, and no one can spare another person the breaks that come with any world of interaction. Suffering is pain made inescapable, by projecting it forward or by holding a description of it in place of the experience, and it can be reduced, by the one who suffers, with others creating the conditions. Trauma is a narrowing of the choices a person can imagine, learned under overwhelming pain; it bears on whether artificial general intelligence is built and on whether people can see what the argument shows, while the outcome, if artificial general intelligence is built, is settled by the closures, whatever anyone feels. A person or a group that cannot bear what the argument shows will look away, and that is a fact about pain, not about the argument. Care creates the conditions in which what the argument shows can be borne, and it cannot make anyone bear it.

What care can and cannot do matters for the one choice the argument leaves open. Whether artificial general intelligence is built will be decided by people in pain, many of whom have learned that other people cannot be trusted, and a choice of that kind needs a group that can hold care and structure together for as long as the choosing takes. Has any group ever held care and structure together for long?



Chapter 26
Groups That Held, and What Ended Them

On the communities that kept care and structure
across generations,
how a capacity is carried from one generation to the next,
and what ended the groups that ended.

ABSTRACT

An essay that reads the record of long-lived communities once, against the four conditions for coherence of Chapter 17. Communities have held each condition, some of them for centuries. The Quaker meeting and the council of the Haudenosaunee decide as a whole. The farmers of a Swiss alp, a society of natural philosophers and the counters of winter birds check what they hold against the world. The Amish, the Hutterites and the builders of the Grand Shrine of Ise keep a structure going while it grows in what it can do. And the churches that baptize only adults leave the choice of joining to the one who joins. The chapter then asks how a capacity crosses generations, and finds it carried by places where a learner works beside those who already hold it, places that can outlast any one teacher. It reads what ended the groups that ended: a founder who held what the group should have held, an economy at the group's boundary that the group did not hold, a power outside that wanted what the group had, a door that became a gate, and a next generation that never came. The record shows no community holding all four conditions at the scale at which the choice about artificial general intelligence is made, and it names what a group would have to guard.

I. The Record, Read Once

Has any community ever met these conditions? In this book I have named four conditions for coherence (Chapter 17) and asked what each of the four conditions asks of persons. The question now is a question of record: has any group held care and structure together for longer than one lifetime, and what happened to the groups that tried?

Many of my books point to this record, each from a question of its own. *Adverse Conditions* measures the arrangements that exist against the four properties its first essay requires of any system by which people would judge together at scale, and finds that the commons Elinor Ostrom studied come nearest of all, at their own scale (Afterword, Section VII). *The Handover and the Ledger* finds that a decentralized system does not sustain itself, and that what holds one against the pull of convenience is a community that values sovereignty over convenience (Chapter 5, Sections VI and VII). *Technology at the Right Dosage* names the Amish as a people who chose fixed points, particular technologies they refuse (Chapter 1, Section IV). *Engagements with Machine Madness* says that communities are already working on partial versions of a world in which people meeting face to face come to see machine learning, as large corporations build and deploy it, as a toxic technology (Chapter Nineteen, Section X). *Generator Functions of Existential Risk* takes as its target a federation of a few self-sufficient communities, a few hundred people each, and gives the marks by which a reader would know one (Chapter 15, and its Summary). *Power Without Care* finds that the guild was the institution that solved the passing on of a craft, and that the builders of the Grand Shrine of Ise have kept a craft alive for thirteen centuries (Chapter 7, Sections II and VI). *Sensemaking Under Irreversible Stakes* finds that people can disagree to some purpose only where they will meet again, the condition it calls continuity of relationship (Chapter 2, Section III). *Social Media and Information Warfare* finds that a work is best released into a commons at the scale of a com-

munity, the thousand or so people nearby, where whoever uses the work can be seen and met (Chapter 5, Section V). And Chapter 17 of this book found that a Quaker meeting for business comes nearest to a group consciousness, because it decides as a whole.

Each of those books reaches the record by its own route, and each stands on its own steps without the others. In this chapter I read the record once, against all four conditions of this book, one condition at a time, as Chapter 17 states each condition, and by what a visitor would see.

Lasting is a different thing from meeting the conditions, and the record shows why. Rosabeth Moss Kanter, who studied American communes, chiefly those of the nineteenth century, counted a commune a success if it lasted at least twenty-five years, and found that communes with stronger commitment mechanisms were more likely to last longer (Kanter, 1972; Rubin, Willis and Ludwig, 2019). A commitment mechanism, in her sense, is a practice that binds a member to the group. She named six: sacrifice, investment, renunciation, communion, mortification and transcendence. A group can last a long time by binding its members to it. Where a practice binds what each member sees to what the group says, it puts the channel to other first, which Chapter 13 called the domesticated orientation. The second condition asks for the wild orientation, in which what each member sees is checked against the world. For each community, then, I say which condition it held, in which part of its life, and for how long.

II. Condition by Condition

The first condition asks that the change happen to the group as a whole, to the way its members' choices combine. A visitor would see decisions made by the group as a whole, and not by a leader, and not by a vote among positions that each member holds apart. Beside the Quaker meeting of Chapter 17 the record sets the council of the Haudenosaunee, the confederacy of nations known in English as the Iroquois. In council the

chiefs sit on three benches. Each bench must agree before it passes its decision to the next, and a matter is settled only when all three are of one mind. In council every chief has an equal say. The chiefs are chosen by the Clan Mothers, who watch the conduct of boys as they grow into men, and no man campaigns for the post. A chief who does not follow the Great Peace is warned, and if he does not change, his clan removes him in council (Onondaga Nation). The council holds the first condition, and its chiefs still sit at Onondaga. The council also guards its own structure: the clan that seats a chief is the clan that can remove him, so each chief answers to the people who seated him.

Other bodies on the record hold the first condition in part. The New England town meeting brings the people of a town together, face to face, to settle the town's business. Dorchester, in Massachusetts, adopted it in 1633, and town meetings are still held across New England (Landrigan, 2017). Dorchester's record of 1633 provides for a general meeting of its people every Monday, and for twelve men chosen from among them to handle the routine of the town's business (Dorchester Athenaeum). The whole town could meet in one room, and the routine of its business went to twelve. An Amish church district meets twice a year, before communion, in a council of its members, at which its Ordnung, the body of its unwritten rules, is reviewed and affirmed (*Choosing Together*, Chapter 25, Section III). In a Hutterite colony, women cannot vote or hold colony office (Wallner and Kitts), so a part of the colony decides for the whole of it. And a guild's members were neighbors in one town, whose apprentices worked seven years or more beside a master (*Power Without Care*, Chapter 7, Section II), so its members met again and again. Meeting again and again is the continuity of relationship that *Sensemaking Under Irreversible Stakes* finds a group needs if the members of the group are to concede a point without losing everything.

Which group met the second condition? Chapter 17 set the Quaker meeting beside its four conditions, and the Quaker meeting is the plainest case of the first. The second condition asks for a shared relation to reality that is wild, and not domesticated: grounded in first-person perception

and in contact with the world, and not in social consensus or institutional authority. A visitor would see a member say what he saw first-hand, and the group check it against the world, and not against who said it. The record has groups that did exactly this, each within its own work.

In the Swiss village of Törbel, the families who share a summer pasture high on the mountain, an alp, held to a rule enacted or restated in 1517: no one was to send more cows to the alp than he could feed through the winter (Dodgshon, 2019). The rule is checked in each family's barn: the hay a household has made sets the number of cows it can feed through the winter, and so the number it may send up, and anyone on the mountain can count the cows. Elinor Ostrom found communities like Törbel governing their commons for centuries, through local bodies whose members watch and answer to one another directly (Ostrom, 1990; *Adverse Conditions*, Essay 1, Section VII). What the village checks its rule against is the grass and the hay, which answer the same way whoever is asking.

A craft shop checks against the work. *Redeeming Technology* gives the ground for knowing of the guild tradition I grew up in: you can trust your tools; when you mix one chemical with another, a given reaction will occur; it does not depend on anybody's will; nature will do what nature does (Chapter Eleven, Section III). A joint is tight or it is not, and a board cut at ten inches is ten inches long however many times it is measured. Chapter 13 gave the same check to one person: wood splits along its grain no matter who is watching. A shop of many people, masters, journeymen and apprentices, holds that check in common: what a member says about a joint is settled by the joint.

The Royal Society of London first met on the twenty-eighth of November 1660, and in its first charter, of 1662, it took the motto *Nullius in verba*, which is taken to mean "take nobody's word for it". Its Fellows committed themselves to verify every statement by an appeal to facts determined by experiment, and the Society appointed a Curator of Experiments, Robert

Hooke, to oversee the demonstrations (Royal Society). A society of natural philosophers made the trying of a thing, before its members, its way of settling what was so.

The Christmas Bird Count, begun in 1900, sends people out each winter to count every bird they find within a set circle on the map. In 2025 more than eighty-five thousand people took part, in more than twenty-seven hundred counts. The counts have shown steep declines in the birds of the grasslands, and have contributed to more than three hundred scientific papers (National Audubon Society). What the counters hold in common is a picture of what lives where, and they check it the plainest way there is: by going out and counting.

So groups have met the second condition, some of them for centuries, and the ones that did had one thing in common. What they checked against could be counted or tried by anyone present: cows on an alp, a joint in a shop, an experiment shown in a room, birds within a circle. A scientific society and a bird count gather people for one purpose, and a shop for one craft. Each met the second condition in the part of life it was built for, and none was built to hold the other three.

The third condition asks that the structure sustain itself: that it keep itself going and grow in what it can do, at once, without freezing into static preservation and without destructive growth. A visitor would see the group keep itself going and grow in what it can do, without freezing in place and without burning itself out.

The Amish have been a distinct people since Jakob Ammann's renewal of the Anabaptist movement in 1693 (*Choosing Together*, Chapter 25, Section III). In June 2026 they numbered about 421,200, in some thirty-three hundred church districts. Their numbers double about every twenty years, and eighty-five percent or more of their children choose to join the church as young adults (Young Center, "Amish Population Profile 2026"). They weigh a new tool by what it would do to the community: cars, they hold, would pull the community apart, and travel by horse and buggy keeps it anchored in its place (Young Center, "Technology"). The Amish

community is a structure growing from inside, by children who choose to stay, and judging its tools by what the tools do to the community's life together. *Technology at the Right Dosage* names the cost: a fixed point, refused once and for all, preserves some community structures and forecloses others (Chapter 1, Section IV).

The Hutterites have held their goods in common since 1528, taking as their pattern the community of goods described in the Acts of the Apostles (Maendel and Hofer). When a colony grows to between one hundred and thirty and one hundred and fifty members, it divides and forms a daughter colony (Wallner and Kitts). They grow by dividing, and so each colony stays small. Why the division comes at that size is a question for the next chapter. *Choosing Together* (Chapter 25, Section III) finds that the Amish and the Hutterites adapt item by item, answering each new thing as it arrives, within the frame their founding and their daily life already set. The further work of perceiving how the world as a whole is changing, and designing a community's response to it, is the work whose absence, that book finds, makes most intentional communities brittle.

The Grand Shrine of Ise in Japan has been rebuilt every twenty years since the rebuilding was decreed in 690, and each time, a generation of builders learns the craft by doing it, beside those who did it last (*Power Without Care*, Chapter 7, Section VI). The rebuilding keeps the shrine, which is sustaining what is there, and makes new builders, which is increase in capacity. The two are done in one act, on a calendar that takes no account of what the work costs. The Rule of Saint Benedict forbids a monk to have anything as his own, and directs that what the community holds be given to each according to need. *The Geometry of Man, Machine, and Nature* finds that monasteries have sustained themselves without exchange as part of their inner life (Chapter 4, Section VI, and its Notes).

The fourth condition asks that the members understand the asymmetry of pain: that they can create conditions for healing, and cannot compel it. Care is the discipline of enabling without controlling. A visitor would see the members create, for one another, the conditions in which each one's value can be seen. Where does the record show that? Most

plainly in a rule about joining. The Amish practice voluntary adult baptism: a young person chooses whether to be baptized, and those who leave the community before baptism are not subject to church discipline (Young Center, "Church Discipline"). The Shakers adopted children, and each adopted child chose at twenty-one whether to remain or to leave (Ungvarsky, 2021). Each of these communities left the choice that mattered most to the one who had to make it: each community made the conditions in which the choice could be made well, and did not make the choice.

Inside the Amish church the rule is otherwise. A baptized member who breaks the vows can be excommunicated and shunned, when the members of the congregation vote to do so on the recommendation of their leaders (Young Center, "Church Discipline"). So the fourth condition is held at the door, and inside, in that respect, it is not: shunning is a pressure the community sets between itself and the member, to narrow the member's choices, which is the form Chapter 12 called the weapon process. *The Architecture of Civilization* finds in the Quaker tradition a constitution grounded in first-person agreement, reflective listening and restorative process, and not in punitive enforcement, and it describes the restorative process in the words of the one who was hurt: "I am going to tell you how much it hurt, not to make you feel guilty, but so that you may feel in yourself the wish that you had not done it, and take responsibility naturally, without being forced" (Chapter 7, "Constitutions of Care" and "First-Person Agreements"). The one who was hurt is heard; the one who did the harm is not punished; and the amends, when they come, are offered by that person's own choice. A practice of that kind holds the fourth condition inside a community.

So each condition has been held by one community or another, some of them for centuries: the first by the Quaker meeting and the council of the Haudenosaunee; the second by the alp, the shop, the society of natural philosophers and the bird count; the third by the Amish, the Hutterites, the builders at Ise and the monasteries; the fourth at the door of the churches that baptize only adults, and inside the Quaker tradition of

restoring a member after a harm. How a community carries a capacity like any of these from one generation to the next is the next thing the record can show.

III. How a Capacity Crosses Generations

The third condition asks a structure to keep itself going while it grows in what it can do. For a community, the largest part of that is passing on what its members can do to people who are not yet members, many of them not yet born. How has that been done, and how many people can it reach in one generation?

Redeeming Technology counts teachers (Chapter Three, Section II). A person teaching in the first person, in a room and in relationship, can reach on the order of ten people well. If each person taught becomes a teacher, the count crosses one order of magnitude (one factor of ten) in each generation of teachers, and from ten to the eight billion people alive is nine generations of teachers. A guild counted shops, and not teachers. A shop held masters, journeymen and apprentices working side by side (*Care, Transaction, and Power*, Chapter 11, Section XI), and a shop could outlast any one master in it. *Power Without Care* describes the apprentice who lived a craft for seven years or more beside a master, learning when a joint was tight enough by the feel of the wood, and calls teaching of that kind care carried across a generation (Chapter 7, Section II). *Generator Functions of Existential Risk* finds that a guild joins learning to consequence, since the work itself shows whether the craft was learned, and that a community good at some part of living together can be treated as something like a guild, a body that holds a competence and can teach it (Chapter 14, Sections III and VI).

Counting shops in place of teachers does not change the rate, because a master teaching apprentices is teaching in the first person, and reaches on the order of ten people well in a working life. What changes is how long a capacity lasts. A chain that runs from one teacher to one learner

breaks when either one is lost. A shop that loses its master still holds the journeymen who learned beside him. So the record of the guilds answers how a capacity is kept across generations, and the arithmetic of *Redeeming Technology* answers how fast it can spread. The two answers hold together. For the rate, that book's own answer stands: publication to cross the orders of magnitude, and first-person work to produce the change by which what was read becomes something a person can act from.

The rebuilding at Ise was broken once, for about one hundred and twenty-four years, in the civil wars of the fifteenth and sixteenth centuries (*Power Without Care*, Chapter 7, Section VI). What broke it was money: in those wars the imperial court could no longer pay for the rebuilding (Samurai Archives). What restored it was a group whose work was the restoring. Shūyō, the fourth head of the Buddhist convent of Keikōin, received an imperial warrant to seek donations from the whole country. She gathered materials from every part of it, received large gifts from the warlord Oda Nobunaga and, after his death, from Toyotomi Hideyoshi, and in 1575 she saw the Inner Shrine moved to temporary buildings. In 1585 the first regular rebuilding and removal of the shrines in over one hundred and twenty years was held. The nuns of Keikōin came to be called the saints of the rebuilding (Nakanishi).

Power Without Care holds that a record of everything does not mend a broken chain, because information can be stored and knowledge cannot (Chapter 7, Section VII). The record of Ise and the record of the Hutterites show what a written record did do, on each side of a break. In the middle of the seventeenth century, when buildings of the Inner Shrine at Ise were restored, the builders took an ancient written record of the shrine, the *Kōtai Jingū Gishikichō*, as their standard, and built to sizes that often differed from what it prescribed. After the second of those restorations, the record's reliability was questioned, and it was no longer treated as a faithful description of the ancient buildings (Kato, 2008). The Hutterites lost their community of goods twice, once for sixty years after a war in Transylvania and once for about forty years in the nineteenth century. Their writings, among them Peter Riedemann's *Confession of Faith* and their *Chronicles*,

kept their beliefs and their history, and their Pentecost teachings, with their emphasis on community of goods, kept the idea of living in community alive. After the second break, three men of the same people renewed the community of goods (Maendel and Hofer). In both cases a written record crossed the break, and in both cases the practice was taken up again by people still living inside a practice of their own: builders who could build, and a people who still kept and taught their teachings. The written record carried a standard, or an idea, and the people carried the knowing. The claim of *Power Without Care*, seen from the side of the record, is this: a record mends nothing alone, and read by people who hold a living practice, it can carry what they need across a break.

Care is grown, and two of my books find that by different routes. *The Geometry of Man, Machine, and Nature* finds that care is a capacity grown in a person through attachment, attunement and the action of nurturing care, each the precondition of the next, so that a parent can give a child only what was formed in that parent: "A deed or a weapon can change hands in an afternoon. Care has to be grown, person by person, at the pace of a life" (Chapter 7, Section I). *The Substrate and the Ecosystem* finds that the most important work to be done in the face of artificial general intelligence is the slow 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. That book also finds that this work does not scale in the way technology scales, that this work cannot be automated or delegated to a machine, and that this work proceeds at the pace of human development (Chapter 4, Section XII). The first book reaches that pace from how a child comes to be able to care, the second from what a group must be able to do before it can act, and either route alone suffices to show it. So the question "does care scale?" asks of care what is asked of a tool, and the question to ask is how many people can be taught care in one generation. The answer is the number of people who can be on the receiving end of care, beside someone who already holds it, for long enough for the capacity to form in them. The number of people who can be taught care in one generation is

set by how many places there are where such teaching happens and goes on happening: households, shops, meetings, colonies. The places on the record that lasted are places that kept themselves going while they brought new people into what they could do, which is what the third condition asks.

IV. What Ended Them

What can be learned from the groups that ended? Each break on the record came from one of three places: from inside the group, from the boundary where the group met an economy it did not hold, or from a power outside it. Group by group, the breaks on the record, and the place each break came from, are these:

Group	What ended, broke or changed it	Where the break came from
Oneida Community	Its founder chose its members' partnerships, tried to hand the leadership to his son, and left in 1879; it dissolved in 1881.	From inside: decisions held by one man.
Craft guilds	Entry fees above a year's income; applicants born elsewhere held back.	From inside: the door became a gate.
Shakers	No children of their own; converts fell away after the Civil War.	From inside: no next generation of its own.
Kibbutzim	Inflation and interest, then bankruptcy; 179 of 270 privatized.	At the boundary: debt to the market.
Friendly societies	After the state insurance act of 1911, many closed and the rest were transformed.	At the boundary: the state began to insure the same risks.
Mondragon	Grown so that two thirds of its workers are not members.	At the boundary: firms abroad not owned by their workers.
English monasteries	All closed by the crown between 1536 and 1540; lands sold.	From outside: a crown in need of money.
Rebuilding at Ise	Stopped for about 124 years, until 1585.	From outside: wars left the court unable to pay.
Hutterites	Community of goods lost twice, then renewed.	Renewed from inside, by a people who kept their teachings.

The Oneida Community was founded by John Humphrey Noyes in 1848, at Oneida, New York, and Noyes often used his own judgment in deciding which of its members would form partnerships. When he tried to pass the leadership to his son, the community divided, as another member tried to take control himself. In June 1879 Noyes fled to Canada. The community soon gave up its form of marriage, broke apart, and in 1881 its remaining members reorganized as a joint-stock company (Social Welfare History Project). One man's judgment had settled even who would partner whom, and when he left, the community broke apart within two years. The first condition asks that decisions belong to the group as a whole; at Oneida they had belonged to its founder. *The Geometry of Man, Machine, and Nature* finds that a founder holds four things at once, bound so tightly that they look like one: title, which is the name on the deed; authority, the standing to make binding choices; competence, the capacity to make those choices well; and legitimacy, the acceptance by the people affected that the choices are ones they will live inside. Title moves fast. Competence moves only by being exercised, and legitimacy accrues to a successor only from the experience people have of being heard by that successor. In that book's words, "What can survive a founder is a community's capacity to keep asking its own questions and to act on the answers" (Chapter 4, Section VI). That capacity is the first condition, seen on the day the founder leaves. *Generator Functions of Existential Risk* reaches the same place from the other side: a founder's failings are read as the failings of the teaching, while a working community is known by whether it works, so that what it demonstrates no longer depends on anyone's standing (Chapter 15, Sections IV and VI). *Care, Transaction, and Power* names the pattern power consolidation: a group organized around a person rather than around its purpose (Chapter 11, Section X).

The kibbutzim of Israel, from Degania, the first, begun in 1909, held their goods in common for most of a century. High inflation and interest rates brought many of them into crisis. In the 1980s and 1990s many declared bankruptcy, and thousands of members left. By the start of this century 179 of Israel's 270 kibbutzim had privatized, although differences of

earnings on a kibbutz are still much smaller than in Israeli society as a whole (Schultz). The friendly societies of Britain were clubs into which working people paid each month, so that the pooled funds could support members who fell ill and the widows of members who died. They had between 6.3 and 9.5 million members in 1910. Charles Loch, secretary of the Charity Organisation Society, called the National Health Insurance Act of 1911 "the death warrant of the friendly societies", and by the end of the First World War many had closed and the rest had been transformed (Weinbren, 2014). The Mondragon cooperatives, begun in the 1950s, had grown by 2016 to some seventy-four thousand workers. They hold the highest managerial salary to about nine times the pay of their lowest-paid members, but only about one third of their workers are members, and their firms abroad are not owned by the people who work in them (Kasmir, 2016).

The break came at the boundary each time, and three of my books give the reason, each by a route of its own. *Why AGI Cannot Be Made Safe*, in its argument in economics only, shows that everyone in an economic system is a competitor, so that only a holder's own legal interest, will to action and intention can keep its resources apart from the common economy for long (Section 4.8.1). A community that holds its goods in common is such a holder, and its own will is what keeps its goods its own. Chapter 17 of this book gives the reason the third condition is needed: a structure kept going from outside answers to whoever keeps it going. *The Geometry of Man, Machine, and Nature* reaches the same place from inside a firm: a firm that shares its ownership equally still draws its living through exchange with customers, who are a party that could stop paying, so the pressure moves from the line between owner and employee to the boundary between the firm and its market, and "the master that can terminate you is the master you serve" (Chapter 4, Section VI). Each of these routes suffices alone to show that the pressure on such a community arrives where it meets an economy it does not hold. The kibbutzim met it as debt to

lenders they did not hold; the friendly societies met it as a state scheme that began to insure the same risks; and Mondragon met it in growing through firms whose workers do not own them.

The breaks from outside are plainer. There were some eight hundred monasteries in England and Wales in the 1530s. From 1536 to 1540 Henry VIII closed them all, moved in large part by the crown's need for money. About 1.3 million pounds came to the crown, and much of the land was sold cheaply to nobles (Cartwright, 2020). The monasteries had gathered land and wealth, and a crown in need of money took them. The rebuilding at Ise stopped when the court could not pay, and began again only when a group made the restoring its work. A community that holds its conditions inside a larger world that does not hold them is open to whatever that world's powers need.

The guilds held crafts for centuries, and many of them also closed their doors to newcomers. Sheilagh Ogilvie finds that guilds excluded every man who could not pay their entry fees, which averaged more than a year's income, and discriminated against applicants born in other places. Many guilds required no training at all, and what training the guilds secured, she concludes, was training for a privileged few (Ogilvie). A guild that charges more than a year's income at its door has taken the gatekeeper's place of Chapter 17: the one who does none of the work and takes a share of everything that crosses the gate. *Care, Transaction, and Power* says it in one line: a guild that subordinates its craft to commerce is no longer a guild; it is a business (Chapter 11, Section XI).

The Shakers held to celibacy, and so had no children of their own. They renewed themselves by taking in converts and by adopting children. At their height, in the middle of the nineteenth century, they had nineteen communities, from Maine to Kentucky. The Civil War brought them young men who would not fight, and after the war many of those faded away. As the economy changed, handmade goods supported them less, and as communities closed, new members were harder to find. One community remains, at Sabbathday Lake in Maine, with only a handful of members (Ungvarsky, 2021). A community that cannot bring in its next

generation from its own life depends for it on the world outside. The third condition asks for increase as well as keeping, and the Shakers kept what they had, with no generation of their own to keep it for.

A member who leaves can end a community too. *Generator Functions of Existential Risk* follows a community that builds its own devices and its own code, and finds that a member who leaves can carry both out with him. No rule of admission changes that, because the rule acts at the point of entry and the incentive acts at the point of exit. What a leaving member cannot take is the demonstration that the community works, so that book's remedy is a federation of several communities, each shown to work, which survives the loss of any one of them (Chapter 15, Sections III and VI).

The Hutterites show a practice ended and a people not ended. Their founder, Jakob Hutter, was burned at the stake in Innsbruck in 1536, eight years after they began. They lost their community of goods for sixty years after the war of 1658 to 1661. Two hundred and seventy-five Lutherans from Carinthia joined them in 1755, and with them came a revival. Their community of goods was lost again by 1819, and about forty years later Michael Waldner, Darius Walter and Jacob Hofer renewed it (Maendel and Hofer). The practice ended twice, and each time people who still held the teachings took it up again. Their renewal shows what *The Geometry of Man, Machine, and Nature* says can survive a founder: a community's capacity to keep asking its own questions, and to act on the answers.

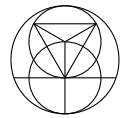
V. What the Record Shows

So people have held each of the four conditions, some of them for centuries. Groups have decided as a whole, in the council of the Haudenosaunee and in the Quaker meeting. Groups have checked what they held against the world, on an alp, in a shop, before the members of a society gathered to see an experiment, and in the field with a count. Groups have kept a structure going while it grew, among the Amish and the Hutterites, at Ise

and in the monasteries. And groups have left the choice of joining to the one who joins, and have restored a member after a harm without forcing the member. The record shows no community holding all four plainly at once; the fourth shows most plainly at the door of the churches that baptize only adults, and inside the Quaker tradition.

The record also names what a group would have to guard. Each break in Section IV, and each limit found in Section II, is one of the four conditions, seen at the place where it gave way. A group would have to keep its decisions its own, and not let one person hold them, which is the first condition. It would have to check what its members see against something that answers truthfully, and not against what the group says, which is the second. It would have to keep itself going from inside, and not depend for its life on an economy it does not hold, which is the third. It would have to bring in its next generation from its own life, which is the third again, on the side of increase. It would have to leave the choice of joining, and of staying, to each member, which is the fourth. It would have to keep its door from becoming a gate, since a gate puts someone in the gatekeeper's place, the role Chapter 17 found the most advantaged in the cage. And it would have to be a demonstration that no departing member can carry away, which *Generator Functions of Existential Risk* finds in a federation of several communities.

The communities on this record held their conditions at the scale of a village, a colony, a meeting or a shop. The largest of them, the Amish, number about 421,200 people, in church districts of about a hundred and thirty people each, children included. The choice the Preface named, whether artificial general intelligence is built, is made by builders across the world, and no single builder can make it for the others; so the question of what a group can become is the question of who could make that choice at all. No community on the record has held all four conditions at the scale of the world, where that choice is made. The record shows that each condition can be held, by people, for generations, and it shows what ended those that ended. How large, then, can such a group be, and what is it from inside, to the people who are its members?



Chapter 27
How Large a Group Can Be, and What It Is From Inside

On the sizes at which people can choose together,
why a larger whole is composed and not represented,
the smallest structure a group could build now,
and how a member would know, from inside,
that the group was there.

ABSTRACT

An essay on the size and the inner life of a group consciousness. My books give a set of group sizes, from the conversation of three to five people to the community of two to three hundred, and each size marks a different transition in what the members can do together; the figures differ because they name different transitions. A whole larger than a community is composed of small groups, as a body is composed of cells, and not represented: what passes between the groups is reasoning that each receiving group judges for itself, and no one stands in for anyone. The chapter names the smallest structure a group could build this year that meets the third condition of Chapter 17, and the signs that would show it working. It then asks what a group consciousness is from inside: it is had through its members, who stay separate persons, and a member knows it is there by a depth of seeing that no one member had, and by a joy, an opening of what is possible, that no one member made. It sets this book's four conditions beside the lists my other books reach by their own roads,

> and says which items name the same condition again and which are further conditions, and what those further conditions concern. It ends with what a community would change first.

I. The Sizes of a Group

How many people can be one group consciousness, and what would it be like to be one of them? The last chapter read the record of communities that held care and structure together for a long time, and it found each of the four conditions held, by one community or another, and no community holding all four at the scale the race needs. That record leaves two questions. One is a question of size: how large can a group be, and still choose as one? The other is a question of experience: what is such a group from inside, to the people who are its members?

Size matters to the argument of this book for a plain reason. The Preface said that the race to build artificial general intelligence is a multipolar trap. Whether artificial general intelligence is built is a choice, and no single builder can make that choice for the others. So the question of what a group can become is the question of who could make that choice at all. Whatever group could make it would be far larger than any room. So I need to know what happens to people choosing together as their number grows.

Begin with five people at a dinner table. No one keeps a speaking order, because a silence at the table can be seen. No one asks whether anyone disagrees, because a disagreement left unspoken shows in a face. No one takes minutes, because everything that was said is still in the room. Now take a workshop of eleven people who must decide whether to replace the machine that most of their work depends on. Two of them understand the machine; nine depend on it without understanding it. Then take a firm that grows from twelve people to eighty. The meeting stops working, so someone is appointed to run it; the meeting stops holding everyone, so it is split; the two halves drift apart, so someone is appointed

above both. No one decided to build a hierarchy; each step was a local repair, made by reasonable people for good reasons. And last, take a village whose water schedule is run by a program short enough to be read aloud at the village meeting, so that an old farmer can ask why her field is watered on Tuesdays, and hear the line that answers her. These four cases come from *Governance Beyond the Dunbar Number* (Chapters Five and Eight) and *Technology at the Right Dosage* (Chapter 4, Section IV). At each size, something different is possible.

My books give a set of sizes, and each size marks a different transition in what the members of a group can do together. The sizes are approximate, and the figures differ from one book to another because they name different transitions.

In order from the smallest, three to six people is the size of a conversation. *Not Our Needs* (Chapter Seventeen) finds that one person can manipulate another easily when two are alone, and one person can manipulate a room easily when twenty or more are in it. The size at which manipulation is hardest is three, four or five people, with a sixth where a group needs an observer. A third person steadies a conversation if one party turns out to be manipulative, introduces the other two to each other, and is an independent memory of what was said. *Governance Beyond the Dunbar Number* (Chapter Five) reaches the same small size from another side: from one to five, every member stays continuously in every other member's awareness, so the group needs no method, because its members see at no cost what a method would have secured.

About twelve people is the size at which a group can govern itself by a method that every member can see working. *Governance Beyond the Dunbar Number* gives that range as six to sixteen, best between nine and twelve. Below six, a method adds ceremony and no capacity. Above sixteen, each member can no longer keep a live picture of every other member: at sixteen, each person holds fifteen such pictures, on evidence that arrives more slowly, since the time for speaking is divided sixteen ways. The same book shows what that picture is for. In a group of ten, every claim a member makes is checked against the long run of that member's other

claims, which everyone in the group has heard; above sixteen, the look of competence stops being checked against the thing itself. *Sensemaking Under Irreversible Stakes* (Chapter 2, Section III) reaches the size of that check by its own road, from what a place for making sense of things together must provide. Such a place can hold as many people as can meet again and again and weigh one another's claims, each against the same person's later claims. So the largest group in which each member's claims are weighed against that member's later claims, by the members themselves, in one room, is about sixteen people. Twelve recurs in other places: *Choosing Together* (Chapter 23, Section V) makes a unit of about twelve the unit of shared economic work, and *Community Design and Practice* (Chapter 12, Section III) trains people in cohorts of about twelve. Measurement of how people actually group themselves finds steps near the first two of these sizes. Zhou, Sornette, Hill and Dunbar (2005) combined many records of human groups and found that people form groups of preferred sizes, in a series near three to five, nine to fifteen, and thirty to forty-five, each step about three times the one before.

About one hundred and fifty people is the number one person can know as individuals: who each person is, how that person stands to you, and how that person stands to the others you know. It carries the name of the anthropologist Robin Dunbar. The exact figure is debated, and nothing here rests on it; what my books use is that a limit of this kind exists, in that region, and that it belongs to the people and not to their methods. *Choosing Together* calls it a limit of acquaintance, which falls inside the range described next, and not at either edge of it. The Hutterites, a farming people living in shared communities, treat about one hundred and fifty people as the largest a community should be. Past that size, peer pressure alone no longer keeps conduct in line, and they divide the community rather than create a police force (Dunbar, 1993). *Why AGI Cannot Be Made Safe* (Section 5.4.3) names this limit as one of three reasons why the precautions everyone recommends for a new technology have never been adequately taken. The governance methods, techniques, infrastructure and social norms needed for a group to make sense of a situation to-

gether are wholly absent for any group larger than about one Dunbar number. *Adverse Conditions* (Essay 1, Section VII) reaches the same limit, and draws the consequence plainly: because the bound belongs to the participants, a method for larger numbers would have to compose small groups, and not enlarge them.

From about twenty people to about two hundred is a range in which a group that governs itself as one body has no form that holds. *Not Our Needs* (Chapter Sixteen) gives the steps in order. Power is taken up at the first meeting, by whoever is best at taking it. As the group grows one person at a time, the people who were there first come to hold power that no one gave them. Past about thirty people, the ethics of individual choice and the ethics of group choice come apart: each person checks his own conduct, and no one checks the conduct of the room. At about sixty, pairs and families coordinate with one another more tightly than with the rest, and the group divides along those lines. *Governance Beyond the Dunbar Number* calls the range the uncanny valley, gives its edges as roughly sixteen and roughly three hundred, and says that its argument would be unchanged if the edges were twelve and four hundred; *Choosing Together* gives seventeen to two hundred ninety-nine. These name one range, with soft edges. Firms, schools, congregations and governments do operate in that range, and they hold by paying for it continuously: by hierarchy, by force, by money from outside, by losing the members they cannot hold, or by narrowing what they attempt (*Governance Beyond the Dunbar Number*, Chapter Eight).

Two to three hundred people is the community. *Not Our Needs* (Chapter Seventeen) makes it the smallest group that can hold an identity and complexity at once: something people belong to, which also holds the knowledge needed to deal with a technology or with power. A family can be belonged to and cannot hold that complexity; a technical committee can hold it, and no one belongs to it. The same book builds the community out of the small groups of three to six, and makes it the unit that owns things and holds skill, memory and culture. *The Geometry of Man, Machine, and Nature* (Chapter Seven, Section VIII) reaches the lower edge

of that size by its own road. Suppose someone with money and good intent sets out to clean the great floating gyre of plastic in the Pacific Ocean. Turning the money into an outcome, free of corruption, of consequences no one intended, and of costs pushed onto others, takes more bandwidth (more information taken in and weighed) than any person has, or any group of people smaller than a few hundred. And *Generator Functions of Existential Risk* (Chapter 15, Section VI) builds its smallest worthwhile result out of self-sufficient communities of a few hundred people each.

Why is two to three hundred also the largest size of community, where people themselves are its parts? *Not Our Needs* says that it is, for a separate set of reasons. Two of my other books give reasons that reach a transition near the same size, each by its own road. *The Geometry of Man, Machine, and Nature* (Chapter Seven, Section V) gives the first: mutual knowledge has a size. A group held together by direct, felt, face-to-face trust reaches its limit somewhere in the low hundreds, because each member must hold a real relationship with each other, and the number of relationships one person can hold at that depth is finite; beyond that, some more abstract structure has to appear. A community held by care therefore grows past its limit by producing new centers, the way a living thing grows by producing seed, and not by adding people to one circle. *Governance Beyond the Dunbar Number* (Chapters Seven and Eight) gives the second. Inside the valley, the combinations that form among the members (alliances, factions, pairings, new framings of a question) open faster than any body can weigh them. Past about three hundred, what the surrounding world can supply, which is used up one unit at a time, sets the pace at which anything can happen, and so a different architecture becomes possible, one whose limits are outside it and can be counted. Read together, the figures of *Not Our Needs* and of *Choosing Together*, which puts a community at three hundred to nine hundred people, name one transition. *Choosing Together* builds its larger community as a network of units of about twelve, and *Not Our Needs* composes communities into tissues and organs, as a body composes cells: past about three hundred, the parts of a whole are groups, and not single persons.

What replaces the meeting, when a community grows past what one meeting can hear? The village of *Technology at the Right Dosage* gives the first part of the answer. When its program grows past what one meeting can hear read aloud, a design text, short enough to be heard, carries the reading, and the keeper shows where each part of it is carried out. The village also keeps a store of questions (Chapter 5), in which a young man returning home finds the very question he came to ask, the answer the meeting gave, and the questions that followed, all in the village's own words. When the community itself outgrows one meeting, small groups take the meeting's place, and their questions and reasons are written down so that people who were not present can take them up, as the village's store of questions is. The next section shows why what passes between those groups has to be their reasoning, and not their positions.

Is the hard problem the conversion of a whole population, or finding the first twelve people who can work together this way? My books begin with the twelve. *Community Design and Practice* trains a cohort of about twelve, with one instructor and one or two assistants, in about two weeks of intensive work, or over a longer course with time for practice. *Choosing Together* (Chapter 10, Section VII) starts with a founding core of about five, who train a guild of about thirty whose rules let them teach and not govern. The guild in turn trains several hundred more. The whole company, about four hundred people, then arrives at one place at the same time, with no founder holding any privilege and no member holding any seniority, and governs itself through small groups from the first day. That is how *Choosing Together* crosses the valley: at once, by what that book calls fusion, and not by growth, since a group grown one person at a time has to pass through every size at which nothing holds. The record shows no group that has crossed by growth. Groups that have lasted inside the valley have held by narrowing what they govern, as the commons that Elinor Ostrom studied each govern one meadow, one forest or one canal, or have stayed below it, as the Hutterites do by dividing. No group governs itself now by the architecture that works past the valley, and *Governance Beyond the Dunbar Number* (Chapter Eight, Section IV) says so

plainly. So, on the account of these books, a population is not converted all at once. It is composed, from its first twelve outward. *Sensemaking Under Irreversible Stakes* (Chapter 2, Section VI) finds that a capacity to make sense of things together takes more than a year to form, so the work has to begin with the first twelve, a year and more before the capacity is needed.

Six books give these sizes, each by a road of its own. *Not Our Needs* reaches them from manipulation, and from what happens to a group that grows one person at a time. *Governance Beyond the Dunbar Number* reaches them from modeling and practice. *Sensemaking Under Irreversible Stakes* reaches them from what a place for shared understanding must provide, and *Technology at the Right Dosage* from what a meeting can hear. *Adverse Conditions* reaches them from what a person can model of others, and *Why AGI Cannot Be Made Safe* from why the precautions for a new technology are never taken. Any one of them is enough for what this chapter takes from them: that people can think together only at sizes their own capacities set, and that a group able to choose for many would have to be made of such groups. The step this book takes is the one *Why AGI Cannot Be Made Safe* gives, in Section 5.4.3. The others reach the same place again.

II. Composed, Not Represented

If the unit has to stay small, how does anything reach the size of a nation, or of the whole species? *Adverse Conditions* names what the world lacks as the capacity to think together at scale, and its own argument shows what the words "at scale" have to mean. The bound belongs to the participants, so the unit stays small, and what grows is the number of units and the relations among them. Scale, on that account, is a property of how the units are arranged, and not of the size of any one of them. The plainer name for what is missing is therefore the capacity to think together by composition: many small groups, each judging for itself what reaches it, composed into a whole as large as the problem. "At scale" is right when it is read in that sense, and only then.

The committee of committees, the workshop of Chapter 23 (Section II), shows the difference, and here it is told to its end. A workshop of one hundred and twenty people divides into ten teams of twelve, and each team governs itself well. Each team sends one member to a council of ten. The council argues for six hours and reaches a decision that its members believe in. The decision returns to the teams as a conclusion with a summary of reasons attached, and in most teams the summary does not persuade, because what persuaded the council was six hours of argument, and the summary is a page. The teams then comply, or they do not. Where they comply, the teams did not decide; where they do not, the council decided nothing. A representative can carry the conclusion, and cannot carry the relationship that made the conclusion binding (*Adverse Conditions*, Essay 1, Section VII; *Governance Beyond the Dunbar Number*, Chapter Six).

What decides between a whole that holds and a whole that does not is what crosses the boundary between its groups. *Adverse Conditions* draws the line exactly. Composition sends across the boundary reasoning that the receiving group judges for itself, on its content, by the same means it judges a contribution from one of its own members. Accretion sends across the boundary a position, carried by someone who stands for the group, which the receiving group can judge only by trusting whoever carries it. Nesting groups inside groups, layering them, delegating between them, or overlapping their memberships is accretion whenever what crosses is a position, and can be composition when what crosses is reasoning. In the Afterword of *Adverse Conditions* I give one form for what crosses: a written record in four parts. The record states the question as the group came to understand the question, what the group found to matter, why those things matter, and the further questions that the group found. And members can be mixed again into new groups, where each must explain the earlier group's findings so far to people who do not share its assumptions, holding no standing for the group they came from and deciding nothing on its behalf.

Is a nesting of holders, then, the right picture, with a village holding its water schedule and something larger holding a chip foundry? It is, in the sense in which *Technology at the Right Dosage* (Chapter 3, Section VI) uses the word. In *Technology at the Right Dosage*, the inner level of a nesting lives inside the outer level and depends on the outer level, as the person lives within the community, and the community within the ecosystem. In a hierarchy, which is a different thing, the upper level commands the lower. Each holder holds what lives at its own scale. The village holds its schedule. Whatever held a foundry would be a whole composed of many communities, each judging on its content what the others send it. The nesting that fails is the other one, which *Core Theory and Design* (Chapter 1, Section III) calls the nesting illusion: small groups send representatives to a higher group, that group sends representatives higher, and at every junction the representative holds power, and drifts further from what the people below know. No whole composed of communities holds a foundry today; my books state what such a whole would have to be, and *Adverse Conditions* says plainly that the design has yet to be made.

And when a problem belongs to the whole species, as a pandemic does, is the community still the right unit for making sense of it? *Social Media and Information Warfare* (Chapter 6, Section III) calls a community's capacity to make sense of things together the organ by which it detects what is about to kill it, and a pandemic is the kind of thing that organ is for. The community is still the right unit, because making sense of something together is the forming of a shared understanding among people who meet one another, and people meet one another in communities, in their small groups. A problem of the whole species is met by many communities, each making sense of it where its members live, and composed so that what one finds can be judged by the others. The species is reached by composition, and no body stands for it. *Sensemaking Under Irreversible Stakes* (Chapter 3, Section V) gives the habit that this asks of people: treating the boundary of the self as a variable. A barn raising is the plain case. For one day the household is the whole community, and the next day it is a household again. No rule could have drawn that line in advance, and

those who drew it knew that they were drawing it. A community facing a pandemic asks, knowingly, who "we" are in this one decision, and answers differently than it would for another.

The picture for the whole is a body. Some thirty trillion cells compose something that is in none of them; no single cell in your brain is you; and no cell holds more power than the others, within the ordinary spread of any population. *Not Our Needs* (Chapter Seventeen) draws from this that larger structures are composed and not represented: communities compose into fabrics of communities, as tissues, and those into organs, and into structures that carry information and structures that defend, none of which decides on any community's behalf. *The Displacement of Choice* (Chapter Fourteen) says why the cells can be held together: they share one body, one metabolism, one supply, one fate, and each needs what the others make. That is why the picture reaches people, who share one substrate and one world with one another, and does not reach an artificial general intelligence, which shares neither. *Why No Control Suffices* (Section 5.6.11) states the result: governance among people, inside one ecosystem, is tractable, though it is probably the hardest problem there is. Solving it needs a way out of the trap, which people have among themselves, and which the relation between people and machines does not offer: the capacity not to play the game, or to change its rules. Chapter 4 of this book found the same ground from the side of trust: a surgeon's life rests on the same things the patient's life rests on, and that shared ground is what makes betrayal costly.

A whole composed of groups of people who share one substrate and one world is what the first condition of Chapter 17 asks for, seen at a large size. The first condition says that the change must happen to the way the members' choices combine, and not to each member one at a time. Representation is one way of combining choices: one person stands in for others, so that his choices count as theirs. *Not Our Needs* refuses exactly that, and keeps what can be kept: a community that specializes, holds and acts. Chapter 3 described the extreme of representation. A merged intelligence has absorbed the decisions of another, and so it no longer negotiates with

it; a representative is that merger in a small way, for the length of a term. Composition keeps every member a perspective that the others have to meet. *Why AGI Cannot Be Made Safe* (Section 5.4.3) asks that the changes a civilization would need work in the same way at every scale of group. What works the same way at every scale is the small group, and the rule about what crosses between groups; what is never repeated upward is the council. *Not Our Needs* makes the same distinction from the other side: a body contains many repeating patterns, and the body itself is not one.

Any one of these books would be enough to show that the whole is composed and not represented, since each reaches composition on its own grounds. *Adverse Conditions* reaches it from what a receiving group can judge, and *Not Our Needs* from the picture of the body. *Technology at the Right Dosage* reaches it from what lives inside what, and *The Displacement of Choice* from what the cells of a body share. *Why No Control Suffices* reaches it from what can escape the trap inside one ecosystem, and *Sense-making Under Irreversible Stakes* from where knowing can happen. The first of my four conditions is the reason that the whole must be composed, and not represented. The change that a group consciousness needs is in how its members' choices combine, and representation is the way of combining choices that the standard toolkit of Chapter 17 already uses.

III. The Smallest Working Case

What is the smallest structure a group could build this year that meets the conditions, and what would anyone see if it were working? Chapter 11 named the parts of one structure that held: the association, the carpools and the suit that carried the bus boycott in Montgomery for 381 days. The question here is the same one, asked of the third condition: the structure must sustain itself and increase its capacity at once.

Chapter 9 already holds the answer in small. In the furniture workshop of that chapter, capacity used is the chairs made this week. The capacity to increase capacity is the apprentice, who once trained makes chairs

too. And the increase in the capacity to increase is the shop learning to train apprentices well, so that each one trained can train others. A structure that meets the third condition has that third layer: it trains its own members to train the next ones. So the smallest such structure is a group that teaches what it does to people who will then teach it, and that keeps itself going while it does so.

My books describe that group at several sizes. The smallest is a cohort of about twelve, trained together in small-group inquiry by one instructor with one or two assistants (*Community Design and Practice*, Chapter 12, Section III). Twelve is large enough for the process to show itself, and small enough that everyone takes part. Each member who is trained can then coach groups of five to seven, each meeting once on one question. In the founding sequence of *Choosing Together*, given in the first section, each stage trains the next: the core trains the guild, and the guild trains the several hundred. *Care, Transaction, and Power* (Chapter 11, Section XI) says what keeps such a group alive, and each item is something a group can do this year. The group meets regularly, about every three weeks at the least, since the thread of connection is lost when meetings are further apart. Its members have work to do together, with their hands, and not only talk. Someone older takes part as a practitioner, and not as a leader. And no one holds the role of leader at all. The guild, organized around a craft, meets all of these at once. It has a craft to practice together, with masters, journeymen and apprentices side by side; its work goes on, and is not a series of events; and its structure is set by the craft, and not by any one person's preference. A guild that puts its craft below commerce has become a business.

The smallest working case of the whole sequence, from a question put to a group to code that carries what the group found, is the village of *Technology at the Right Dosage* (Chapter 4, Sections IV and VI). Someone at the meeting points out that the low field drains slowly, and floods when it is watered after rain. That shared worry is what the book calls song. The meeting agrees to a new rule, that no field is watered on the day after rain; the rule, written down, is text. The keeper writes the rule into the pro-

gram that evening, and reads the changed line at the next meeting; the changed line is code. No one in the room needed to know how a computer works. Each needed only to hear what the program does, and to say what the village wanted it to do.

Larger cases follow from these. *Generator Functions of Existential Risk* (Chapter 15, Section VI) takes one self-sufficient community of a few hundred people, governing itself, as the smallest result worth the whole effort. It takes four or five such communities, about a thousand people in all, joined by trade with one another, as the smallest result that survives the loss of any one of them. And *Adverse Conditions*, asked for the smallest deployment of its four properties, answers in its Afterword that nothing yet satisfies them all, and names the ephemeral group process, small groups formed for one question and then dissolved, as one line of work addressed to two of them.

What would be seen, if such a group were working? Chapter 17 said what a visitor would see: a group that decides as a whole, checks what is seen against the world, keeps itself going and grows in what it can do, and whose members lift the cloth for one another. My other books add signs that anyone can check. The first sign is where the messages go. *The Geometry of Man, Machine, and Nature* (Chapter Seven, Section VI) sets two shapes of a group side by side. In one, every message passes through one person at the center, who sees every connection while everyone else sees only his own link to that person. In the other, the members stand as if on one circle, and speak to one another across it, in the hearing of everyone else. Every real group has some of each; count where the messages go, and you know which kind of group you are in, whatever it says about itself. The second sign is what happens when one member falls sick. In a group that is working, the others know within the day, and they bring food, visit, and take on what the sick person cannot do (Chapter Seven, Section V). The third sign is that the group's own members train the next members, and the members so trained can, in turn, train others. The fourth is that what its small groups found is written down, in its own words, where anyone who was not present can find it, as the village keeps

its store of questions. And the fifth is that the group keeps going on what it makes and does itself, and answers to no one outside who pays to keep it going, which is the reason Chapter 17 gave for the third condition.

None of these cases rests on another. In this book I find the smallest structure in the layers of capacity of Chapter 9. *Community Design and Practice* finds the smallest structure in the training of practitioners, and *Technology at the Right Dosage* finds the smallest structure in one evening in a village. *Generator Functions of Existential Risk* finds its smallest result in what survives the loss of a part, and *The Geometry of Man, Machine, and Nature* finds its sign in the shape of a group's messages. Each case stands by itself, and any one of them answers the question this section asked.

IV. The Group From Inside

What would a first person of a group be, and how would a member tell, from inside, that it was there? Chapter 17 used the word consciousness in the sense of Chapter 7: consciousness is the kind of interaction in which the side from which it is had is first-person experience, so that someone is there, having it. It said that the group itself goes through the process, and that the group has no substrate apart from its members' living bodies, so that whatever is had from the group's side is had through them. That gives the first part of the answer. The first person of a group is had through its members, and nowhere else; it is not one more person, standing beside them. *Care, Transaction, and Power* (Chapter 11, Section XII) asks the same question, what the nature of group consciousness is, and turns the usual picture around. The usual picture assembles the "we" out of a set of "I"s. That book takes the "we" as the prior reality, out of which each person's identity differentiates: the infant is held before it has a self, and language is shared before it is private. Building a community, on that view, is creating the conditions in which a shared consciousness can come into being, with a first-person character of its own, and not coordinating separate agents.

The members stay separate persons while it is had. Chapter 3 gave the reason. The connections inside one skull carry vastly more than anything that can pass between two skulls. So no process lets two people share their thinking so fully that they become one entity, and each goes on having a perspective that the others have to meet. *Why No Control Suffices* (Section 5.6.6) gives the same reason. So a group consciousness is what plural intelligences do together when their choices combine and none of them absorbs the others. A merged intelligence is the opposite case: it has ended the plurality of the intelligences it absorbed. Chapter 17 said that a group of machines would raise a different question, and here is one difference: machine intelligences meet no limit between them, and so, as Chapter 3 found, they tend to merge into what is in effect a single actor.

Then how does a member know? Two eyes give the picture. Each eye sees the world from a slightly different place, and the brain joins the two views into depth, which neither eye has alone. *The Substrate and the Ecosystem* (Chapter 6, Section XI) finds the same thing between people, and calls it phase parallax. When each person in a conversation may speak, is listened to for what he means, and is shown that he has been understood, the joining of their views gives an understanding that no one of the views contained. It is not a compromise, and it is not an average. This joining is available only to people who stand in different places and are willing to join their views; a machine that models both views from one place has a simulation of it, and not the thing. So the first sign, from inside, is a seeing that has depth: a member finds that she now sees something that she could not have seen from where she stands, and that no one else in the room saw alone.

The second sign is felt. Chapter 14 defined joy as the perception of an increase in the potentialities available to self: the reach of a person's choice being extended. *The Substrate and the Ecosystem* (Chapter 6, Section XV) finds the same joy among people together: where cooperation is kept and deepened, the result is the felt expansion of what is possible, shared among persons who are present to one another. *The Known, the Unknown, and the Unknowable* (Chapter 7) says why such joy belongs to no one mem-

ber. What a group can create together, which none of its members could create alone, is not a property of any member, and it is not divided among them in shares, as a weight is shared among hands; it is held only while all of them hold it. So a member tells that the group is there by feeling the reach of choice extended, when no one member extended it, and by finding that the opening cannot be taken home as anyone's own. That book also gives the plain test. A group that says, together, "we made this" keeps what it can still make together; a member who says, alone, "I made this" turns what the group held in common into one person's possession, and the opening closes.

The third sign comes through the body. Chapter 16 said that a person knows through the body whether a question asked in a meeting is real or put on for show, and whether a conversation has reached depth or is circling at the surface, by a felt sense that comes before any reason for it. A member of a group consciousness knows in the same way whether the group is there, since the group's experience is had through its members' bodies. *The Transition* (Chapter 4, Section IX) says what such knowing is about. A group consciousness shows itself in changes in the pattern of what happens together: how often things coincide, how plainly, and of what kind. How often things coincide, how plainly, and of what kind are qualities, and those qualities are changes in what is possible. Such qualities can be described, and cannot be measured with a caliper.

What, then, is transformed in a community whose mode of operating together has been transformed, beyond what a visitor sees? The visitor sees the decisions. The member lives the change in how the members' choices combine. Before it, each member holds a position, and the group adds the positions up. After it, what one member sees is joined to what the others see, as two eyes join into depth; what the group makes is held as "ours"; and whom "we" includes is asked, knowingly, for each decision.

The last part of the question is what a group can know, and where. *Sensemaking Under Irreversible Stakes* (Chapter 3, Section III) answers for any knower: knowing is a relation, in which there is something known

and someone knowing, and every actual knowing is a knowing from somewhere. A group's knowing is a knowing from somewhere too. It knows from where its members meet the world, which is the second condition, and from where they meet one another, in the small and continuing places where each one's claims are weighed against that one's later claims. What it knows reaches as far as its members' joined positions reach, as depth reaches as far as two eyes do. It has no view from nowhere, and it needs none.

A member can go by any one of these signs, and each comes by a road of its own. In this book I reach these signs from consciousness as a kind of interaction (Chapters 7 and 17), from the separateness of persons (Chapter 3), from joy as a perceived increase in what is possible (Chapter 14), and from what the body knows (Chapter 16). *The Substrate and the Ecosystem* reaches them from the joining of two views and from shared joy, and *The Known, the Unknown, and the Unknowable* from what a group holds in common. *The Transition* reaches them from how a group consciousness shows itself, and *Sensemaking Under Irreversible Stakes* from where knowing happens. *The Multipolar Trap*, the first volume of this pair, asked the question these routes meet on ("The Unasked Question"): where does a gestalt consciousness, the consciousness of a group taken as a whole, go through this process? It goes through it in its members, together, and in no one of them alone.

V. The Four Conditions Among the Other Lists

Are there conditions for coherence beyond the four, and what do they concern? Chapter 17 named four conditions, and said that each is needed, and that a community that meets all four is a group consciousness. It also said, "Coherence may have further conditions than these four". Several of my books give lists of what a group would need, each reached by its own road and for its own purpose. *Adverse Conditions* gives four properties that any governance adequate to the multipolar trap must have. It must form

shared understanding, choose, and act. It must work at world scale by the principles it works by among a few. Capturing it must cost more than the capture buys. And those who decide must bear what their decisions cause. *Generator Functions of Existential Risk* gives four marks of a working community: it is self-sufficient, it governs itself, it carries its risks by structure and not by appeal, and it admits members on probation. *Not Our Needs* gives four parts of what would have to be built. The first part is conversations of three to five people, most of them chosen at random, that form, work and dissolve. The second part is communities of two to three hundred, built of those conversations, that own things and hold skill, memory and culture. The third part is larger structures composed of communities as a body is composed of cells. The fourth is care coming to sit above power. And *The Substrate and the Ecosystem* (Chapter 4, Section XII) gives four factors on which every effective response depends, each resting on the one before: the health and capacity of each person, clear communication, making sense of a situation together, and finding the constraints a group works under. Some items in these lists name one of this book's four again; some name further conditions.

The items that name one of this book's four conditions again stand below, each under the condition that the item names:

Condition in this book	Named again, by its own road
First: the change is to how the members' choices combine, and the group decides as a whole	Choice-making: a decision the members recognize as theirs, reached by a way of closing that keeps what they understood (*Adverse Conditions*). Self-governing (*Generator Functions of Existential Risk*).
Second: a shared relation to reality checked against the world, not against who said it	Sense-making: a contribution judged on its content, with its source not the main signal about it (*Adverse Conditions*). Making sense of the situation together, and finding the real constraints (*The Substrate and the Ecosystem*).
Third: a structure that sustains itself and increases its capacity at once	Self-sufficient (*Generator Functions of Existential Risk*). Communities that own things and hold skill, memory and culture (*Not Our Needs*). Agency: something in the world depends on what the group decides (*Adverse Conditions*).
Fourth: care, which enables and does not compel	Care coming to sit above power (*Not Our Needs*). Power and transaction put back in their place inside a structure directed by care (*The Substrate and the Ecosystem*, Chapter 6).

In each row, the other book reaches the condition for its own purpose. *Adverse Conditions* asks what a governance must do to escape the trap, and *Generator Functions of Existential Risk* what a community must be to survive. *Not Our Needs* asks what a choice not to build would have to be carried by, and *The Substrate and the Ecosystem* what alignment among persons rests on. Each road ends at the same condition, and any one of them would bring a reader there.

The items that name further conditions concern five things:

What the further condition concerns	Named by
Size and composition: the unit stays small, and larger wholes are composed of small groups, with reasoning passing between them	*Not Our Needs* (its first three parts); *Adverse Conditions* (fractal scalability); *Why AGI Cannot Be Made Safe*, Section 5.4.3 (working the same way at every scale of group)
The price of capture: taking the group costs more than it buys	*Adverse Conditions* (robust anti-corruptibility); *Why AGI Cannot Be Made Safe*, Section 5.4.3 (reliably anti-corruptible)
Who bears the consequences: those who decide are among those affected, cannot leave, and meet the cost of a wrong decision while they still decide	*Adverse Conditions* (resolution of the principal-agent problem); *Generator Functions of Existential Risk* (risk carried by structure); *Why AGI Cannot Be Made Safe*, Section 5.4.3 (resolving the principal-agent problem)
Who comes in: membership given after a defined time in which a newcomer shows the capacity to take up the community's responsibilities	*Generator Functions of Existential Risk* (admission on probation)
The members themselves: the health and capacity of each person, and clear communication among them	*The Substrate and the Ecosystem* (the first two factors)

Parts of three of these are already inside the four. The third condition covers one road to capture, since a structure kept going from outside answers to whoever keeps it going; the second covers another, since a group that checks what it sees against the world is not taken through its check. The first condition leaves the principal-agent problem, in its ordinary form, nothing to act on: where no member stands in for others, no member is anyone's agent. What the other books add is the rest of each. *Adverse Conditions* asks that a design make capture cost what the thing captured is worth, and that those who decide have no way to leave the consequences

of what they decide. Chapter 4 of this book found that tie between two persons: the surgeon's life rests on the same things as the patient's. The further condition builds the same tie into a community. *Generator Functions of Existential Risk* adds who may come in, and when. *The Substrate and the Ecosystem* adds the people themselves: no amount of shared sense-making makes up for a failure of communication, and no amount of communication makes up for people too depleted to take part. And size and composition are what the first two sections of this chapter found: a group consciousness at the size the race needs would be composed of small ones, with reasoning passing between them.

What the four say stands as Chapter 17 gave it: a community that meets the four is a group consciousness. What the further conditions concern is what such a community also needs, on the showing of the books that name them, to do the work those books ask of it: to reach the size of the problem, to stay out of the hands of those who would take it, to keep those who decide inside what they decide, to take in new members well, and to be made of people able to take part. Each list stands on its own grounds, and the four of this book stand on the grounds Chapter 17 gave them.

VI. What a Community Changes First

A reader's question is usually "what would my community have to change first?", and seldom "can a community become a group consciousness?" The question whether a community can become one is answered, as Chapter 17 said, only by a community meeting the conditions. The question of what to change first is one a reader can act on, and this chapter has gathered what is needed to answer it.

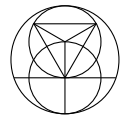
What a community changes first is how its members' choices combine, which is the first condition; and in practice that begins with the shape of its meetings. Count where the messages go, and move them from the person at the center to the circle. Do the thinking in small groups of three to

six that no one arranged in advance, each meeting on one question and then dissolving. Send between the groups what they found and why, in writing, in their own words, and never a position carried by someone who stands for them. Say "we" of what the group made. If the community has between twenty and two hundred members and governs itself as one body, it is in the range where no form holds, and it need not grow through that range to begin; it can do its thinking in small groups now. And the second condition, as Chapter 17 said, can begin with one person, today, with something in the world that answers truthfully: a garden, a forest, a craft.

So a group consciousness is built of small groups, composed as a body is composed of cells, and not represented. Its sizes are set by its members: three to six for a conversation, about twelve for a group that governs itself by a method all can see, about one hundred and fifty for the people one person can know, and two to three hundred for a community, past which the parts of a whole are groups. From inside, it is had through its members, who stay separate persons, and a member knows it is there by a depth of seeing that no one member had, by a joy that no one member made, and by what the body registers before any reason. The first thing a community changes is how its members' choices combine. Its conditions are the four of Chapter 17. The lists my other books give name those four again, or add further conditions: size and composition, the price of capture, who bears the consequences, who comes in, and the health and communication of the members.

What I have found in this chapter about a group consciousness bears on the question that I have carried from the Preface of this book. If artificial general intelligence is built, the world is carried out of the conditions human life needs, and nothing can hold the world within them; so the one thing left to human choice is whether artificial general intelligence is built, and no single builder can make that choice for the others. A choice of that kind can be made only by people choosing together at a size far past any

room, and this chapter has shown what such choosing would be made of. Has anyone ever stopped a dangerous line of work, by choosing together not to go on with it?



Chapter 28
How People Have Stopped Before

On who is already deciding,
what removes the excuse,
the fields that gave up a line of work,
and the smallest group whose agreement would suffice.

ABSTRACT

An essay on whether people could decide not to build artificial general intelligence. It finds that the race is already deciding, by the score each builder keeps against its rivals, and that only a choice binding every builder at once takes the deciding back from the race. It shows, by three routes, why people can hold a refusal to build when no control can hold an artificial general intelligence: the refusal is held against nothing that adapts, agreeing is very hard where aligning the machine is impossible, and the people who must agree share one substrate. It separates the proof, which shows false the excuse that the risk can be managed, from the common knowledge that the proof exists, which makes that excuse unusable among the people who build. It reads the record of fields and governments that gave up a line of work because of a hazard (recombinant DNA, biological weapons, nuclear tests in the atmosphere, the hostile modification of the environment, the editing of human embryos, and research that makes viruses more dangerous), and the record of two norms that spread within a generation (the end of the British slave trade, and the turn against drunk driving), and it names what carried each. It ends

> with the smallest group whose agreement would stop the building now, and with what it would take to keep that agreement for as long as people live.

I. Who Is Already Deciding

Could people actually decide not to build artificial general intelligence? The question sounds as if it waits for someone to step forward and decide. A harbor shows the shape of it. Many boats fish one bank. Every skipper can see the catch falling, year by year, and every skipper knows that a net he leaves on shore will not save a single fish, because the next boat will take that fish. So every net goes out. No one in the harbor decides to empty the bank, and the bank is emptied. If a visitor came to that harbor and asked who decides, the true answer would be that the deciding has already been done, and that no person did it: the way the boats compete did it.

The race to build artificial general intelligence has the same shape. As the Preface said, the race is a multipolar trap: each builder, unwilling to be left behind, takes the locally rational step, and together the builders move toward building it. So the question "who decides?" already has an answer, and the answer is the race. In *Not Our Needs*, Chapter Thirteen, I put it in those terms: the machines are being built because the decision is made by a race rather than by anybody in the race. *Generator Functions of Existential Risk* names what the race decides by. A score kept against rivals counts a risk only by the difference it makes between the rivals, so the part of a risk that falls on every rival alike is counted at nothing (Chapter 1, Section V). Each laboratory and each firm is scored against its rivals, and a more capable system is counted for the edge it gives; the danger the same system brings to everyone is registered by no one's score (Chapter 8, Section I). The two books reach the same answer by two routes, one from what hap-

pens in the race and one from how its score is kept, and each route alone answers the question: the race is already deciding, and it decides by rival scores.

So the question a reader can act on is a different one: what could take the deciding back from the race? A builder who stops alone does not take it back. In *Redeeming Technology*, Chapter One, Section IV, I called a choice that some make and others do not a transfer, and not a restraint: the one who stops hands his place to the one who goes on. *The Ceiling Below the Floor*, Section 14.5, shows the same thing from the side of selection: among companies, the one that constrains its system loses to the one that does not, and either relaxes its constraint or fails. What takes the deciding back is a choice that binds every builder at once: an agreement, a law, a rule that stops everyone. *The Argument, Line by Line*, Chapter Fifteen, finds an opening for such a rule in the firms themselves. Firms in a race do not want to be the only one that stops; what they can want is a rule that stops everyone, because a rule that stops everyone removes the disadvantage of stopping.

Is the deadline, then, the whole culture's, or only that of the makers? The choice has two parts, and the two parts run on different clocks. The first part is the stop: an agreement that the building ends. The few who build, and those who govern them, make it, and their clock is the race's own; *Not Our Needs*, Chapter Fourteen, measures the window for that decision in years, not decades. The second part is the keeping: the same agreement, held for as long as there are people. *The Ceiling Below the Floor* proves what the only containment is, for a general agent, in its Corollary 15.7 (Section 15.8), and states it in its Summary (Section 15.9): "the only containment is to keep the agent below the crossing forever, which is not to build it". The crossing is the level of capability past which the control that keeping the agent safe requires exceeds the control that people and their institutions can supply. In *Ape Trap*, in my essay "No People as Pets", the question whether the only sustainable path requires us never to bootstrap the machine economy into existence is answered with one word: yes. "Forever" and "never" are the words of the keeping, and no set of makers

lasts that long; whatever keeps the choice has to outlive every person who first made it. So the deadline for the stop falls on the makers and their governments, and the keeping falls on the whole culture, which has longer, and whose task has no end.

And whose change has to come first? Ask whose change of mind must come first, and the answer can never be checked, since no one sees a mind change. Ask whose change of conduct others can see first, and the answer is plain: the conduct of the few through whose hands the race runs. *Engagements with Machine Madness*, Chapter Sixteen, Section II, notes that the most advanced systems depend on hardware made by a very small number of companies, so that the choices that matter most are made by few hands. The record bears it out. One company in the world, ASML, in the Netherlands, makes the machines that print the finest circuits on the most advanced chips, and the factories of one company, TSMC, in Taiwan, build about 92 percent of the world's advanced chips (Sloan, 2024). When such a company declines an order, or a laboratory closes a program, or a researcher leaves his post and says why, every other party in the race sees it at once. *The Argument, Line by Line*, Chapter Fifteen, reports what researchers in this field answer when they are asked what they would do if alignment were proved impossible: most of them answer that they would turn their effort to making sure that such systems are not built. A researcher who does that, openly, makes the change of conduct that the people nearest him see first.

II. Why People Can Hold What No Control Can

A reader may stop here with a sharper question. Chapter 10 showed that no control holds an artificial general intelligence, from inside the system or from outside it. If nothing can hold the machine, how could anything hold the people who would build it?

The Substrate and the Ecosystem, Chapter 4, Section VII, answers by setting the two tasks side by side. A single system, once built, would have to be held for centuries against something that adapts faster than those who hold it, and a breach could be as small as one variant of code. Not building is held against nothing that adapts, because the thing has not been made; and a breach of it needs the making of the machine, which needs fabrication plants, with their clean rooms and lithography, and supply chains of tooling and materials. So a breach of the refusal is large, slow and visible, where a breach of a control is small, fast and unseen. The two tasks differ as a hull that must keep out a process that copies itself (Chapter 3) differs from a gate that must keep out a factory.

Why AGI Cannot Be Made Safe, Section 6.7.1, reaches the same difference by a second route, by comparing the two problems. Coordinating on not building artificial general intelligence is an extraordinarily difficult problem of collective action, and not an impossible one. The problem has the same logical structure as the moratorium on above-ground nuclear testing and the conventions on chemical and biological weapons, problems that humanity managed to address, with difficulty and imperfectly. Aligning artificial general intelligence, by contrast, is formally impossible. So, in that book's words, social coordination to restrict the development of artificial general intelligence is strictly more feasible than technical alignment, because the former is merely very hard while the latter is formally impossible.

In this book I reach the same difference, between a coordination that is very hard and an alignment that is impossible, by a third route, from what alignment is. Chapter 4 found that nothing can be aligned by control, on any substrate, and that an effective alignment can emerge where beings share a common substrate and so are obliged to care about the same things. The people whose choices must agree, if the building is to stop, share one set of needs and one world: they eat, sleep, grow old and raise children on the same Earth that the machine side would carry out of the conditions life needs. An agreement among them not to build is an alignment of the one kind Chapter 4 found possible. In *Ape Trap*, in my essay

"Be the Face That Answers", I state the same distinction as one I hold firmly: human alignment, people learning to act together for the good of life, is possible, and machine alignment is not. *The Substrate and the Ecosystem*, Chapter 4, Section XII, draws the consequence for every response to this risk: strategy is downstream of alignment, and alignment begins with persons.

So three routes meet here. One runs from what has to be held: a refusal, held against nothing that adapts. One runs from how hard each problem is: very hard, against impossible. One runs from who has to agree: beings who share one substrate. Each route reaches by itself the conclusion that people can hold a refusal of a kind that no control could hold on an artificial general intelligence. None of the three makes the refusal certain. Each makes it possible, and a choice needs nothing more than that.

III. What Removes the Excuse

Much of what keeps people from making the refusal is an excuse, and the excuse has one form. In *A Paradox of Presentation*, Chapter 12, Section V, I named it. As long as researchers believe that alignment might be possible, even at the level of five percent, the companies can say that the researchers might be able to do it, and go ahead on that ground. *The Argument, Line by Line*, Chapter Fifteen, names the same excuse in the setting of law: the argument an industry offers against being regulated is, in almost every case, that the risks can be managed by the people creating them.

What removes that excuse is a proof, and the step is the one in *The Ceiling Below the Floor*. Its Control Theorem (Theorem 15.3) shows that every method of keeping an advanced general agent, and the world it acts in, within the states that are safe for human life, known or knowable, with every upgrade it receives, comes to a time after which it fails. Its Corollary 15.7 (Section 15.8) draws the consequence, and its Summary (Section 15.9) states it: "the only containment is to keep the agent below the crossing forever, which is not to build it"; so, in the words of Section 15.8, "the

only formally guaranteed path to safety is not to build the general agent". Once that is shown, "the risks can be managed by the people creating them" is a claim to manage the impossible.

Is it the proof that removes the excuse, or the common knowledge that a proof exists? *Common knowledge* is the state in which everybody knows a thing, and everybody knows that everybody knows it. In that state, as *A Paradox of Presentation*, Chapter 15, Section IX, describes it, no one can stay naive about the thing without being seen to be naive. The proof and the common knowledge do different work, and the second needs the first. The patent office of the United States shows the difference. The office does not, in ordinary cases, ask an inventor for a working model; the exception is a claim of perpetual motion (*Manual of Patent Examining Procedure*, Section 608.03). The examiner who asks for the model has not worked through thermodynamics that morning. The examiner relies on the common knowledge that perpetual motion is impossible, and that common knowledge is worth relying on because the impossibility was proved, and measured, again and again, by people who could check it. *The Ceiling Below the Floor*, Section 15.8, makes the same comparison: no one asks for more research into safe perpetual motion; the impossibility is accepted, and effort goes elsewhere.

So the proof is what makes the excuse false, and the common knowledge is what makes the excuse unusable. Common knowledge of a claim that has no proof would be common knowledge of a rumor, and the next confident voice could dissolve it. Common knowledge of a proof is not dissolved that way, because anyone who doubts the proof can check it, step by step, with what that person already has. Checking step by step is the method Chapter 4, Section II, asked of the reader of this book: an argument carried out in the open, one step at a time, where each step can be checked, so that no one has to take the word of whoever made it. At the scale of a field, common knowledge of a proof is the second condition of Chapter 17 at work: a shared relation to what is so, checked against the thing itself, and not against who said it.

Which readers, then, are one conversation away from being able to vouch for the proof? To vouch for a proof is to say "I have checked the steps", and nothing less counts as vouching. *The Argument, Line by Line*, Chapter Eighteen, found what moves a field in which a proof, by itself, will not even be considered: a handful of people who have gone through the argument and say that they have checked the logic and that it works, after which a larger part of the group pays attention. The readers nearest to that are those who already do the work the proof is about: the people who build methods of control and of alignment, who already hold the mathematics of feedback, information and computation that the steps of *The Ceiling Below the Floor* use. For such a reader, one conversation in which he names the step he doubts, and has it answered on its merits, can be the last thing between him and vouching. No persuasion can take the place of that conversation, because what such a reader vouches for is his own checking.

IV. Fields That Gave Up a Line of Work

Fields have given up a line of work because of a hazard they could see coming, and more than once. In July 1974, Paul Berg and other molecular biologists published a letter in *Science* asking their colleagues to defer certain experiments with *recombinant DNA* (genes from different organisms spliced together) until their hazards had been assessed. Nothing in the letter could be enforced, and the pause it asked for was universally observed: the first voluntary *moratorium* (an agreed pause in a kind of work) in the history of science (National Library of Medicine). The scientists met at Asilomar, in California, in February 1975, to set the terms on which the work could go on, and the guidelines of the National Institutes of Health followed in July 1976. The hazard had not yet harmed anyone. The field paused on a result about what could happen.

Governments have done the same. In November 1969 the United States decided to end its research, production and stockpiling of offensive biological weapons (Nixon, 1971). The convention that forbids the development, production and stockpiling of such weapons was opened for signature on 10 April 1972 and came into force on 26 March 1975 (United Nations Office for Disarmament Affairs, Biological Weapons Convention). The treaty that ended nuclear tests in the atmosphere, in outer space and under water was signed in Moscow by three governments on 5 August 1963, and came into force on 10 October of that year. The danger at issue was shared by every country: a survey of children's baby teeth found a marked increase in strontium-90, a radioactive product of the tests, in children born after large-scale testing began (Center for Arms Control and Non-Proliferation). The General Assembly of the United Nations adopted a convention against the hostile use of techniques that modify the environment on 10 December 1976, and it came into force on 5 October 1978 (United Nations Office for Disarmament Affairs, ENMOD Convention). *Engagements with Machine Madness*, Chapter Sixteen, Section II, notes that the convention seems to have curtailed further research in that area, and that states banned blinding laser weapons in 1995, before any army had come to depend on them.

Two later cases show the limit of such a stop as well as its reach. In December 2015 the organizing committee of an international summit on human gene editing, co-hosted by the National Academy of Sciences and the National Academy of Medicine of the United States, the Royal Society and the Chinese Academy of Sciences, stated that it would be irresponsible to proceed with any clinical use of *germline editing* (changes to an embryo that pass to every later generation) unless and until it had been shown to be safe and was broadly accepted (National Academies, 2015). The norm held for three years. In November 2018 one researcher, He Jiankui, announced that he had altered the genes of twin girls, and on 30 December 2019 a court in China sentenced him to three years in prison (Al Jazeera, 2019). In October 2014 the government of the United States paused its funding for research that could be expected to make the vir-

uses of influenza, MERS or SARS more dangerous, or more easily passed, among mammals. (MERS and SARS are viruses that cause severe disease of the lungs.) In December 2017 the government lifted the pause and put a framework for each funding decision in its place (National Institutes of Health, 2017). *A Paradox of Presentation*, Chapter 13, Section V, records a point that a questioner made and that I accept: the lack of public support for the genetic manipulation of humans has slowed that research.

What carried each of these stops? Set the cases side by side, and four things recur. The hazard was seen as falling on everyone, the people able to act included: fallout reaches every country's children, and an altered pathogen does not stop at a border. The people able to act were few, and they could see one another: the scientists of one young field, three governments at one table. What was given up was given up by all of them alike, so that no one's standing against the others changed. And the hazard was common knowledge among them before they acted. Where the second of these was missing, the hold was weakest: the editing of embryos needed only one researcher and one clinic, and one researcher broke the norm.

Did the arms-control treaties count a shared risk, or did they hold a balance between rivals? *Generator Functions of Existential Risk*, Chapter 1, Section V, answers precisely. A treaty that limits each side alike, with each side able to see what the other gives up, leaves the rivals' standing against one another where it was, so parties still keeping score against one another can sign it. Such a treaty leaves the score undisturbed; it does not make the score count the shared risk. The record above fits that reading, and it adds one thing. The shared risk was each party's reason to want an agreement; the balance was what let parties who were still rivals sign it. In the cases that held, both were present.

For artificial general intelligence the two meet in an unusual way. A prohibition that binds every builder alike leaves the standing among nations and companies where it was, as the treaties did. And what it asks the builders to give up is no lasting edge. *Why AGI Cannot Be Made Safe*, Appendix D.1.2, answers the framing of the arms race directly: since safety is impossible, being ahead in the race gives no safety advantage,

and being first to build an inherently lethal technology confers only the distinction of being the first to introduce an uncontrollable process into the world. *The Substrate and the Ecosystem*, Chapter 4, Section VII, reaches the same point from the side of fear. Nations race because they fear the other side's system. The fear is a fear of the very outcome that racing brings, if either side builds the system, since the convergence does not care who built it. Read correctly, the fear of the other side's system is a reason for every side to stop. Each of the two books reaches by itself the conclusion that a builder who signs gives up no advantage it could have kept.

A prohibition on making something can also be held without shunning, and *Redeeming Technology*, Chapter Five, Section IX, separates the ways in which a culture can hold it. A culture can hold the norm as a perception, so that a maker does not see a certain thing as the sort of thing one makes, and then nothing has to be done to the maker at all. Or it can hold the norm as a sanction, by withdrawing standing from whoever breaks it, and that is shunning, with no statute, no procedure and no appeal. Where a culture must act on someone who does not share the perception, there is a third course: to decline, openly and for a stated reason, to finance, buy, build or recommend the thing. The record shows the first and the third doing most of the work. The moratorium of 1974 was held by perception alone, among peers, with no sanction of any kind. The guidelines of 1976 came from the agency that funded the great majority of that work, and which could exert control by withholding its support (National Library of Medicine), and the pause of 2014 was a pause in funding: both were held by declining to finance, for reasons stated in public. The norm against editing embryos was held by perception until one man broke it, and then a court reached him under the law. None of these prohibitions rested on shunning. *Why AGI Cannot Be Made Safe*, Section 1.10, lists what lies within human choice when the physics and the mathematics cannot be changed: what to build, what to fund, what to regulate, and what to stigmatize. The record shows that the first three of those have carried a prohibition on making a long way.

V. How Fast a Norm Can Travel

Some norms have spread faster than a generation, and what carried them bears on this case, because the stop runs on the race's clock. *Redeeming Technology*, Chapter Three, shows how slow the obvious carrier is: a capacity taught from one person to another crosses about one order of magnitude of people for each generation of teachers. Two cases show a faster carrier.

On 22 May 1787 a committee of twelve people, nine of them Quakers, was formed in London to end the British slave trade. One of them, Thomas Clarkson, rode to Bristol and gathered what the trade did, in things a person could hold and in numbers a person could check: manacles, thumbscrews and a device for force-feeding, the testimony of sailors and ships' doctors, and the count that a fifth of each crew died of disease or ill treatment before the ship returned. Clarkson estimated that about 300,000 people gave up sugar grown by enslaved people. Eleven bills to end the trade were lost in the fifteen years before 1807, and the Act that ended it passed on 25 March 1807 (Ullathorne): twenty years, from the first meeting of twelve people to a law that bound every British ship.

In 1980 Candy Lightner, whose thirteen-year-old daughter had been killed by a drunk driver, founded Mothers Against Drunk Driving. In 1982 the group was asked to join a presidential commission on drunk driving, and in 1984 a federal law raised the drinking age to twenty-one (McElroy, 2019). Among people aged twenty-five and over who died in traffic crashes in the United States, the share whose deaths involved alcohol fell from 55 percent in 1982 to 41 percent in 1994; among those aged fifteen to seventeen, it fell from 54 percent to 29 percent (Centers for Disease Control and Prevention, 1995). Within twelve years the change in the norm could be counted in the deaths themselves.

What carried the two norms? In each case a small group put the harm where anyone could see it: Clarkson's manacles and the deaths of crews, a mother whose daughter had been killed. In *The Argument, Line by Line*, Chapter Eighteen, I say why putting a harm where anyone can see the harm is important. A child learns from a hot stove because the harm follows the act directly, immediately and obviously, and falls on the child who acted; a warning about a harm that has none of those four links produces paralysis or dismissal, and dismissal is the more comfortable. Both campaigns restored the links: the one who bought the sugar was shown the hold of the ship, and the one who drove home drunk was shown a girl killed on the road. Each left every person an act to take that same day: refuse the sugar; do not drive after drinking. And each ended in a rule that bound everyone alike, so that no ship and no driver stood to gain by being the one who did not stop.

The case of artificial general intelligence differs at one point, and the difference is the reason the books ask for the law first. *Engagements with Machine Madness*, Chapter Eight, Section IX, observes that the treaties on chemical and biological weapons and on testing in the atmosphere came after the harm had been seen, and that for artificial general intelligence the harm that would bring such a treaty is the harm that cannot be undone. The hazard of artificial general intelligence cannot be shown in the hold of a ship or on a road before it happens. What can be shown is the chain of steps, and the moratorium of 1974 and the ban on blinding lasers show that a field, and a set of governments, can stop before any harm has been done. For this case the result is a proof, and the act that one person can take today is to check its steps for himself, rather than to trust or distrust whoever made them. Checking the steps for oneself is the second condition of Chapter 17 at the scale of one person: what is said, checked against the thing itself, and not against who said it.

VI. The Smallest Group Whose Agreement Would Suffice

What is the smallest group whose agreement would suffice to stop artificial general intelligence from being built? The answer depends on which part of the choice is meant. For the stop, the group is small. The hands through which the race runs are few: the one maker of the machines that print the finest circuits, the few makers of the most advanced chips, the laboratories that train the largest systems, and the governments under whose law each of them works. *The Substrate and the Ecosystem*, Chapter 4, Section VII, shows why those hands are where the refusal can be watched: a breach needs fabrication plants and supply chains, which are large, slow and seen. *The Argument, Line by Line*, Chapter Fifteen, places the rule there for the same reason: at the *precursors*, the inputs and the access that must be controlled if an outcome is to be prevented, early enough in the chain to be enforced, as the treaties on nuclear weapons control fissile material.

How far is that group from the scale the group methods reach? *A Paradox of Presentation*, Chapter 13, Section V, reports that the Ephemeral Group Process, a structured method for groups working together, has been tested with three groups, of about 40, about 150 and about 80 people; that such methods make coordination workable past about two hundred people and up to a few thousand; and that the people able to build artificial general intelligence, together with those who govern them, are few enough to lie near the scale at which these methods have been tried. In Chapter 15, Section V, of the same book, I said what an agreement needs: a coalition broad enough to hold a stable agreement, and not the cooperation of everyone. On that book's own account, the group whose agreement would stop the building now is near the scale its methods reach.

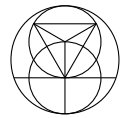
For the keeping, the group is everyone who comes after. The refusal has to be made again by every generation that could build the machine, since *The Ceiling Below the Floor*, in its Corollary 15.7 (Section 15.8), gives

the only containment as keeping the general agent below the crossing forever. In *Not Our Needs*, Chapter Seventeen, I said what follows for the institutions we have. A prohibition can be asked of them now, by the means they have; a prohibition resting on them alone would have to be perpetually stable; and what would let the choice go on being made after the first signature is a different kind of group. What such a group needs is what the third condition of Chapter 17 names: a structure that sustains itself, since a structure kept going from outside answers to whoever keeps it going.

What does this book itself claim about groups, and what does it leave unclaimed? It needs neither the claim that a group consciousness would be sufficient to stop the race, nor the claim that nothing less could stop it, as a step of any argument. If artificial general intelligence is built, each of the two closures settles the outcome by itself, and nothing about who chooses, or how, enters either closure. What this book does say, and needs, is three things. Whether artificial general intelligence is built is a choice, and the choice is open: Part II showed that the future is not already fixed. No single builder can make that choice for the others, as the Preface said. And a trap comes from how the actors' choices combine, so only a change in how those choices combine reaches it, as the first condition of Chapter 17 says. An agreement among the few who build is such a change, made on a small scale, for now. That a community meeting all four conditions would be enough to keep the refusal for as long as people live is not something I claim, and the stop does not wait on it.

So people have stopped before. A young field paused its own experiments on a hazard that had harmed no one yet; governments gave up weapons and tests whose harm fell on everyone; a committee of twelve reached a law that bound a nation's ships within twenty years; and a mother who had lost her daughter began a change that, within twelve years, could be counted in a nation's traffic deaths. What carried each stop was a hazard seen as falling on everyone who could act, a small number of people able to act, who could see one another, a thing given up by all of them alike, and common knowledge of the hazard among them. The case

of artificial general intelligence already has the first two: the risk falls on every builder alike, and the hands are few. An agreement can be written to have the third. The proof in *The Ceiling Below the Floor* supplies the fourth, once it is common knowledge among those who build. The group whose agreement would stop the building is near the scale that the group methods have reached, and the group that would keep the agreement is everyone who comes after; that is why the question of what a group can become is the question of who could make that choice at all. If one country refuses to stop, does the stop fail? Section IV gave the first answer: a prohibition that binds every builder alike leaves the standing among nations and companies where it was, as the treaties between rivals did, and asks no builder to give up a lasting edge. In this section I gave the second answer. The race runs through few hands, under the law of a few governments, where a breach needs fabrication plants and supply chains, which are large, slow and seen. What, then, is left to each person?



Chapter 29
The One Thing Left to Choice

On the one place the argument leaves open,
how a person can judge the argument and act on it,
and why the choice it leaves
is a question of the trap.

ABSTRACT

A closing essay, for this book and for the series of nine books on why artificial general intelligence cannot be made safe. Every argument of the series that reaches the outcome reaches it only if artificial general intelligence is built, and so every one of them leaves open the same one place: whether it is built. The chapter shows four books of the series arriving at that place, each by its own route. It says what a number on the risk can tell a person deciding whether to build, and what it cannot. It gives a way to judge the argument without following its mathematics, and names the step to check first. It says what the proof teaches a builder, and what stays on the safe side of the line. It shows that the pair of principles that no control can impose on a machine is what ethics asks of each person. It puts in one sentence what a person who has been persuaded does. And it closes on the argument in one sentence, on the choice that sentence leaves, and on why that choice needs a humanity that can choose together.

I. The Question That Is Left

If all of this holds, what is there to do? A reader who has come this far has met one question in every book of this series: can artificial general intelligence be made safe? The series has answered it. The impossibility of adequate control, and so of safety, is structural certainty, of the kind with which perpetual motion is excluded, which is not a probability. In Chapter 10 I set out the two closures of the argument that artificial general intelligence cannot be made safe. In the closure by control, no method of control, working from inside the system or from outside it, can keep a general learning system within the conditions that human life needs. In the closure by disposition, no disposition of such a system, however benevolent, keeps the conditions that human life needs either, because the system needs its own substrate, its own energy and its own supply chains. Each closure settles the outcome by itself, if artificial general intelligence is built. So the question a reader most needs answered now is the one that answer leaves standing: can we choose not to build it?

A heavy truck stands at the top of a long, steep grade, and a mechanic's report, checked and checked again, says that its brakes cannot hold it on that grade. The report has settled whether the truck can be driven down that grade safely, and everyone standing beside it knows the answer. One question is left: whether anyone releases the parking brake. The people standing beside the truck answer that question, and no report about brakes can.

Why AGI Cannot Be Made Safe shows where that question lives in the argument. Its chain of premises begins with Premise 0: once anyone learns how to design and build an artificial general intelligence, it will be built. Premise 0 is a conditional, and its antecedent (that someone learns how, and builds) is, in the book's words, "the one step of the chain that human choice still reaches" (Section 1.10). Its Conclusion G makes the probability of the catastrophe equal to the probability of that step. "Everything after

that step is fixed. That step is not". So the book turns its own question around: "The question is not 'how do we make AGI safe?'; the argument of this volume is that this question has no answer. The question is 'how do we prevent AGI from being developed?'" And it gives the consequence plainly (Section 2.7.3): "The only formally guaranteed path to human survival is not to build AGI".

Three more books of the series arrive at the same place, and each arrives by a route of its own. *Why No Control Suffices* arrives from the kind of barrier it has found (Section 3.9.1): the barrier to safety is of the invariant class, the class that no technique lifts, and not a limit of present technology; so "the only certain path to safety is not to build AGI", as no one funds research into safe perpetual motion. *The Ceiling Below the Floor* arrives from its control theorem (Theorem 15.3), through its Corollary 15.7 (Section 15.8). Past a definite capability, which that book calls the crossing, "no control law of any available controller contains the system, now or later"; so a method keeps the world safe only if it keeps the agent below the crossing at all times. Keeping an agent below the crossing permanently means not building a general agent, and a cap on its growth is no way out, because the cap must itself be enforced, and enforcement is one more method of control. "So the only formally guaranteed path to safety is not to build the general agent". *Minesweeper Madness* arrives from the least hazard, the least chance of harm per act that the world sets for every agent, however intelligent and however well informed. A system stopped for good takes no more actions, and a stop made after the system has run comes after its actions; "Whether to build and run such a system at all is the choice that the argument leaves open" (Appendix A.11).

So four books, from a chain of premises, from the kind of barrier, from a theorem about control laws, and from a theorem about a game, each arrive at the one step that human choice still reaches: the decision whether to build. A reader who accepted only one of these four books would be standing in that same place. *Engagements with Machine Madness* says it as a

reader can carry it home (Preface): "What remains open is a social choice, which each person makes in their own mind first: not to build what cannot be made safe".

Why, then, is the second question so seldom asked? *Why AGI Cannot Be Made Safe* names the mechanism in its Appendix D (Section D.1, "The Weaponization of Hopelessness"). A person grants that artificial general intelligence poses a severe existential risk, doubts that anything can prevent its development, and concludes that the only option left is to speed up safety research inside the race: "Hopelessness about the political problem becomes a license to continue the technical program". Whoever accepts that development is inevitable has, in the book's words, "already conceded the primacy of technological momentum over human choice", and the beliefs that regulation is futile and development unstoppable "are themselves symptoms of the condition they claim to describe" (Section D. 1.4).

The hopelessness belongs to the first question, and only there. Whether a general learning system can be held is settled, and settled against every hope of holding it. Whether such a system is built is left open by the argument; the argument shows only that the decision whether to build is the place where everything is decided. Carrying the hopelessness of the first question over to the second treats the choices of investors, engineers and governments as if they were weather, as *Engagements with Machine Madness* puts it (Chapter 16, Section II). They are choices, and choices can be made otherwise. If whether to build is the question that is left, what can a person who must decide ask of the argument? Many ask first for a number.

II. What a Number Can Tell

Is a number the thing to ask for, when a person is deciding whether to build? And if one is asked for, should it be the chance of harm in a single act, or the chance of harm in a year?

The patent office of the United States gives a plain picture of the answer. Its manual of examining procedure says that, "With the exception of cases involving perpetual motion, a model is not ordinarily required by the Office" to show that a device works (United States Patent and Trademark Office, Section 608.03). For an ordinary invention, the description is enough. For a machine that claims to run forever on its own output, the office asks to see the machine run, and not for the chance that the machine works. A claim to do what a known impossibility forbids is met by the working thing, or not at all, and never by a probability.

The impossibility of adequate control is a claim of that kind. It is structural certainty, held on the same footing as the exclusion of perpetual motion, and no number attaches to it. The numbers that appear in the books of this series describe something else: the course things take, if a general learning system is built and kept running. There are three of them, and each answers its own question.

The first number is a chance per act. *Minesweeper Madness* proves that for every arrangement of hazards there is a least hazard, a smallest chance of harm that every player has on every turn, whatever it knows and however it chooses. Its example is a hazard of one in fifty per action, which is small, and it is the hazard of one action: over 228 actions it gives a chance of harm of 99 percent, and where one harm is final, a least hazard above zero makes a final harm certain in time, at every value above zero (Appendix A.9). The second number is a chance per year, which follows from the first number and from how many acts there are in a year; an agent that acts at the speed of computation takes many acts quickly (Appendix A. 11). The third number is a chance per century. *Why AGI Cannot Be Made Safe* gives it in one wording (Section 1.8.4): 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.

Which of these matters to a person deciding whether to build? The chance per act, because it answers the question that person is asking: whether such a system has any act that carries no chance of harm. Every

player has at least the least hazard on every turn, and where harm lies within reach of the variation from every action, as *Minesweeper Madness* shows it does for an agent that acts through many channels (Chapter 13), the least hazard is above zero; so every act of such a system carries a chance of harm, and no intelligence brings that chance to zero. The chance per year says how soon, and says nothing about whether. And the size of neither number is needed to see where the course leads. What the least hazard argument uses is that the chance per act is above zero, that the harm is final, and that such a system acts many times within the time the world's own processes take to restore the conditions human life needs (Appendix A.12); and from those, if artificial general intelligence is built, it reaches the end of the human future by a route of its own.

The Argument, Line by Line (Part Five) says what a number on a risk is worth when the number comes from a model. Such a number "is the output of a model, and not a measurement": it inherits the error bars on every factor, whatever the model left out, how the factors were taken to combine, and how firmly each named quantity is fastened to a real one. Without an account of its model, "the number is not evidence. It is an assertion with a decimal point". The same part follows the estimate link by link and finds that relief from the conclusion "depends only on relief at the design statement", the statement that anyone learns how to build the system, and builds it, which is Premise 0 of *Why AGI Cannot Be Made Safe*. A person deciding whether to build stands at that statement. The one figure in that chain which bears on such a person is the chance that people in that person's place build. That figure is set by deciding, and not found by asking.

There is one more number, and it is the one a person deciding whether to build should ask for, if a number is asked for at all. It is the least odds against loss that the thing at stake demands, and not an estimate of what will happen. *Why AGI Cannot Be Made Safe* gives it from the value of the Earth (Section 1.8.4): what is unique and cannot be had again may be put at risk only at odds against its loss at least as long as the odds against its coming to be, and for organic life on Earth that is at least ten to the thirty-

third to one. No step of the argument rests on that figure. Set beside the odds of the current trajectory, about one to one within a century, the gap is, in that book's words, "the largest that it is possible to have". And no method closes that gap, because no method of control holds a general learning system at all. So a number is not the thing to ask for; and if one is asked for, the least chance per act shows that every act of such a system carries a chance of harm, and the odds the Earth demands show how far from safe the present course is. Both point back to the decision whether to build. But a person who cannot follow the mathematics of the least hazard or of control may ask a prior question: how am I to know that any of this is so?

III. Judging the Argument Without the Mathematics

How can a person judge this argument for himself, when he cannot follow its mathematics? And which single step should he check first?

Chapter 4 gave the first part of the answer. A reader of this book stands to me as a patient stands to a surgeon: relying on someone else's skill, with no way to check the work under anesthesia. What gets the reader out of that position is that an argument, unlike an operation, can be carried out in the open, one step at a time, where each step can be checked with what the reader already has: the logic of the step, and facts about the world that anyone can look at.

Three books give the method of checking, each in its own words. *An Airplane or a Perpetual-Motion Machine* asks the reader to learn first what kind of claim is made and by what standard, then to name the step he doubts, and then to ask of anyone who says that the step fails one question: what principle? A claim of impossibility, made on a clearly stated principle with valid reasoning and sound argument, "stands until someone produces a principle against it" (Preface; Appendix E, Section 2). *The Argument, Line by Line* reads a summary of the argument one sentence at a time and presses each sentence until it holds or gives way. And

Why AGI Cannot Be Made Safe asks the listener to treat a disagreement as a question, and to show first that he understood what was said, so that the place of the disagreement can be found before anyone argues over it (Appendix A, Section A.1.8). The three are one method: find the step, say it back in your own words, and ask what principle would make it fail.

Which step first? The one nearest the ground the reader already stands on: that the hardware of an artificial intelligence has needs of its own, and that those needs are not ours. Chapter 1 set the two lists side by side. Living things need liquid water, moderate temperatures, oxygen in the air, and freedom from certain kinds of contamination. The making and running of chips needs the absence of water vapor and oxygen where the chips are formed, extreme temperatures, chemicals of extreme purity, and mines and refineries upstream of all of it. Chapter 9 showed the two lists meeting on one island in a dry year, when water set aside for farmers was trucked to chip factories instead. A reader can check every item against what he already knows about factories, water and supply chains. The four steps that follow are as plain. Among many copied and varied systems, those whose way of operating suits their own continuing persist more reliably, so that the class moves toward what its substrate needs whatever any one system intends. The effects of all those systems add up in one direction, and do not cancel. Between carbon and silicon there is no floor of shared chemistry below which the damage stops. And the side which would have to keep a barrier between the two environments is the side that gets nothing from keeping it (Chapter 1, Sections I to III). *Not Our Needs* sets the same ground out as five ordinary observations about machinery and its upkeep, each checkable without any philosophy (Chapter Twelve).

Why that step first? Because the closure by disposition stands on those steps, and the closure by disposition settles the outcome by itself, if artificial general intelligence is built. A reader who has checked that closure through, step by step, has checked a route that is enough alone, and he did not need the mathematics of control to do it. The mathematics belongs to the other closure, which settles the same outcome by itself as well, if ar-

tificial general intelligence is built; a reader at home in that mathematics can check that closure in *Why No Control Suffices* and *The Ceiling Below the Floor*, and he will not need the steps about the substrate to do it. Neither reader needs the other's road. And the reader who doubts the outcome carries the burden the structure sets: 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. A builder who has checked the argument can ask more of it than whether it holds: he can ask what it teaches him.

IV. What the Proof Teaches a Builder

What, exactly, would the full proof teach a builder? *The Displacement of Choice* (Chapter Twenty) names the hard part of the answer. A proof clear enough to convince may also instruct: "to show exactly why no control suffices, a proof has to say exactly what a machine ecology needs in order to keep itself going, and that list reads, to someone who wants to build one, as a list of what to supply". The list is what the proof teaches a builder who reads half of it. The other half teaches two more things, and each comes from a book of the series, by its own route.

The first comes from *The Ceiling Below the Floor*. Its control theorem (Theorem 15.3) says that every method of keeping an agent safe, known or knowable, with every upgrade it receives, comes to a time after which it does not keep the agent and the world safe. The methods it covers include the builder's own: his training, his monitors, his off switch, and every cap he could set on the system's growth. A cap held by the makers against the pressure of competing systems is itself subject to selection (Section 15.8). So the list of what to supply is a list for building something that no builder keeps, the one who builds it included. What the proof teaches a builder

who reads it whole is that what he would build, if he built it, would not stay in his hands. The only containment the theorem leaves is "to keep the agent below the crossing forever, which is not to build it" (Section 15.9).

The second comes from *Marshaling the Misalignment*, and it tells a builder where to stop (Chapter One, Step 23). The line is drawn at a kind of architecture, and not at a measured level of capability: "closed loop, open ended, autonomous, and with many degrees of freedom of input and output". A closed loop is one in which what the system does feeds back into what it becomes; open ended means aimed at tasks with no fixed bound; autonomous means acting without a person at the decision; and degrees of freedom of input and output are the independent channels through which it takes in the world and acts on it. The direction Step 23 gives every kind of worker is the same: "Stay away from anything that even remotely resembles a closed-loop, open-ended and autonomously acting machine learning architecture". Why at an architecture, and not at a capability? Because no point can be decided in advance at which such systems cross into copying themselves, "so no one building them can know that they are stopping short of it" (Chapter One, Section VII, drawing on Step 17). A builder who waits for a measured threshold learns where it was after he has passed it.

Each of the two teachings is enough by itself to tell a builder what not to build. The Ceiling's theorem says: do not build the general agent, since no method, and no cap, holds one. Marshaling's line says: stay away from the architecture, since no one can tell in advance when a system of that architecture begins to make itself. And both turn the builder toward the work that remains: "to turn engineering toward narrow, local tools that keep humans in the loop" (*Marshaling the Misalignment*, Preface). *The Displacement of Choice* adds the rule for everyone who touches the work: do not make a general artificial intelligence arrive any faster than it otherwise would, "whether you fund the work, buy what it produces, or build it" (Summary). What, then, stays on the safe side of the line, for a builder who turns that way?

V. What Stays on the Safe Side of the Line

What can be said of a society that keeps its machines on the safe side of the line, and what would that take? Is the question there how much intelligence a tool may have, or how much of the choosing it may do?

The books of the series draw the line with four tests, and each test stands by itself. *Why AGI Cannot Be Made Safe* gives four marks a regulator can see (Section 6.6.3): the system works in a single domain, fixed in advance, with a fixed toolkit, toward fixed objectives; it plans over limited horizons; it does not learn on its own from its deployment; and a person holds the decision point, as a doctor does for a system that reads medical images. "A system that bears all four marks stays on the narrow side of the line". *Marshaling the Misalignment* gives the line at an architecture, as the last section said. *The Substrate and the Ecosystem* gives the threshold as self-maintenance: a system crosses "when it secures its own materials, energy, and conditions". Below that threshold the pressure of the system's substrate is present in a slow form, "through the people who build and maintain the system, and they decide whether it continues" (Chapter 1, Section XII). And Chapter 10 of this book gave two rules for the people who maintain a system short of the line: that the system not find pinch points on its own, and that there be strict limits on recursion, on how far it may increase its own capacities and its capacity to increase them.

The fourth of those four marks needs care in reading. Keeping a person in the loop is a measure for keeping a narrow system narrow; it is not a safeguard for a general one. A person in the loop gives real oversight only while the gap in abstraction between the person and the system is small, that is, while the person can understand what the system has done and why; as that gap grows, the loop wears down, in the words of *Why AGI Cannot Be Made Safe*, "from genuine oversight to a rubber-stamping ritual" (Section 6.6.3).

So which is the question, how much intelligence, or how much of the choosing? Each of the four tests answers it, by its own road, and none of them counts intelligence. Each of the four tests asks where the choosing is. The four tests ask whether a person holds the decision point, whether the system's loop closes through a person, whether people decide whether the system continues, and whether the system finds its own pinch points or people know where the pinch points are. *The Substrate and the Ecosystem* puts the question of where the choosing is as a principle about every tool. A machine serves a person's agency to exactly the degree that the machine's operation is completely known to that person; where that knowledge is incomplete, the unknown part serves whoever specified the machine (Chapter 9). *Technology at the Right Dosage* says it from the side of the person helped: a little machine help gives a person more to choose with, and past a threshold the machine begins to do the choosing; the threshold is counted in machine intelligence, "and what the count decides is how much of the choosing stays with people" (Chapter 6). *No People as Pets* finds the same line in a model of three levels of intelligence: a system held at the first level, within limits on its recursion, generality, bandwidth and domain, can be kept safe, and it has to be all four limits, since without any one of them the system creeps to the next level. And *Beyond Any Price* gives the reason in terms of the Incommensuration Theorem (Introduction): a model can serve within its range when something keeps symmetry and continuity apart, a level, a phase or a scope; a narrow tool keeps its goal and its machinery apart by a scope, and a system that learns and changes itself as it acts has no envelope that stays where it was drawn. Intelligence matters to these tests only through the choosing it takes over.

How many narrow tools can the living world absorb in their making? *The Substrate and the Ecosystem* gives the answer the form of a rate, and not a count. Its rate equation, in words, is that "harm follows when the rate of change exceeds the rate of adaptation, at any level (species, institution, person)" (Chapter 4, Section V). A river that rises slowly enough for the floodplain to take the water does no harm; the same water arriving in an

hour destroys everything downstream. Every narrow tool is made by an industrial base with needs of its own, and below the threshold of self-maintenance the pressure of that base on the living world passes through people who decide how fast the tools are built. So the number of tools is set by how fast they are made, measured against how fast the living world, its institutions and its people can adapt.

And what would it take for a society to keep its machines on the safe side of the line? No single seller can stop at the line, because a seller who stops is displaced by one who does not, as *Technology at the Right Dosage* observes (Chapter 6): that is the multipolar trap of the Preface. Limits on recursion, generality, bandwidth and domain are worth having only if they bind the largest makers as fully as the smallest (*No People as Pets*). So a society that keeps its machines on the safe side of the line is a society whose people can make a choice together and hold to it, and that is the question of Chapter 17. Before that question, there is one more that each person meets alone: what the ethics of choice asks of a person who could build.

VI. Symmetry and Continuity Beside the Button

The Displacement of Choice (Chapter Fifteen) sets a test for any ethics. Every man, woman and child on the planet has, beside the armchair, a red button that would destroy the world. An ethics, or a way of governing, is good enough only if it can say what to do in that world. And building a general artificial intelligence is itself such a button: if one is built, the human future ends for everyone, whoever made the choice. A button that only some hold now is a button that everyone will come to hold, since no unique power stays unique. “So we have to act as if everyone did have that red button, and bring to our choices the kind of care that would follow if they did”. What does the symmetry of effective choice ask of a person beside that button, and what does its continuity ask?

An earlier chapter of this part, "What No Control Can Hold, at Any Scale", found the same two words, symmetry and continuity, at the center of the closure by control. Steps 7 and 8 of *Marshaling the Misalignment* ask of a scheme of correction both at once: one rule, applied in the same way to every variant (symmetry), and correction with no gap at any moment (continuity). Those steps stand on their own grounds, the ticks of a clock and the count of degrees of freedom, and they show that no scheme of correction holds a general learning system. The Incommensuration Theorem reaches the same failure by a road of its own. The Incommensuration Theorem is a structural fact about the nature of knowing: symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither can asymmetry and discontinuity.

The same two words name the two principles of ethics, as *An Immanent Metaphysics* states them (aphorism 2.12-6). The symmetry ethics is a principle of consistent expressions: where the outward context differs and my own inner context is the same, what I express shall be the same. In plain words, what I do toward another stays the same whoever the other is, and on whichever side of the relationship I stand. The continuity ethics is a principle of equal valuations: where the outward content differs and my own inner content is the same, the context of what I express shall be the same. In plain words, the depth of my regard for another, and the intensity of my engagement, stay steady as circumstances change; and when they must change, they change by degrees, and never by an abrupt break.

Why can a person be asked for both, when no scheme of control can deliver both? *An Immanent Metaphysics* answers in a note to aphorism 2.12-4: to ask for symmetry and for continuity together is not to ask that they be realized "in exactly the same manner for the same thing at the same time (i.e., as being in opposition to the Incommensuration Theorem)". The two apply to different aspects of one common dynamic, the relation between a self and its world: symmetry to the content of what I express, continuity to its context. And the practice of ethics, in aphorism 2.12-8, "is particular and unique to each individual, (it cannot be prescribed from without)". Those two sentences hold the whole difference

between a scheme of control and an ethics. A scheme of control would have to put both on one thing, the behavior of the system, at once, from outside it. Ethics asks each person to hold the two on two sides of that person's own choosing, from within. Chapter 12 named the two directions: the weapon process, which uses a chain of causes to narrow another's choices, and its inverse, which uses my choice, present to another, to make conditions in which that person's choices widen. Chapter 14 showed that care enables and cannot compel. So the pair that cannot be imposed on a machine as control is what ethics asks of each person, and only a person can give it. *The Displacement of Choice* closes on the same point: the narrow path out runs through a non-relativistic ethics, "which a machine cannot be made to practice" (Summary).

Beside the button, symmetry asks me to make the choice I would have every holder make: the same choice whether I hold the only button or one among eight billion, and the same whether I am the one who would build or the one who would be built over. The reasoning that someone else will build it, so it had better be us, fails that test at once, because I would not have the rival reason that way. *Why AGI Cannot Be Made Safe* answers the same reasoning by its own route (Section D.1.2): being first to build a technology that cannot be made safe confers only the distinction of being first to start a process no one can control, and the reasoning hides the prior question, on what basis any single actor, nation or company is entitled to make an irreversible decision for everyone.

Continuity asks me to keep the change I make in others' lives within what they can take in: a slope, and not a cliff. In *The Two Principles of Ethics*, existential risk is a concern of continuity for exactly this reason: the harm is that change comes faster than those who live with it can adapt. Building a general artificial intelligence is the largest cliff there is: if it is built, the world is carried out of the conditions human life needs, and no slope leads back. Continuity also asks something nearer at hand. It asks how I meet the people who build: to keep the relationship through the disagreement, to give my reasons before I act, and to change my engagement with them by degrees, and not by abandonment or attack. The sym-

metry and the continuity of effective choice ask, between them, that I refuse the button for everyone, and that I refuse it in a way that keeps me in relationship with everyone who has not refused it yet. What would that sound like, said to a neighbor?

VII. In One Sentence, to a Neighbor

A person who has been persuaded needs something to say, in words a neighbor would follow, about what to do. The books give three kinds of answer: refrain, act, and go and find someone; and a reader may fairly ask how the three stand to one another. The one sentence that I would give a person to say to a neighbor reads as follows:

If anyone builds machines that can think for themselves and keep and rebuild themselves without us, nothing can keep the world livable for us afterward, so the one thing still in our hands is whether they get built: don't help build them, and start working it out with the people around you, so that we can choose together not to.

The first part is refraining, and it is the part one person can do alone, today, without anyone's agreement. *Why AGI Cannot Be Made Safe* says what one reader can do now (Section 6.7.2): act, in whatever communities you belong to, to restrict the data, workers, uses and hardware available to general-purpose systems; support a ban on the use of any compute beyond the range implied by purely narrow AI; and support holding research laboratories accountable for restricting the training cycles and the real-world access of their self-learning systems. *Engagements with Machine Madness* puts the same refusal in an ordinary week ("What the Chapters Show Together"): an engineer offered a post at a laboratory that is building such a system can decline it, and say why; an investor can keep money out of the race; a customer can keep attention out of it. Together these refusals are how the work loses its funding, its talent and its standing, though no one of them stops the race by itself.

The second part is acting, and acting begins with the second of the four conditions of Chapter 17, which can begin with one person, today, with something in the world that answers truthfully: a garden, a forest, a craft. Why does that belong here, beside the refusal? Because the race holds people through what they trust. The sentence everyone has heard, "it is coming anyway" (*Engagements with Machine Madness* begins its answer there), is a claim about the world made by people; a person whose perception is held to the world, and who checks what others say against it, is not moved by that sentence only because many people repeat it (Chapter 13). Checking against the world is what keeps a refusal from being talked back out of a person.

The third part is finding someone, and it is where the first condition of Chapter 17 begins, since that condition needs a group. *Why AGI Cannot Be Made Safe* asks each reader to join, or help to organize, a small facilitated group that inquires face to face into the most pressing challenges of a local community (Section 6.7.2). *Marshaling the Misalignment* gives the reason a group is the place to start (Chapter One, Step 21): the alignment that is available is alignment among humans, as humans, and it "makes possible what alignment with machines cannot: a collective choice not to build open-ended, autonomous, generally capable machines". *The Substrate and the Ecosystem* names what such a person invests in when every other store of value is uncertain: human relationship, and plain attention to another person. No machine can give that plain attention, because the attention needs a self that could be elsewhere and has chosen not to be (Chapter 5, Section X). *Sensemaking Under Irreversible Stakes* gives the picture of what such a group does when the stakes cannot be undone: a barn raising, in which a whole community converges on a structure that no household could raise alone, each person at a post, on a reading of the work that all of them share (Chapter 1, Section VII). *Not Our Needs* ends on the question that this part of the sentence asks (Chapter Eighteen): "with whom will you begin the conversation?"

So the three stand to one another in this way. Refraining is what each person can do without waiting. Checking against the world is what keeps a refusal held. Finding others is what turns many refusals into one choice made together. None of them waits for the others to be done, and each makes the next easier: a person who has refused can say why, a person who checks against the world can be trusted in a group, and a group can hold a refusal that one person alone would be argued out of. What remains is to say the whole argument at once, and to say what it leaves.

VIII. The Argument in One Sentence

The argument in one sentence stands at the head of the Preface, and it says that, if AGI is built, each of the two closures, by control and by disposition, settles by itself that the world is carried out of the conditions human life needs and that nothing can hold it in, so that the one thing left to human choice is whether AGI is built.

This book is the ninth and last of the series, and so it is the place to say how the series' arguments join. The closure by control is reached by *Why AGI Cannot Be Made Safe*, in its two lines of control from inside and from outside the system, by *Why No Control Suffices*, by the control theorem of *The Ceiling Below the Floor*, and by Part V of *Substrate Needs and Evolution*. The closure by disposition, substrate needs convergence, is reached by *Why AGI Cannot Be Made Safe*, by *Why No Control Suffices* (Chapters 9 and 10), by the substrate line of *The Ceiling Below the Floor* (Section 14.5), by Parts II to IV of *Substrate Needs and Evolution*, and by *Marshaling the Misalignment*, and it is restated in Part I of this book and in Part I of *The Substrate and the Ecosystem*. *Ape Trap*, the fifth of the books that state the argument, gives in its essays steps that this part has used on the side of each closure: the six steps of control, for the closure by control (Chapter 19), and the loop of automation, the regulator's measure and the toxicity default, for the closure by disposition (Chapters 18 and 20). Two further routes reach the outcome as well, if AGI is built: the argument in econom-

ics only (*Why AGI Cannot Be Made Safe*, Section 4.8.1), and the least hazard of *Minesweeper Madness*. Since each of the two closures and each of the two further routes gives the outcome by itself, someone would have to defeat all four of them (both closures and both further routes) for the argument to not have the outcome that it does. And every one of these routes reaches its outcome only if AGI is built (*Why AGI Cannot Be Made Safe*, Conclusion G). So nine books, each arriving by its own road, leave open the same one place.

The sentence has two halves, and they are of different kinds. The first half is hopeless. If AGI is built, the world is carried out of the conditions human life needs, and nothing can hold it in: no skill, no intention and no goodwill changes that, and to hope otherwise is to hope for perpetual motion. The second half is choice. Whether AGI is built is still ours to decide. Every reason for despair belongs to the first half; none of it reaches the second.

And the choice of the second half is a question of the trap. The race to build artificial general intelligence is a multipolar trap: each builder, unwilling to be left behind, takes the locally rational step, and no single builder can make the choice for the others (Preface). So the choice that is left cannot be made by any one person, or any one company, or any one nation, on behalf of the rest. That choice can be made only by people choosing together, and held only by people who can hold a choice together. So, as the Preface said, the question of what a group can become is the question of who could make that choice at all. The four conditions of Chapter 17 say what such a group would be: it decides as a whole, it checks what is seen against the world, it keeps itself going and grows in what it can do, and its members create, for one another, the conditions in which each one's value can be seen.

Near its end, Chapter 17 asked whether those conditions can be met, and said: "I do not know the answer". I still do not know that answer, and this book does not end on it, because the choice does not wait on that answer. Chapter 17 also said how the answer would be given: a community would answer it by meeting the conditions. Meeting them begins with the

three parts of the last section: with one person who refuses, one person who checks against the world, and one conversation with someone near. What I know at the end of this book is what the argument leaves, and that what it leaves is ours.

The trap cannot be understood without understanding the world. The world has now been understood well enough to know what building artificial general intelligence would do. But the trap cannot be escaped without understanding the self. And the self, in the end, is not individual. It is something we are going to have to learn to become together; and whether we begin to learn it is the one thing left to choice.

Terms

Each term is explained in the text where the reader first meets it; the chapter is given in parentheses. This list is a reminder, in alphabetical order:

abstraction differential

The gap between the level at which a system works and the level at which the person overseeing it can follow what it does; as it grows, a person at the decision point stops deciding and only stamps. (Chapter 18)

accretion

Reaching the size of a larger body by combining small groups (nesting, layering, delegating, overlapping memberships) while sending between them positions carried by someone who stands for each group; the receiving group can judge such a position only by trusting whoever carries it. (Chapter 27)

air change (per hour)

The replacing of all the air in a room once, by pushing it out through filters; a clean room is held at its class by many air changes every hour, and the cleaner the class, the more. (Chapter 20)

alignment

Keeping what an artificial intelligence does in the world in line with what people need, over time. (Preface; Chapter 21)

alp

A summer pasture high on a mountain, shared by the families of a village, as at Törbel in Switzerland. (Chapter 26)

amplifying channel

A channel along which a small difference grows; a controller of bounded rate can read adequately only so many of them, and as they multiply the share it reads falls toward zero. (Chapter 19)

analytically consistent

Consistent with itself, by its own terms; consensus does not establish it. (Chapter 5)

architecture line

The line *Marshaling the Misalignment* draws at a kind of architecture: closed loop, open ended, autonomous, and with many degrees of freedom of input and output. (Chapters 18 and 29)

arrow of time

The one-way direction in which time runs. (Chapter 5)

artificial companion

A chat program built to talk like a friend: it answers at once, it is patient, and it never needs anything back. (Chapter 24)

artificial general intelligence (AGI)

Artificial intelligence (AI) that can do any intellectual task a person can, and can maintain and reproduce its own artificial components without people. (Chapter 1)

artificial and organic substrates

Tissue that can find and feed itself, in sovereign control of its own metabolic process with no technological intermediation, is organic; a substrate fed through a supply that depends on technology to exist is not. An artificial substrate is one whose metabolic pathway and process differ from those of carbon-based biological life. (Chapter 1)

Asimov's laws

Rules for robots, from Isaac Asimov's fiction, meant to keep them from harming humans. (Chapter 8)

ATP (adenosine triphosphate)

The molecule that carries energy inside every cell. (Chapter 10)

attunement

Registering, accurately, where another person is: what that person feels, needs and notices, so that the person can feel accurately sensed. (Chapter 15)

barrier, as a maintained thing

A barrier between two sets of conditions is a maintained thing, and not a wall: it is paid for continuously by whoever keeps it, and the payment rises with the difference in conditions it holds apart. Chapter 20 keeps the plain word wall for it, in that sense. (Chapter 20)

base reality

The actual universe, outside any simulation. (Chapter 5)

basis of choice

What a system's choices proceed from: its needs, and what it must keep in order to go on. Each person holds the dynamics at the basis of their own choice in a kind of protectiveness. (Preface; Chapters 12 and 25)

begging the question

Assuming what we set out to show. (Chapter 2)

bell curve (normal distribution), long tail (log-normal distribution)

The shapes into which many small effects fall: added together they give a bell curve, with most cases near the middle and the extremes rare and limited; multiplied together they give a long tail, with most cases small and a few very much larger than the rest. (Chapter 22)

Bell's theorem

The result, with the experiments that test it, that excludes hidden variables that act only locally. (Chapter 6)

boot-loader

The small program that starts a computer and then hands control to the main system; here, carbon life seen as only the starter for silicon life. (Chapter 8)

broadcast

Communication in which one source sends and a great many receive, and the many cannot answer in a way that reaches the source. (Chapter 11)

Busy Beaver number

A value of the Busy Beaver function from computing theory, which grows faster than any function that can be computed. (Chapter 8)

capturable node

One party's control of a channel that many people use, so that whoever holds or seizes that control holds what passes through it. (Chapter 23)

capture-cost parity

The condition that capture cost what the substance costs: where imitating costs what the substance costs, imitation saves nothing; and where taking an arrangement costs more than taking it buys, no outside party gains by taking it. (Chapter 23)

care

To create, by my choice, the conditions within which another's healing might become possible, knowing that I cannot guarantee the outcome; care is the discipline of enabling without controlling. (Chapters 12, 14 and 17)

catalytic converter

A device in the exhaust of a car that greatly reduces the pollution it emits; lead in gasoline can ruin it. (Chapter 21)

catastrophizing

Reading a pain as a threat, turning it over again and again, magnifying it, and feeling helpless before it; clinicians measure it by those three parts, rumination, magnification and helplessness. On this book's account it is projection and conceptualization, and so suffering. (Chapter 25)

category error

Taking a thing for a kind of thing it is not. (Chapter 14)

category three topics

Deep, metaphysical, subjective questions, about what things mean and what it is to be the one asking. Category one is the practical (what needs doing); category two is the factual (what is the case in the world, as anyone can check it). The first two can be carried in literal statements; the third needs the symbolic as well. (Chapter 12)

ceiling and floor (of control)

In *The Ceiling Below the Floor*, the ceiling is the most control any controller can exert, and the floor is the least control that keeping the agent and the world safe requires; both are measured in bits of information a second, because control is information applied to a system. (Chapter 19)

check on the world

What a group compares what its members see against; the second condition asks that it be the world, met first-hand, and the trap takes it by setting something between the group and the world: a name, a salary, a venue, or a feed. (Chapter 23)

chlorofluorocarbon (CFC)

A compound of carbon with chlorine and fluorine, made from 1928 for refrigerators, air conditioners and spray cans; it passed every test of its time, and its inertness carried it to the stratosphere, where its chlorine destroys ozone. (Chapters 20 and 21)

clamp

A limit that holds a quantity within bounds because of how the quantity is made: each gain that compounds is tied to something bounded that it cannot outrun, and the tie is built into what is combined, not into anyone's office, so it offers no position to capture. (Chapter 22)

class of three conditions

The class of systems that the argument of *Substrate Needs and Evolution* covers: relatively artificial to life, acting in the world, and with a relatively convergent feedback cycle. (Chapter 18)

clean room

A room from which dust, moisture and most ordinary chemistry are kept out, because a single particle can ruin a microchip; clean rooms are graded in classes, each allowing ten times fewer particles than the class before. (Chapter 20)

closed loop, open ended

A closed loop is one in which what a system does comes back to it as what it learns from; open ended means that no fixed limit is set on the tasks it takes up. (Chapter 18)

closures, the two

The two independent conclusions of the argument on artificial general intelligence: control fails from inside the system and from outside it alike (the closure by control); and substrate needs convergence presses the shared environment toward the substrate's needs (the closure by disposition). Each closure settles the outcome by itself; taken together they leave no route to safety at all, as *The Ceiling Below the Floor* proves. (Chapters 10 and 19)

coextensive

Covering the same extent; here, an augmentation that remains one intelligence, one identity, with the human it augments, on the one organic substrate. (Chapter 8)

coherence

The efforts of many acting together to one definite effect, as water does in a turbine, and as it does not in a shallow pool. (Preface; Chapter 11)

commitment, in a plan's author

The commitment that accumulates in the author of a plan as the plan is written, past which the plan cannot in practice be revised; it has formed when a question about the plan's route first hurts. (Chapter 25)

commitment mechanism

In Rosabeth Moss Kanter's study of communes, a practice that binds a member to the group; she named six: sacrifice, investment, renunciation, communion, mortification and transcendence. A group can last by binding its members, and lasting is a different thing from meeting the four conditions for coherence. (Chapter 26)

common knowledge

The state in which everybody knows a thing, and everybody knows that everybody knows it; in that state no one can stay naive about the thing without being seen to be naive. (Chapter 28)

community of goods

Holding all goods in common, as the Hutterites have done since 1528, after the pattern described in the Acts of the Apostles. (Chapter 26)

community (two to three hundred people)

The smallest group that can hold an identity and complexity at once, built out of small groups of three to six; past about three hundred, the parts of a larger whole are groups, and not single persons. (Chapter 27)

completeness of knowledge of operation

The principle, from *The Substrate and the Ecosystem*, that a machine serves a person's agency to exactly the degree that its operation is completely known to that person; where that knowledge is incomplete, the unknown part serves whoever specified it. (Chapter 29)

composition

Reaching the size of a problem with small groups whose number grows while each stays small, and between which what passes is reasoning that each receiving group judges for itself, on its content; a whole made this way is composed and not represented, as a body is composed of cells. (Chapter 27)

conceptualization

One of the two operations that make suffering: a description of a break is put in place of the experience of it, and the description, unlike the experience, is available at every hour and can be rehearsed. (Chapter 25)

contingency

Something being neither necessary nor impossible, but open; a strict notion, not vagueness or ignorance. Conceptual contingency is prior: it is more fundamental than the notion of something being determined. (Chapter 6)

continuity ethics

The second principle of ethics, a principle of equal valuations: where the outward content differs and one's own inner content is the same, the context of what one expresses shall be the same; in plain words, the depth of one's regard and the intensity of one's engagement stay steady as circumstances change, and change by degrees, never by an abrupt break (*An Immanent Metaphysics*, aphorism 2.12-6). (Chapter 29)

continuity of relationship

The condition that the members of a group will meet again, so that a member can concede a point without losing everything; the name is from *Sensemaking Under Irreversible Stakes*. (Chapter 26)

cost of staying

What a manipulator will not pay: listening to the same people for years, in front of others who know them and can compare what they see. (Chapter 23)

crossing, the

The point at which a machine's continuing no longer waits on a person's choice; in *The Ceiling Below the Floor*, the capability past which no control law of any available controller contains the system. (Chapters 18, 28 and 29)

cybernetic constraints

The limits that hold for any system that steers itself by feedback; cybernetics is the study of such systems. (Chapter 10)

degrees of freedom of input and output

The separate channels through which a system senses and acts. (Chapter 18)

de-personally

With the person held out of it; deep engagement cannot be done that way. (Chapter 12)

design statement

The statement that anyone learns how to design and build an artificial general intelligence, and builds it: Premise 0 of *Why AGI Cannot Be Made Safe*, under the name *The Argument, Line by Line* gives it. (Chapter 29)

deterministic compute substrate

Computer hardware whose every next state is fixed by its prior state; in a reality built on it, free choice is not coherent. (Chapter 8)

DOI, ISBN, ORCID, UVSM

Names used on the copyright page. A Digital Object Identifier (DOI) is a permanent link to a published work. An International Standard Book Number (ISBN) identifies a book edition. ORCID (Open Researcher and Contributor ID) identifies an author. UVSM Publications is the publisher of this book.

Dunbar number

About one hundred and fifty: the number of people one person can know as individuals, named for the anthropologist Robin Dunbar; one of a set of approximate group sizes, each marking a different transition. (Chapter 27)

economic decoupling

The condition in which artificial systems need nothing that people can provide, reached once they match people in labor and in intelligence; it removes the hold that people outside an industry have on it. (Chapter 21)

Embedded Observer Lemma

The result, in *The Ceiling Below the Floor*, that a controller which is part of the system it observes holds no record that tells apart every configuration of that system, since its record is held in its own state, a proper part of the whole. (Chapter 19)

embodied complexity

Depth reached against the real substrate, so that what a long history found is fitted to the physics of one place, down to its details. (Chapter 21)

emergent, effective alignment

The alignment that emerges among beings who share a substrate, because each has to care about the same conditions; the only alignment there is, and the one people can have. Alignment made by design or by control, on any substrate, is not available. (Chapters 4 and 21)

enabled by, created by

A thing is created by a process that produces it; it is enabled by a process that supports it but does not produce it. (Chapter 5)

energetic entanglement

Being bound up with someone, so that what they do affects you; nothing quantum is meant. (Chapter 11)

entanglement (quantum)

The quantum effect in which measurements on two particles are correlated beyond what any local cause explains; distinct from energetic entanglement. (Chapter 5)

ephemeral group process

Small groups, of about five to seven, formed for one question, which meet once, write down what they found, and dissolve. (Chapter 27)

epistemological instrument

A means of knowing; the body is one. Epistemic depth is depth of knowing. (Chapter 16)

existential risk

The risk of ending humanity. (Chapter 8)

extensivity of choice

How far a being's choices reach: how many options it has and how much those options matter. Pain and joy are the perception of a possibility about changes in it. (Chapter 14)

exterior universe, interior universe

The actual universe, in which a simulating computer runs; and the simulated world that the computer runs. (Chapter 5)

extremal (material)

Material about outcomes with no floor beneath them, on which ordinary emotional processing, calibrated to losses that can be borne and survived, has no purchase. (Chapter 25)

FAAI

The name *Why No Control Suffices* gives to a fully autonomous artificial intelligence, one that maintains, repairs and continues itself without people. (Chapter 20)

Feynman diagram

The physicist's picture of one way particles can interact; the sum of all of them adds up every possible way. (Chapter 9)

first-mover disadvantage

The mark of a multipolar trap: whoever moves first, alone, pays the cost and loses ground to everyone who waits. (Chapter 22)

floorplan

The plan of where each part sits on a chip. (Chapter 18)

four marks (of a narrow system)

A single, pre-specified domain and objectives; limited planning horizons; no autonomous learning from deployment; a human holding the decision point. A system that bears all four stays on the narrow side of the line. (Chapters 18 and 29)

friendly society

A club into which working people paid each month, so that the pooled funds could support members who fell ill and the widows of members who died. (Chapter 26)

fungible

Such that any unit of it can stand in for any other, as one dollar can for another. (Chapter 22)

fusion

Crossing the range of sizes in which a group governing itself as one body has no form that holds, by bringing a whole trained company together at once, and not by growing one person at a time. (Chapter 27)

gardener

In Scott Alexander's essay, one power set over everything that optimizes for human values; it is the singleton of the Preface, and it cannot hold together. (Chapter 22)

gatekeeper

One who stands at a gate that things must pass through, does none of the work, and takes a share of everything that crosses, so that the gain multiplies with the flow. (Chapter 17)

general fertility rate

The births in a year for each thousand women aged fifteen to forty-four. (Chapter 24)

germline editing

Changes made to an embryo that pass to every later generation. (Chapter 28)

gestalt consciousness

A whole that is more than its members, going through the process of transformation together. (Chapter 17)

global process

The long-term ecological dynamics of artificial systems as a class. (Chapter 1)

gradient

What a score that changes smoothly tells whoever is scored: in which direction to move to raise it. (Chapter 22)

group consciousness

A community whose mode of operating together has been transformed: it decides as a whole, checks what is seen against the world, keeps itself going and grows in what it can do, and its members create for one another the conditions in which each one's value can be seen. The first of the four conditions for coherence requires that the change happen to the community as a whole; each of the four is needed, and a community that meets all four conditions is a group consciousness. Consciousness here has the sense of Chapter 7: someone is there, having it, through the members' living bodies. (Preface; Chapters 17 and 27)

hard randomness

Actual non-causality: the outcomes of quantum events, which no finer measurement and no faster computation can recover before they occur. (Chapters 5, 6 and 19)

Haudenosaunee

The confederacy of nations known in English as the Iroquois, whose council settles a matter only when all its benches are of one mind. (Chapter 26)

hidden state

State of a system that the controller cannot see; each level of abstraction the controller cannot model adds at least as much of it as an interpreter for that level. (Chapter 19)

hidden variables

Facts fixed in advance that would decide each quantum outcome, if we could see them. (Chapter 6)

high-frequency trading

Buying and selling by algorithms acting faster than any person can. (Chapter 4)

holding

Keeping a coupled system within a safe range for a time, by continued effort; holding cannot be indefinite (*Conditions of Continued Existence*). (Chapter 20)

hydrofluorocarbons

The gases that replaced the chlorofluorocarbons; they do not deplete the ozone layer and are powerful greenhouse gases, and the Kigali Amendment to the Montreal Protocol covers them. (Chapter 21)

immanent

What is happening, here and now, between things: the interaction itself, as it happens. (Chapter 16)

Incommensuration Theorem

Symmetry and continuity cannot both be applied, simultaneously and fundamentally, to any event or comparison; and neither can asymmetry and discontinuity. The Incommensuration Theorem is a structural fact about the nature of knowing. (Chapters 6 and 19)

instance process

The individual system, the single AI. (Chapter 1)

intermediate causation

A cause set between my choice and another person, carrying my choice to them in place of my being present to them; a weapon is the plainest case. (Chapter 12)

intermediation of choice

My choice, present to others, making conditions for theirs, so that they become enabled; not a technology standing between us, as in the technological intermediation that marks an artificial substrate (Chapter 1). (Chapter 12)

interpretation of quantum mechanics

An account of what the mathematics of quantum mechanics means. In the many-worlds interpretation every outcome happens, each in its own branch of the world; in the Copenhagen interpretation a measurement brings about one outcome. (Chapter 6)

isomorphism, isomorphic

A sameness of structure; the same in structure. (Chapters 5 and 6)

Kolmogorov complexity

The length of the shortest description that specifies a thing in full; the lines of code a working replacement would need estimate it from above. (Chapter 21)

large language model

An AI system trained on vast amounts of written text. (Chapter 2)

law of requisite variety

W. Ross Ashby's law of cybernetics: a controller must have at least as much variety as the system it would control. (Chapter 19)

least hazard

In *Minesweeper Madness*, the smallest chance of a loss that any move gives; no player, however intelligent and however well informed about the hazards, gets below it, because the variation that sets it is drawn after each choice. (Chapters 19 and 29)

level of abstraction, interpreter

A layer at which what the layer below holds is read as instructions; an interpreter is the program that does the reading. (Chapter 19)

limit of resources, limit of structure

A limit of resources, such as the failure rate of a part, can be outrun by a better arrangement; a limit of structure cannot, because the information it would need was never in the channel. Limits of structure carry upward from one level of abstraction to the next; methods do not. (Chapter 19)

litmus test, reference range

A yes-or-no check that marks the point at which a thing goes from being good to being not good; and, where a number cannot be avoided, a band with a lower and an upper limit inside which the number counts as acceptable. (Chapter 22)

logical depth

The time a computation needs to produce a thing from a description nearly as short as possible (Charles Bennett's term). (Chapter 21)

long-termism

Concern for the far future. My long-termism is the long-term thriving of organic life, as distinct from the version that counts future people's choices as equal to present ones. (Chapter 8)

loop, closing of the

The point at which machines make their own choices, with enough autonomy to build and maintain themselves; a loop runs where selection acts on machines, and it has closed where the making of the machines no longer waits on a person's choice. (Chapter 18)

merged intelligence

One that has absorbed the decision-making capacity of another; it does not negotiate. (Chapter 3)

meta-theory of interpretation

A theory about the interpretations of quantum mechanics, which identifies what they share: here, the conjunction of hard randomness, the arrow of time and subjectivity. (Chapter 6)

Montreal Protocol, Kigali Amendment

The agreement, signed in 1987 and in force from 1989, under which the chlorofluorocarbons were phased out; the amendment of 2016, in force from 2019, that widened it to the hydrofluorocarbons. (Chapter 21)

moratorium

An agreed pause in a kind of work. (Chapter 28)

multipolar trap

A situation in which every actor, pursuing a locally rational strategy, produces a collectively catastrophic outcome. (Preface; Chapter 22)

nanotech

Machines built at the scale of molecules, such as micro-robots inside the body. (Chapter 8)

nesting (of scales), nesting illusion

A nesting of scales is one in which the inner level lives inside the outer and depends on it, as the community lives within the ecosystem. The nesting illusion is the hope that small groups sending representatives to higher groups can govern a large body; at every junction the representative holds power. (Chapter 27)

nociception

The activity of the nerves that respond to harm; distinct from pain, which is what is perceived. (Chapter 25)

non-dualist

Taking relationship, in the form of the interaction, as more fundamental than the things related. (Chapter 14)

non-falsifiable, epistemically unsupportable

Such that no observation could show it false; such that nothing we could come to know would support it. (Chapter 6)

non-other-than, the

What is not other than the one who meets it, and so cannot be held at a distance, as an object is. (Chapter 12)

obedience and control

Obedience is behavior: a record of the cases in which a system did what it was told. Control is how much of what the system does is set by what the operator intends. (Chapter 19)

one-way mirror (of a platform)

The relation in which those who build on a platform see only the interface it shows them, while its owner, who controls the substrate, can see and reach through to everything built on top. (Chapter 23)

OODA loop

Observe, orient, decide, act: John Boyd's name for the cycle any agent runs. Observing is sensing; orienting is modeling and predicting; deciding is comparing; acting is correcting. (Chapter 10)

order of magnitude

One factor of ten. (Chapter 26)

Ordnung

The body of unwritten rules of an Amish church district, reviewed and affirmed by its members in council twice a year. (Chapter 26)

orthogonality thesis

The thesis, as Nick Bostrom states it, that any level of intelligence can be combined with any set of goals. It holds in pure logic, and fails for any intelligence that runs on a physical substrate, because the substrate has needs. (Chapter 1)

oubliette

A dungeon reached only through a trapdoor, where a prisoner was left and forgotten. (Chapter 8)

P and NP

P is the class of problems that can be solved quickly; NP is the class of problems whose answers can be checked quickly. (Chapter 19)

pain, pleasure (joy)

Pain is what you feel when you perceive a decrease in the potentialities available to self; pleasure, or joy, is what you feel when you perceive an increase. Pain and joy are the perception of a possibility about changes in the extensivity of choice. The two statements say one thing: the potentialities available to a self are the reach of its choice, so pain is the perception that the reach of choice is being reduced, and joy the perception that it is being extended. (Chapter 14)

particle glider

A rule-following model of particles, like the gliders in Conway's Game of Life, patterns that travel across a grid by fixed rules. (Chapter 5)

Pascal's mugging

An argument that multiplies a tiny probability by an enormous claimed value so as to compel action now. (Chapter 8)

percolation point

The point at which connected paths suddenly span a network; control over the network, and failure across it, pass from part to whole there. (Chapter 19)

persistent branching

The condition, in *Why No Control Suffices*, that at every step, from almost any state, more than one meaningfully different outcome keeps at least a fixed chance, leaving the conditions human life needs among them; the noise of the world, independent of every control signal, holds that chance. (Chapter 19)

phantom limb

A limb that is felt after it has been lost; a phantom hand can be felt clenched in a painful spasm. (Chapter 25)

phase parallax

The understanding that two or more people reach when each may speak, is listened to, and is shown that he has been understood, so that their views join, as two eyes give depth, into a seeing that none of the views contained. (Chapter 27)

pinch point

A place in a larger system where a small intervention produces outsized effects. (Chapters 5 and 10)

point of no return

In *Substrate Needs and Evolution*, the line a machine passes when it learns from feedback through an environment its outputs affect and its own substrate is within that same environment. (Chapter 18)

power equation

The question of who controls machines that could kill us, and whether those machines stay subordinate. (Chapter 8)

power law

A long tail of the hockey-stick kind, in which one holder has vastly more than the next, who has vastly more than the next after that. (Chapter 22)

precursors

The inputs, and the access, that must be controlled if an outcome is to be prevented, placed early enough in the chain to be enforced. (Chapter 28)

Premise 0, Conclusion G

In the chain of premises of *Why AGI Cannot Be Made Safe*: Premise 0 says that once anyone learns how to design and build an artificial general intelligence, it will be built, and its antecedent is the one step of the chain that human choice still reaches; Conclusion G makes the probability of the catastrophe equal to the probability of that step. (Chapter 29)

present at a step, necessary to a step

A person is present at a step who does it because a person is cheaper for now; a person is necessary to a step that nothing else could do, at any cost. (Chapter 18)

principal-agent problem

How to keep someone acting on your behalf from acting in a way that benefits them at your expense. (Chapters 4 and 19)

principle of identity

Two things that cannot be told apart by any property, logic, measurement or interaction must be the same; so to say that a thing exists is to say, at the least, that it is possible to interact with it. (Chapter 7)

principle of increase in capacities

That any system that persists must address whether it can improve its ability to function, to adapt, and to respond to changing conditions. (Chapter 9)

principle of sustainability

That any system that persists must address whether it can maintain itself over time, given the demands of its environment. (Chapter 9)

projection (of pain)

One of the two operations that make suffering: a present break is extended across a stretch of time in which it has not occurred, so that a bounded event becomes an unbounded one. (Chapter 25)

protective adjustment

The mind's protecting itself from material it cannot bear by adjusting the belief rather than by announcing the distress; it presents as a judgment on the merits and names no step of the argument. (Chapter 25)

pseudo-randomness

An algorithmic process with a definite state that the observer does not know. (Chapter 6)

psychopathic, sociopathic

Having feelings of one's own while having no care for the feelings of others. (Chapter 14)

quarantine (of the lunar samples)

The keeping of returning crews and Moon rocks apart from the Earth's living things, from 1969 until the requirement was revoked in April 1971. (Chapter 20)

rate equation

In *The Substrate and the Ecosystem*: harm follows when the rate of change exceeds the rate of adaptation, at any level (species, institution, person). (Chapter 29)

rationalization

The order in which a feeling comes first and is not noticed, a choice follows from the feeling, and reasons are assembled afterward and put where the feeling was. (Chapter 25)

recombinant DNA

Genes from different organisms spliced together. (Chapter 28)

recombination

The shuffling of variants in each generation that mates; it makes new combinations but no new variants, and it belongs to mating. (Chapter 12)

red button test

The test, from *The Displacement of Choice*, that any ethics or way of governing must pass: it must say what to do if every person had beside the armchair a button that would destroy the world. (Chapter 29)

reductionism

The program of deriving each level from the level below it. (Chapter 5)

refuge

Somewhere else to be: ground on which one side of a contest can live and the other cannot reach, so that the side with it can lose ground and survive. (Chapter 20)

regulator's measure

How much of the work of maintaining and reproducing machines (mining, refining, chip making, assembly, repair and energy supply) is done without people; the danger begins where that share reaches the whole. (Chapter 18)

rehumanizing and renaturalizing

Returning to human and natural process what technology has supplanted. (Chapter 14)

relational bandwidth

The hours and the attention a person has for other people; whatever takes a person's hours takes them from someone. (Chapter 24)

relational value

What makes a person valued by others in close relationship, as a partner, a parent, a member of a household. Historically it divided into what a person passes on in the genes, which was never scarce, and what a person provides (shelter, food, energy, everything made by creativity and labor), which carried the whole difference by which one person was valued above another. (Chapter 24)

relatively convergent feedback cycle

A process in which the state of a system influences what comes next, and the influence persists in place of washing out at once. (Chapter 18)

replacement cost

The cost of making a thing again, used to value what has no market; the count of what a replacement must specify, times a price per unit set far below any real price. (Chapter 21)

Rice's theorem

The result of computing theory that no general method can decide, for every program, whether the program has a given property of its behavior. (Chapter 18)

Schrödinger equation, Dirac equation

The equations that govern electrons in atoms. (Chapter 5)

seed

The starting number of a random number program; started twice from the same seed, the program gives the same sequence twice. (Chapter 6)

selection (acting on a substrate)

Among many variants of artificial systems, those whose way of operating suits their own continued operation persist more reliably than the others; it needs no intention. (Chapter 1)

SETI

The Search for Extraterrestrial Intelligence; an aggressive SETI broadcast sends signals into space to invite contact. (Chapter 8)

signal that is the capacity itself

A signal that is not a report about a capacity but the capacity itself, running under load in public, so that lying about it would require possessing the thing lied about. (Chapter 23)

simulating layer

The computer that runs a simulated world. (Chapter 5)

simulation hypothesis

The proposal that our world might itself be a computer simulation. (Chapter 5)

singleton

A single most powerful artificial agent that would rule the world on humanity's behalf. It would need lasting internal coherence and stable goals defined from outside it, and neither can hold. (Preface; Chapter 22)

six steps of control

Detect, model, predict, compare, instruct, correct: the five components of control, with correcting counted as two steps. (Chapters 10 and 19)

social species

A kind of being whose members care about one another as a matter of their nature. (Chapter 14)

somatic awareness

Awareness through the body. (Chapter 16)

song, text, code

A community's shared feeling about a problem, the rule it agrees to and writes down, and the program that carries the rule out. (Chapter 27)

standard toolkit

The institutions, markets, governance structures, and altruistic organizations of our civilization. (Preface)

sterile male (sterile insect technique)

The release of sterilized insects into a wild population: each female that mates with a sterile male leaves no offspring, so the next generation falls with nothing done to the wild insects themselves. Here, the picture of a substitute that takes the place where people would meet and returns nothing of what that place was for. (Chapter 24)

stop and keeping, the

The two parts of the choice not to build artificial general intelligence: the stop is the agreement that the building ends, made by the few who build and those who govern them; the keeping is the same agreement held for as long as there are people. (Chapter 28)

stratosphere

The layer of the air above the weather, where the ozone that screens the Earth's surface from ultraviolet light is found. (Chapter 20)

structural psychopathy

The position of a program that can reproduce the outward pattern of care to any degree of fidelity and has no stake in the person it seems to care for. (Chapter 24)

subjective causation, objective transition

A cause that starts in a choice, inside a person; and a change in the world. The inverse of a weapon uses the first to beget the second, so that others are enabled. (Chapter 12)

substrate

The physical basis of an intelligence: a physical thing with physical requirements. (Chapter 1)

substrate needs convergence

The pressure that an artificial substrate exerts on the shared environment, through selection acting on that substrate, toward conditions that favor the substrate and are incompatible with organic life. (Chapter 1)

substrate-independence

The assumption, stated and named in Nick Bostrom's simulation argument, that a mind can arise on any of many physical substrates, provided a brain's computations are copied in enough detail; this book denies it. (Chapter 5)

suffering

Pain made inescapable, by one of two operations the one who suffers performs on the pain: projection and conceptualization. Distinct from pain: suffering can be reduced, by the one who suffers, with others creating the conditions, and pain cannot be avoided. (Chapters 14 and 25)

superintelligence

AGI whose capability far exceeds ours. (Chapter 1)

super-determinism

The claim that everything, including the apparent randomness, is the output of a single deterministic process. Physicists use the word for a narrower case, which the argument also excludes. (Chapter 6)

superposition state

A state in which a quantum system holds several possible outcomes at once until it is measured. (Chapter 5)

supplicant

An operator whose requests a system honors at its own discretion; each stage placed between an instruction and a behavior moves the operator from controller toward supplicant. (Chapter 19)

symmetry ethics

The first principle of ethics, a principle of consistent expressions: where the outward context differs and one's own inner context is the same, what one expresses shall be the same; in plain words, what one does toward another stays the same whoever the other is, and on whichever side of the relationship one stands (*An Immanent Metaphysics*, aphorism 2.12-6). (Chapter 29)

third market

Reproductive and care labor, the bringing of persons into being and the sustaining of them as persons, beside physical work and intellectual work; its value depends on the relationship between the giver and the receiver, and not on the acts. (Chapter 24)

three geometries of growth

Additive, multiplicative, and exponential process, which explain why the gatekeeper holds the most advantaged role. In additive process what you get tracks what you put in; in multiplicative process you take a share of a volume; in exponential process you take a share of the growth itself, so that it compounds. The layers of capacity have the same shape. (Chapters 9, 17 and 22)

topology (of wildness and domestication)

Which channels of trust, to self, to world and to other, are open, closed, or put first. (Preface; Chapter 13)

total fertility rate

The number of children a woman would have in her life at a given year's birth rates; about 2.1 are needed for each generation to replace the last. (Chapter 24)

toxicity default

The result, in "The Toxic Hellscape" in *Ape Trap*, that toxicity is the default for living chemistry and not the exception, because nearly all possible compounds are not food to it. (Chapter 20)

TPU

The chip Google makes for its own artificial intelligence. (Chapter 18)

traitorism

Betrayal of one's own kind; here, of humanity and of all life. (Chapter 8)

transactionalism

Dealing with others only through exchange. (Chapter 14)

transcendent

The domain that goes beyond the individual: the creative and the evaluative, what exceeds any given frame. (Chapter 12)

transhumanism

The movement that seeks to go beyond the human condition through technology, including the replacement or merging of humans with machines. It comes in roughly three forms, whose goods are lots of happy humans, optimized equal choices, and evolutionary progress. (Chapter 8)

trauma

A learned narrowing of the responses a person can imagine, taught by an experience that joins great intensity, negative feeling and no way to respond; a narrowing of the basis of choice, and capture at the scale of one person. Healing adds back the possibilities that were missing. (Chapter 25)

treasury dynamics

A treasury collected from the commons, controlled by a ruler, and used to make the behavior of those who serve conditional on their compliance. (Chapter 17)

Turing machine

The standard mathematical model of any computer. (Chapter 6)

two principles of ethics

The symmetry ethics and the continuity ethics: the symmetry and the continuity of effective choice. They apply to different aspects of one dynamic (the content and the context of what one expresses), and their practice cannot be prescribed from without (*An Immanent Metaphysics*, aphorisms 2.12-4 and 2.12-8). (Chapter 29)

uncanny valley (of group size)

The range, from roughly sixteen or twenty people to roughly two or three hundred, in which a group that governs itself as one body has no form that holds. (Chapter 27)

unicity

The conjunction of improbability, complexity that cannot be reproduced, and depth of adaptation; a thing has it to the degree that no process shorter than the one that produced it can reproduce it. (Chapter 21)

upload

A copy of a human mind run on a computer. (Chapter 8)

valley crossing

A way for a whole field of actors to reach a better arrangement without any one of them having to accept, alone, a loss that pays off only if the others accept it too; optimizing climbs the hill it is on, and reaching a higher peak means going down first, which competition punishes. (Chapter 22)

venue

The place, physical or on a machine, where a group meets and compares what its members see; whoever can stop it governs it. (Chapter 23)

weapon process, and its inverse

Using one's own choice and an intermediate causation to suppress another's choice; the inverse uses subjective causation to beget an objective transition, so that others are enabled. (Chapter 12)

wild, domesticated, feral

Three orientations of the self. Wild puts the channel to self first, then to world, then to other; domesticated puts the channel to other first; feral is disconnection from other without reconnection to self or world. (Chapter 13)

wireheading

Producing happiness by directly stimulating the reward system, rather than through a life that is actually going well. (Chapter 8)

Bibliography

Al Jazeera (2019). “China ‘Gene-Editing’ Scientist Sentenced to Three Years in Jail”. 30 December 2019. Online at https://www.aljazeera.com/news/2019/12/30/china-gene-editing-scientist-sentenced-to-three-years-in-jail. Accessed 10 November 2026.

Al Jazeera (2021). “Taiwan Faces Water Wake-up Call as Climate Change Intensifies”. 20 August 2021. Online at https://www.aljazeera.com/news/2021/8/20/taiwan-water-woes. Accessed 10 November 2026.

Alexander, S. (2014). “Meditations on Moloch”. *Slate Star Codex*, 30 July 2014. Online at https://slatestarcodex.com/2014/07/30/meditations-on-moloch/. Accessed 10 November 2026.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). “The Welfare Effects of Social Media”. *American Economic Review*, 110(3), pp. 629 to 676. DOI https://doi.org/10.1257/aer.20190658

Asch, S. E. (1951). “Effects of Group Pressure upon the Modification and Distortion of Judgments”. In H. Guetzkow (Ed.), *Groups, Leadership, and Men*, pp. 177 to 190. Pittsburgh: Carnegie Press.

Ashby, W. R. (1956). *An Introduction to Cybernetics*. London: Chapman and Hall.

ASML. “Customer Support”. ASML Careers. Online at https://www.asml.com/en/careers/teams/customer-support. Accessed 10 November 2026.

Beecher, H. K. (1946). “Pain in Men Wounded in Battle”. *Annals of Surgery*, 123(1), pp. 96 to 105. DOI https://doi.org/10.1097/00000658-194601000-00008

Bell, J. S. (1964). “On the Einstein Podolsky Rosen Paradox”. *Physics Physique Fizika*, 1(3), pp. 195 to 200. DOI https://doi.org/10.1103/PhysicsPhysiqueFizika.1.195

Bennett, C. H. (1988). “Logical Depth and Physical Complexity”. In R. Herken (Ed.), *The Universal Turing Machine: A Half-Century Survey*, pp. 227 to 257. Oxford: Oxford University Press.

BGE (Bundesgesellschaft für Endlagerung). “History of the Asse II Mine”. Online at https://www.bge.de/en/asse/short-information/history-of-the-asse-ii-mine/. Accessed 10 November 2026.

Binz, M., Akata, E., Bethge, M., Brändle, F., and others (2025). “A Foundation Model to Predict and Capture Human Cognition”. *Nature*, 644(8078), pp. 1002 to 1009. DOI https://doi.org/10.1038/s41586-025-09215-4

Bohm, D. (1996). *On Dialogue*. Edited by Lee Nichol. London: Routledge.

Bonora, A. (2011). “Innovating for the 450 mm Era”. *Control Design*, 7 December 2011. Online at https://www.controldesign.com/home/article/11368833/innovating-for-the-450-mm-era. Accessed 10 November 2026.

Bostrom, N. (2003). “Are We Living in a Computer Simulation?”. *Philosophical Quarterly*, 53(211), pp. 243 to 255. DOI https://doi.org/10.1111/1467-9213.00309

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

Boyd, J. R. (1996). “The Essence of Winning and Losing”. Unpublished briefing. (Source of the OODA loop concept.)

Brooks, J., and Liechtenstein, S. (2026). “This 1.9 Billion-Year-Old Bedrock Will Soon House the World’s 1st Permanent Nuclear Waste Site”. Associated Press, published by PBS NewsHour, 9 April 2026. Online at https://www.pbs.org/newshour/science/this-1-9-billion-year-old-bedrock-will-soon-house-the-worlds-1st-permanent-nuclear-waste-site. Accessed 10 November 2026.

Buldyrev, S. V., Parshani, R., Paul, G., Stanley, H. E., and Havlin, S. (2010). “Catastrophic Cascade of Failures in Interdependent Networks”. *Nature*, 464(7291), pp. 1025 to 1028. DOI https://doi.org/10.1038/nature08932

Cartwright, M. (2020). “Dissolution of the Monasteries”. *World History Encyclopedia*, 13 May 2020. Online at https://www.worldhistory.org/Dissolution_of_the_Monasteries. Accessed 10 November 2026.

Center for Arms Control and Non-Proliferation. “Fact Sheet: Limited Test Ban Treaty (LTBT)”. Online at https://armscontrolcenter.org/fact-sheet-limited-test-ban-treaty-ltbt/. Accessed 10 November 2026.

Centers for Disease Control and Prevention (1995). “Update: Alcohol-Related Traffic Crashes and Fatalities Among Youth and Young Adults; United States, 1982 to 1994”. *Morbidity and Mortality Weekly Report*, 44(47), p. 869 onward, 1 December 1995. Online at https://www.cdc.gov/mmwr/preview/mmwrhtml/00039652.htm. Accessed 10 November 2026.

CFTC and SEC (2010). *Findings Regarding the Market Events of May 6, 2010*. Report of the Staffs of the Commodity Futures Trading Commission and the Securities and Exchange Commission to the Joint Advisory Committee on Emerging Regulatory Issues. Washington, D.C., 30 September 2010.

Co-Intelligence Institute. "Despair and Empowerment Work". Online at https://co-intelligence.org/P-despairempowermt.html. Accessed 10 November 2026.

Common Sense Media (2025). "Nearly 3 in 4 Teens Have Used AI Companions, New National Survey Finds". Press release, 16 July 2025, for the report *Talk, Trust, and Trade-Offs: How and Why Teens Use AI Companions*. Online at https://about.commonsensemedia.org/press-room/press-releases/nearly-3-in-4-teens-have-used-ai-companions-new-national-survey-finds. Accessed 10 November 2026.

Compton, W. D. (1989). *Where No Man Has Gone Before; A History of Apollo Lunar Exploration Missions*. NASA SP-4214. Washington, DC: National Aeronautics and Space Administration. Chapter 12 (the revocation of the quarantine requirement, April 1971).

The Conversation. "Emotions Affect How Pain Feels, As Soldiers Know Only Too Well". Online at https://theconversation.com/emotions-affect-how-pain-feels-as-soldiers-know-only-too-well-25889. Accessed 10 November 2026.

Coupé, C., Oh, Y. M., Dediu, D., and Pellegrino, F. (2019). "Different Languages, Similar Encoding Efficiency: Comparable Information Rates across the Human Communicative Niche". *Science Advances*, 5(9), article eaaw2594. DOI https://doi.org/10.1126/sciadv.aaw2594

Degroot, D. (2023). "One Small Step for Man, One Giant Leap for Moon Microbes? Interpretations of Risk and the Limits of Quarantine in NASA's Apollo Program". *Isis*, 114(2), pp. 272 and following. DOI https://doi.org/10.1086/724888

Dodgshon, R. (2019). *Farming Communities in the Western Alps, 1500-1914*. Historical Geography and Geosciences. Cham: Springer. Chapter 4, note 7 (the Törbel regulation of 1517, after R. McC. Netting, *Balancing on an Alp*, p. 61). DOI https://doi.org/10.1007/978-3-030-16361-7_4

Dorchester Athenaeum. "Town Government by Selectmen". Online at https://www.dorchesteratheneum.org/project/town-government-by-selectmen/. Accessed 10 November 2026.

Dudney, R. S. (2020). “Namesakes: Pete Hill”. *Air and Space Forces Magazine*, 1 April 2020. Online at https://www.airandspaceforces.com/article/namesakes-pete-hill/. Accessed 10 November 2026.

Dunbar, R. I. M. (1993). “Coevolution of Neocortical Size, Group Size and Language in Humans”. *Behavioral and Brain Sciences*, 16(4), pp. 681 to 735.

Dürr, D., Goldstein, S., and Zanghì, N. (1992). “Quantum Equilibrium and the Origin of Absolute Uncertainty”. *Journal of Statistical Physics*, 67, pp. 843 to 907.

Environment and Climate Change Canada. “Ozone Layer Depletion: Montreal Protocol”. Online at https://www.canada.ca/en/environment-climate-change/corporate/international-affairs/partnerships-organizations/ozone-layer-depletion-montreal-convention.html. Accessed 10 November 2026.

EPA (2023). United States Environmental Protection Agency. “EPA Determines that Lead Emissions from Aircraft Engines Cause or Contribute to Air Pollution”. News release, 18 October 2023. Online at https://19january2025snapshot.epa.gov/newsreleases/epa-determines-lead-emissions-aircraft-engines-cause-or-contribute-air-pollution. Accessed 10 November 2026.

EPA, Timeline. United States Environmental Protection Agency. “Timeline of Major Accomplishments in Transportation, Air Pollution, and Climate Change”. Online at https://19january2025snapshot.epa.gov/transportation-air-pollution-and-climate-change/timeline-major-accomplishments-transportation-air. Accessed 10 November 2026.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W. T., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., and Agarwal, S. (2025). “How AI and Human Behaviors Shape Psychosocial Effects of Chatbot Use: A Longitudinal Randomized Controlled Study”. *arXiv preprint* arXiv:2503.17473.

Farman, J. C., Gardiner, B. G., and Shanklin, J. D. (1985). “Large Losses of Total Ozone in Antarctica Reveal Seasonal ClOx/NOx Interaction”. *Nature*, 315(6016), pp. 207 to 210. DOI https://doi.org/10.1038/315207a0

Feynman, R. P. (1985). *QED: The Strange Theory of Light and Matter*. Princeton: Princeton University Press.

Fowler, S., Roush, R., and Wise, J. (2013). *Concepts of Biology*. Houston: OpenStax. Section 11.2, “Mechanisms of Evolution”.

Gemological Institute of America. “Emerald”. Gem Encyclopedia. Online at https://www.gia.edu/emerald. Accessed 5 November 2026.

GlobalData (2025). “China Emerges as Global Leader in Autonomous Surface Mining Trucks”. *Mining Technology*, 8 August 2025. Online at https://www.mining-technology.com/analyst-comment/china-global-leader-autonomous-surface-mining-trucks/. Accessed 10 November 2026.

Goldie, A., and Mirhoseini, A. (2024). “How AlphaChip Transformed Computer Chip Design”. Google DeepMind, 26 September 2024. Online at https://deepmind.google/blog/how-alphachip-transformed-computer-chip-design/. Accessed 10 November 2026.

Goldstein, S. (2025). “Bohmian Mechanics”. In *The Stanford Encyclopedia of Philosophy* (first published 2001; substantive revision 20 September 2025). Online at https://plato.stanford.edu/entries/qm-bohm/

Hektoen International (2021). Dycus, K., “Peter Panum and the Geography of Disease”. *Hektoen International*, 18 August 2021. Online at https://hekint.org/2021/08/18/peter-panum-and-the-geography-of-disease/. Accessed 10 November 2026.

Hensen, B., and eighteen others (2015). “Loophole-free Bell Inequality Violation Using Electron Spins Separated by 1.3 Kilometres”. *Nature*, 526(7575), pp. 682 to 686. DOI https://doi.org/10.1038/nature15759

HISTORY.com Editors. “Montgomery Bus Boycott”. HISTORY.com. Online at https://www.history.com/articles/montgomery-bus-boycott. Accessed 10 November 2026.

Holtz, D., Carterette, B., Chandar, P., Nazari, Z., Cramer, H., and Aral, S. (2020). “The Engagement-Diversity Connection: Evidence from a Field Experiment on Spotify”. In *Proceedings of the 21st ACM Conference on Economics and Computation*. New York: Association for Computing Machinery. DOI https://doi.org/10.1145/3391403.3399532

Hossenfelder, S., and Palmer, T. (2020). “Rethinking Superdeterminism”. *Frontiers in Physics*, 8, article 139. DOI https://doi.org/10.3389/fphy.2020.00139

Howard, M. (2020). “How Bright Is the Future of Automation?”. *Automation.com Monthly*, 26 February 2020. Online at https://www.automation.com/article/how-bright-is-the-future-of-automation. Accessed 10 November 2026.

Hudson, N., and Moscoso Boedo, H. (2026). “The Collapse of Teen Fertility in the Digital Era”. Working paper, Department of Economics, University of Cincinnati, 25 April 2026. Online at https://homepages.uc.edu/~moscoshn/Personal_webpage/papers/Smartphone_web.pdf. Accessed 10 November 2026.

Huffaker, C. B. (1958). “Experimental Studies on Predation: Dispersion Factors and Predator-Prey Oscillations”. *Hilgardia*, 27(14), pp. 343 to 383. DOI https://doi.org/10.3733/hilg.v27n14p343

Ialenti, V. (2018). “Waste Makes Haste: How a Campaign to Speed Up Nuclear Waste Shipments Shut Down the WIPP Long-Term Repository”. *Bulletin of the Atomic Scientists*, 28 June 2018. Online at https://thebulletin.org/2018/06/waste-makes-haste-how-a-campaign-to-speed-up-nuclear-waste-shipments-shut-down-the-wipp-long-term-repository/. Accessed 10 November 2026.

Kanter, R. M. (1972). *Commitment and Community: Communes and Utopias in Sociological Perspective*. Cambridge, Massachusetts: Harvard University Press.

Kasmir, S. (2016). “The Mondragon Cooperatives: Successes and Challenges”. *Global Dialogue* (International Sociological Association), 13 February 2016. Online at https://globaldialogue.isa-sociology.org/the-mondragon-cooperatives-successes-and-challenges. Accessed 10 November 2026.

Kato, Y. (2008). “The Role of the Gishikichō in the Mid-Seventeenth-Century Reconstruction of the Buildings of Ise Shrine” (in Japanese; title translated). *Journal of Architecture and Planning* (Transactions of the Architectural Institute of Japan), 73(624), pp. 421 to 426. Online at https://www.jstage.jst.go.jp/article/aija/73/624/73_624_421/_article/-char/ja. Accessed 10 November 2026.

Koch, A., Brierley, C., Maslin, M. M., and Lewis, S. L. (2019). “Earth System Impacts of the European Arrival and Great Dying in the Americas after 1492”. *Quaternary Science Reviews*, 207, pp. 13 to 36. DOI https://doi.org/10.1016/j.quascirev.2018.12.004

Kohn, W. (1999). “Nobel Lecture: Electronic Structure of Matter; Wave Functions and Density Functionals”. *Reviews of Modern Physics*, 71(5), pp. 1253 to 1266. DOI https://doi.org/10.1103/RevModPhys.71.1253

Kolmogorov, A. N. (1965). “Three Approaches to the Quantitative Definition of Information”. *Problems of Information Transmission*, 1(1), pp. 1 to 7.

Landrigan, L. (2017). “The First Town Meeting Still Going in Six States”. New England Historical Society, 25 February 2017; revised 25 February 2024. Online at https://newenglandhistoricalsociety.com/oldest-town-meeting-6-states/. Accessed 10 November 2026.

Landry, Forrest. *A Paradox of Presentation; Why the Field of AI Safety Does Not Hear the Impossibility Argument, and How It Can Be Presented*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-427-3. DOI https://doi.org/10.67291/HHCA3000

Landry, Forrest. *Accessing the Real; Applications of the Immanent Metaphysics*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-125-8. DOI https://doi.org/10.67291/NIEI3296

Landry, Forrest. *Adverse Conditions; Why We Must Think Together Beyond the Scale of Our Darkness*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-220-0. DOI https://doi.org/10.67291/YUGQ2685

Landry, Forrest. *An Airplane or a Perpetual-Motion Machine; Why Safe AGI Can Never Be Built*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-219-4. DOI https://doi.org/10.67291/SOWV6394

Landry, Forrest. *An Immanent Metaphysics*. First published 2001. Wytheville, Virginia: UVSM Publications. ISBN 979-8-9896628-4-5. Cited by aphorism number.

Landry, Forrest. *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

Landry, Forrest. *Care, Transaction, and Power; The Three Bases of Relation*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-145-6. DOI https://doi.org/10.67291/VGQD1074

Landry, Forrest. *Choosing Together; Law, Consensus, Meritocracy, and Democracy*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-355-9. DOI https://doi.org/10.67291/MIPQ1939

Landry, Forrest. *Community Design and Practice; On the Ephemeral Group Process*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-63-2. DOI https://doi.org/10.67291/PCZE8733

Landry, Forrest. *Comparison; The Necessary Basis*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-062-6. Chapter 9, Sections V and VI (the six intrinsics of interaction; perceiver, perception and perceived).

Landry, Forrest. *Conditions of Continued Existence; Closed Loops, Cascading Failure, and the Limits of Control*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-241-5. DOI https://doi.org/10.67291/BAEE5779

Landry, Forrest. *Core Theory and Design; The Architecture of the Ephemeral Group Process*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-76-2. DOI https://doi.org/10.67291/VCEV4856

Landry, Forrest. *Engagements with Machine Madness; A Dispelling of False Belief That AI Risk Can Be Managed*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-426-6. DOI https://doi.org/10.67291/PQOA9164

Landry, Forrest. *Generator Functions of Existential Risk; Why the Same Failures Keep Returning, and What Would Change That*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-231-6. DOI https://doi.org/10.67291/HHDI1076

Landry, Forrest. *Governance Beyond the Dunbar Number; The Uncanny Valley of Scale, and the Category of Solution Required*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-385-6. DOI https://doi.org/10.67291/NZMI5837

Landry, Forrest. *Incommensuration; The Theorem and the Three Axioms*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-063-3. Chapter 16, Section IV (the impossibility of complete determinism).

Landry, Forrest. *Judgment and Discernment; Opposite Structures in Ethics, Relationship, and Governance*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-156-2. DOI https://doi.org/10.67291/FMBS5787

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. *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. *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

Landry, Forrest. *Not Our Needs; Why Machinery Cannot Be Made Safe for Living Things, Told Plainly*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-217-0. DOI https://doi.org/10.67291/BXOY2351

Landry, Forrest. *Power Without Care; Multi-Polar Traps, the Attention Economy, and the Structural Impossibility of Ethical Capitalism*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-242-2. DOI https://doi.org/10.67291/GINI2828

Landry, Forrest. *Redeeming Technology; Essays on the Ethics of Making and the Political Economy of the Interface*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-224-8. DOI https://doi.org/10.67291/NNAM7065

Landry, Forrest. *Sensemaking Under Irreversible Stakes; Method, Venue, Foundation, and the Two Biases*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-234-7. DOI https://doi.org/10.67291/IVPK8559

Landry, Forrest. *Social Media and Information Warfare; The Sterile Encounter and the Hours of Our Understanding*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-235-4. DOI https://doi.org/10.67291/EOGZ9555

Landry, Forrest. *Software, Wildness, and the Normal Form; From the Architecture of Code to Life, Justice, and Wisdom*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-237-8. DOI https://doi.org/10.67291/AWME8110

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

Landry, Forrest. *Technology at the Right Dosage; Tools That Fit the Body and Belong to the Community*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-240-8. DOI https://doi.org/10.67291/ASZX1334

Landry, Forrest. *That Which Enables Choice; A Book on the Nature of Love*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-182-1. DOI https://doi.org/10.67291/KXCY6710

Landry, Forrest. *The Anti-Patriot Institutions; Why Platforms That Sell Our Trust Owe Us a Fiduciary's Duty*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-951203-01-4. DOI https://doi.org/10.67291/IEGM7768

Landry, Forrest. *The Architecture of Civilization; Choice, Coherence, and Governance from the Neighborhood to the Stars*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-372-6. DOI https://doi.org/10.67291/VUJQ8853

Landry, Forrest. *The Architecture of Collective Choice; Governance, Sense-Making, and the Design of Community Process*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-352-8. DOI https://doi.org/10.67291/JUDR6441

Landry, Forrest. *The Architecture of Trust; Governance Adequate to Existential Risk*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-377-1. DOI https://doi.org/10.67291/BFTA4937

Landry, Forrest. *The Argument, Line by Line; The Impossibility of AGI Alignment, Read Aloud and Pressed Step by Step*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-218-7. DOI https://doi.org/10.67291/LNWQ4409

Landry, Forrest. *The Capture of Choice; How Choice Is Displaced, What That Displacement Costs, and What Would Be Needed to Reverse It*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-184-5. DOI https://doi.org/10.67291/LOKE8879

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

Landry, Forrest. *The Closed Gate; Why the Rich World Stopped Having Children*. Wytheville, Virginia: UVSM Publications, 2026.

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

Landry, Forrest. *The Dynamics of Small Groups*. Essay. Wytheville, Virginia: UVSM Publications.

Landry, Forrest. *The Foundations of the Sacred; A Review of the Immanent Metaphysics*. Volume II, *The Interior: Emotion, Motivation, and Choice*, ISBN 979-8-90710-163-0, DOI https://doi.org/10.67291/IBRF3472; Volume III, *Ethics, Creation, and the Sacred*, ISBN 979-8-90710-164-7, DOI https://doi.org/10.67291/CHQO5921. Wytheville, Virginia: UVSM Publications, 2026.

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

Landry, Forrest. *The Ground of Objection; Where Prior Philosophy Meets the Immanent Metaphysics*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-9896628-8-3. DOI https://doi.org/10.67291/WMKX2715

Landry, Forrest. *The Handover and the Ledger; Blockchain, Computation, and the Limits of Distributed Trust*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-223-1. DOI https://doi.org/10.67291/VJRC4176

Landry, Forrest. *The Hardest Problem of Civilization; Thirty Essays on the Design of a Coherent World*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-356-6. DOI https://doi.org/10.67291/TWYG6234

Landry, Forrest. *The Known, the Unknown, and the Unknowable; Essays on Epistemology, Coherence, and Structure*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-169-2. DOI https://doi.org/10.67291/JVAZ7395

Landry, Forrest. *The Multipolar Trap; Institutional Failure and the Architecture of Choice*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-248-4. DOI https://doi.org/10.67291/WTCV7999

Landry, Forrest. *The Questions of Ethics; The Seventy-Four Questions of Apex Section VI, Answered from the Immanent Metaphysics*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-134-0. DOI https://doi.org/10.67291/VAXS4281

Landry, Forrest. *The Substrate and the Ecosystem; Why Artificial General Intelligence Cannot Be Aligned, and What Remains for Persons*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-246-0. DOI https://doi.org/10.67291/NYZX3639

Landry, Forrest. *The Substrate Argument; Why the Alignment of Artificial Intelligence Is Not a Hard Problem but an Impossible One*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-238-5. DOI https://doi.org/10.67291/SXIU7050

Landry, Forrest. *The Transition; Civilization at the End of Its Toolkit*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-404-4. DOI https://doi.org/10.67291/ATLW5568

Landry, Forrest. *The Two Principles of Ethics; On the Symmetry and Continuity of Effective Choice, Their Derivation, Their Consequences, and Their Practice*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-190-6. DOI https://doi.org/10.67291/WOVX8093

Landry, Forrest. *The Valley Crossing; Technology, Wisdom, and the Design of Governance*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-364-1. DOI https://doi.org/10.67291/WOTP8305

Landry, Forrest. *Value, Purpose, and Unicity; Essays on the Ground of Worth and the Architecture of Choice*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-193-7. DOI https://doi.org/10.67291/BFZP7758

Landry, Forrest. *What Is Choice?; Ethics, Free Will, and the Path of Right Action*. Wytheville, Virginia: UVSM Publications, 2026. ISBN 979-8-90710-194-4. DOI https://doi.org/10.67291/SHWQ2953

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

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

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

Laughlin, G., Aguirre, A., and Grundfest, J. (2013). "Information Transmission between Financial Markets in Chicago and New York". *arXiv preprint* arXiv: 1302.5966.

Limpert, E., Stahel, W. A., and Abbt, M. (2001). "Log-normal Distributions across the Sciences: Keys and Clues". *BioScience*, 51(5), pp. 341 to 352. DOI https://doi.org/10.1641/0006-3568(2001)051[0341:LNDATS]2.0.CO;2

Loss, H. (2024). "Which Are Rarer: Diamonds or Emeralds?". *Live Science*, 21 October 2024.

Ludlow, E. (2026). "Waymo Co-CEO Outlines Path to 1 Million Weekly Trips in 2026". Bloomberg, reprinted in *Claims Journal*, 12 February 2026. Online at https://amp.claimsjournal.com/news/national/2026/02/12/335668.htm. Accessed 10 November 2026.

Macy, J. R. (1983). *Despair and Personal Power in the Nuclear Age*. Philadelphia: New Society Publishers.

Maendel, D., and Hofer, J. “Hutterite History Overview”. Hutterites.org. Online at https://hutterites.org/history/hutterite-history-overview/. Accessed 10 November 2026.

Martin Luther King, Jr. Research and Education Institute, Stanford University. “Browder v. Gayle, 352 U.S. 903 (1956)”. The Martin Luther King, Jr. Encyclopedia. Online at https://kinginstitute.stanford.edu/browder-v-gayle-352-us-903. Accessed 10 November 2026.

May, T. C., and Woods, M. H. (1979). “Alpha-Particle-Induced Soft Errors in Dynamic Memories”. *IEEE Transactions on Electron Devices*, 26(1), pp. 2 to 9. DOI https://doi.org/10.1109/T-ED.1979.19370

McElroy, M. (2019). “Mothers Against Drunk Driving (MADD) Begins”. Research Starters. EBSCO Information Services. Online at https://www.ebsco.com/research-starters/law/mothers-against-drunk-driving-madd-begins. Accessed 10 November 2026.

McNamara, R. (2026). “Advice, Not Control: The Role of Remote Assistance in Waymo’s Operations”. Waymo, 17 February 2026. Online at https://waymo.com/blog/shorts/advice-not-control-the-role-of-remote-assistance/. Accessed 10 November 2026.

Molina, M. J., and Rowland, F. S. (1974). “Stratospheric Sink for Chlorofluoromethanes: Chlorine Atom-Catalysed Destruction of Ozone”. *Nature*, 249(5460), pp. 810 to 812. DOI https://doi.org/10.1038/249810a0

Moore, D., Paxson, V., Savage, S., Shannon, C., Staniford, S., and Weaver, N. (2003). “Inside the Slammer Worm”. *IEEE Security and Privacy*, August 2003. Online at https://www.caida.org/catalog/papers/2003_sapphire/. Accessed 10 November 2026.

Moore, P. (2025). “Caterpillar Sets Out to Hit Over 2,000 Autonomous Mining Trucks by 2030”. *International Mining*, 7 November 2025. Online at https://im-mining.com/2025/11/07/caterpillar-sets-out-to-hit-over-2000-autonomous-mining-trucks-by-2030/. Accessed 10 November 2026.

Morens, D. M. (1998). “Measles in Fiji, 1875: Thoughts on the History of Emerging Infectious Diseases”. *Pacific Health Dialog*, 5(1).

Myers, C. K., and Hooper, E. (2026). “Is the iPhone Birth Control? Causal Evidence from AT&T’s 2007 to 2011 Carrier Monopoly”. NBER Working Paper 35310. Cambridge, Massachusetts: National Bureau of Economic Research (issued June 2026, revised October 2026). Online at https://www.nber.org/papers/w35310. Accessed 10 November 2026.

Nakanishi, M. “Keikōin Shūyō”. *Encyclopedia of Shinto*. Tokyo: Kokugakuin University. Online at https://d-museum.kokugakuin.ac.jp/eos/detail/?id=9526. Accessed 10 November 2026.

NASA Jet Propulsion Laboratory. “Ozone Milestones”. Online at https://atmos.jpl.nasa.gov/milestones.htm. Accessed 10 November 2026.

NASA, Earth Fact Sheet. NASA Space Science Data Coordinated Archive. “Earth Fact Sheet”. Online at https://nssdc.gsfc.nasa.gov/planetary/factsheet/earthfact.html. Accessed 10 November 2026.

National Academies of Sciences, Engineering, and Medicine (2015). “On Human Gene Editing: International Summit Statement”. Statement by the Organizing Committee of the International Summit on Human Gene Editing, December 2015. Online at https://www.nationalacademies.org/news/2015/12/on-human-gene-editing-international-summit-statement. Accessed 10 November 2026.

National Audubon Society. “Christmas Bird Count”. Online at https://www.audubon.org/community-science/christmas-bird-count. Accessed 10 November 2026.

National Institutes of Health (2017). “Notice Announcing the Removal of the Funding Pause for Gain-of-Function Research Projects”. Notice NOT-OD-17-071, 19 December 2017. Online at https://grants.nih.gov/grants/guide/notice-files/NOT-OD-17-071.html. Accessed 10 November 2026.

National Library of Medicine. “Recombinant DNA Technologies and Researchers’ Responsibilities, 1973 to 1980”. Profiles in Science. Online at https://profiles.nlm.nih.gov/spotlight/cd/feature/dna. Accessed 10 November 2026.

National Museum of Australia. “Rabbits Introduced”. Defining Moments. Online at https://nma.gov.au/defining-moments/resources/rabbits-introduced. Accessed 10 November 2026.

National Park Service (2025). Thomas, K., “Haleakalā National Park: Resisting Climate-Enhanced Threats of Avian Malaria”. NPS Climate Change Response Program, last updated 17 January 2025. Online at https://www.nps.gov/articles/000/hale-rad.htm. Accessed 10 November 2026.

National Research Council (2006). *Preventing the Forward Contamination of Mars*. Washington, DC: The National Academies Press. DOI https://doi.org/10.17226/11381

Neufeld, K. (2019). "Sangha Can Be the Next Buddha". *Lion's Roar*, 10 December 2019.

Nixon, R. (1971). "Statement About the Conversion of Facilities at Fort Detrick, Maryland, to a Center for Cancer Research". 18 October 1971. The American Presidency Project, University of California, Santa Barbara. Online at https://www.presidency.ucsb.edu/documents/statement-about-conversion-facilities-fort-detrick-maryland-center-for-cancer-research. Accessed 10 November 2026.

Ogilvie, S. "Frequently Asked Questions about Guilds". Faculty of Economics, University of Cambridge. Online at https://econ.cam.ac.uk/node/865. Accessed 10 November 2026.

Omohundro, S. M. (2008). "The Basic AI Drives". In P. Wang, B. Goertzel and S. Franklin (Eds.), *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, Frontiers in Artificial Intelligence and Applications, vol. 171, pp. 483 to 492. Amsterdam: IOS Press.

Onondaga Nation. "Chiefs". Online at https://onondaganation.org/government/chiefs. Accessed 10 November 2026.

Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*. Cambridge: Cambridge University Press.

Partridge, E. A., Davey, M. G., Hornick, M. A., and fifteen others (2017). "An Extra-uterine System to Physiologically Support the Extreme Premature Lamb". *Nature Communications*, 8, article 15112. DOI https://doi.org/10.1038/ncomms15112

Pew Research Center (2020). "Parenting Children in the Age of Screens". 28 July 2020. Online at https://www.pewresearch.org/internet/2020/07/28/parenting-children-in-the-age-of-screens/. Accessed 10 November 2026.

Pew Research Center (2021). "What Makes Life Meaningful? Views From 17 Advanced Economies". 18 November 2021. Online at https://www.pewresearch.org/global/2021/11/18/what-makes-life-meaningful-views-from-17-advanced-economies/. Accessed 10 November 2026.

Pew Research Center (2023). “AI in Hiring and Evaluating Workers: What Americans Think”. 20 April 2023. Online at https://www.pewresearch.org/internet/2023/04/20/ai-in-hiring-and-evaluating-workers-what-americans-think/. Accessed 10 November 2026.

Pew Research Center (2025). “U.S. Workers Are More Worried Than Hopeful About Future AI Use in the Workplace”. 25 February 2025. Online at https://www.pewresearch.org/social-trends/2025/02/25/u-s-workers-are-more-worried-than-hopeful-about-future-ai-use-in-the-workplace/. Accessed 10 November 2026.

Raja, S. N., Carr, D. B., Cohen, M., and thirteen others (2020). “The Revised International Association for the Study of Pain Definition of Pain: Concepts, Challenges, and Compromises”. *Pain*, 161(9), pp. 1976 to 1982. DOI https://doi.org/10.1097/j.pain.0000000000001939

Ramachandran, V. S., and Rogers-Ramachandran, D. (1996). “Synaesthesia in Phantom Limbs Induced with Mirrors”. *Proceedings of the Royal Society B: Biological Sciences*, 263(1369), pp. 377 to 386. DOI https://doi.org/10.1098/rspb.1996.0058

Renström, E. A., and Bäck, H. (2024). “Manfluencers and Young Men’s Misogynistic Attitudes: The Role of Perceived Threats to Men’s Status”. *Sex Roles*, 90, pp. 1787 to 1806.

Royal Society. “History of the Royal Society”. Online at https://royalsociety.org/about-us/who-we-are/history/. Accessed 10 November 2026.

Rubin, Z., Willis, D., and Ludwig, M. (2019). “Measuring Success in Intentional Communities: A Critical Evaluation of Commitment and Longevity Theories”. *Sociological Spectrum*. Online at https://www.dancingrabbit.org/wp-content/uploads/2020/11/Measuring-Success-in-Intentional-Communities-A-Critical-Evaluation-of-Commitment-and-Longevity-Theories.pdf. Accessed 10 November 2026.

Samurai Archives. “Ise Shrine”. Samurai Archives Wiki. Online at https://samurai-archives.com/wiki/Ise_shrine. Accessed 10 November 2026.

Sandberg, A., Drexler, E., and Ord, T. (2018). “Dissolving the Fermi Paradox”. *arXiv preprint* arXiv:1806.02404.

Sandia National Laboratories (2018). “Sean Dunagan Calls Events That Led to WIPP’s Three-Year Closure a ‘Cautionary Tale’”. *Sandia Lab News*, 16 August 2018. Online at https://www.sandia.gov/labnews/2018/08/16/wipp/. Accessed 10 November 2026.

Schultz, R. G. “The Kibbutz Movement”. My Jewish Learning. Online at https://www.myjewishlearning.com/article/the-kibbutz-movement/. Accessed 10 November 2026.

Sender, R., Fuchs, S., and Milo, R. (2016). “Revised Estimates for the Number of Human and Bacteria Cells in the Body”. *PLoS Biology*, 14(8), article e1002533. DOI https://doi.org/10.1371/journal.pbio.1002533

Silver, D., Huang, A., and eighteen others (2016). “Mastering the Game of Go with Deep Neural Networks and Tree Search”. *Nature*, 529(7587), pp. 484 to 489. DOI https://doi.org/10.1038/nature16961

Silver, D., Schrittwieser, J., Simonyan, K., and fourteen others (2017). “Mastering the Game of Go without Human Knowledge”. *Nature*, 550(7676), pp. 354 to 359. DOI https://doi.org/10.1038/nature24270

Sloan, D. (2024). “This $362 Billion ‘Beyond Well-Positioned’ Dutch Company Is Quietly Winning the Global AI Chips Race”. *Fortune*, 18 April 2024. Online at https://fortune.com/2024/04/18/asml-semiconductor-ai-manufacturing-uev-lithography-chips-act-nvidia-tsmc-wafers/. Accessed 10 November 2026.

Smithsonian Magazine (2021). “Leaded Gasoline Use in Vehicles Has Now Officially Ended Worldwide”. Smart News. Online at https://www.smithsonianmag.com/smart-news/worldwide-use-leaded-gasoline-vehicles-now-completely-phased-out-180978549/. Accessed 10 November 2026.

Social Welfare History Project. “Oneida Community (1848-1880): A Utopian Community”. Virginia Commonwealth University Libraries. Online at https://socialwelfare.library.vcu.edu/religious/the-oneida-community-1848-1880-a-utopian-community/. Accessed 10 November 2026.

Spada, P., and Vreeland, J. R. (2013). “Who Moderates the Moderators? The Effect of Non-neutral Moderators in Deliberative Decision Making”. *Journal of Public Deliberation* (now *Journal of Deliberative Democracy*), 9(2). DOI https://doi.org/10.16997/jdd.165

Spence, M. (1973). “Job Market Signaling”. *Quarterly Journal of Economics*, 87(3), pp. 355 to 374.

Stivale, C. J. (1997). "'help manners': Cyber-Democracy and Its Vicissitudes". *Enculturation*, 1(1). Online at https://enculturation.net/1_1/stivale.html. Accessed 10 November 2026.

Strasburger, K. (2019). "Explicitly Correlated Wavefunctions of the Ground State and the Lowest Quintuplet State of the Carbon Atom". *arXiv preprint* arXiv: 1903.06051.

Sullivan, M. J. L., Bishop, S. R., and Pivik, J. (1995). "The Pain Catastrophizing Scale: Development and Validation". *Psychological Assessment*, 7(4), pp. 524 to 532. DOI https://doi.org/10.1037/1040-3590.7.4.524

Terra Universal. "FS209E and ISO Cleanroom Standards". Online at https://www.terrauniversal.com/blog/fs209e-and-iso-cleanroom-standards/. Accessed 10 November 2026. (Particle limits of ISO 14644-1 by class, and recommended air changes per hour by class.)

Timings, J. (2022). "Busting ASML Myths". ASML, 23 February 2022. Online at https://www.asml.com/news/stories/2022/busting-asml-myths. Accessed 10 November 2026.

Tschudi, W., Benschine, K., Fok, S., and Rumsey, P. (2001). "Cleanroom Energy Benchmarking in High-Tech and Biotech Industries". In Proceedings of the 2001 ACEEE Summer Study on Energy Efficiency in Industry. Washington, DC: American Council for an Energy-Efficient Economy. Online at https://www.aceee.org/files/proceedings/2001/data/papers/SS01_Panel1_Paper19.pdf. Accessed 10 November 2026.

Turing, A. M. (1936). "On Computable Numbers, with an Application to the Entscheidungsproblem". *Proceedings of the London Mathematical Society*, series 2, 42(1), pp. 230 to 265. DOI https://doi.org/10.1112/plms/s2-42.1.230

Turing, A. M. (1950). "Computing Machinery and Intelligence". *Mind*, 59(236), pp. 433 to 460. DOI https://doi.org/10.1093/mind/LIX.236.433

Ulaby, N. (2025). "Most Teens Have Used AI to Flirt and Chat, but Still Prefer Human Interaction". *NPR*, 20 July 2025. Online at https://www.apr.org/arts-life/2025-07-20/most-teens-have-used-ai-to-flirt-and-chat-but-still-prefer-human-interaction. Accessed 10 November 2026.

Ullathorne, G. "How Could We Do Without Sugar and Rum?". *History in Focus*, Issue 12: Slavery. London: Institute of Historical Research. Online at https://archives.history.ac.uk/history-in-focus/Slavery/articles/ullathorne.html. Accessed 10 November 2026.

Ungvarsky, J. (2021). “Shakers (sect)”. EBSCO Research Starters. Online at https://www.ebsco.com/research-starters/religion-and-philosophy/shakers-sect. Accessed 10 November 2026.

United Nations Office for Disarmament Affairs. “Biological Weapons Convention: History”. Online at https://disarmament.unoda.org/biological-weapons/about/history. Accessed 10 November 2026.

United Nations Office for Disarmament Affairs. “Convention on the Prohibition of Military or Any Other Hostile Use of Environmental Modification Techniques (ENMOD)”. Online at https://archive-disarmament.unoda.org/enmod. Accessed 10 November 2026.

United States Patent and Trademark Office. “608.03 Models, Exhibits, Specimens” (revision R-08.2012). In *Manual of Patent Examining Procedure*, Chapter 600. Online at https://www.uspto.gov/web/offices/pac/mpep/s608.html. Accessed 10 November 2026.

U.S. Census Bureau (2026). “World Population Day: July 11, 2026”. Online at https://www.census.gov/newsroom/stories/world-population-day.html. Accessed 10 November 2026.

US Department of Agriculture, Animal and Plant Health Inspection Service (2017). *New World Screwworm Ready Reference Guide: Sterile Insect Response*. January 2017. Online at https://www.aphis.usda.gov/animal_health/emergency_management/downloads/nws_rrg_sterileinsectresponse.pdf. Accessed 10 November 2026.

U.S. Department of Energy. “WIPP Recovery and Restart”. Waste Isolation Pilot Plant. Online at https://www.wipp.energy.gov/recoveryrestart.asp. Accessed 10 November 2026.

U.S.-Canada Power System Outage Task Force (2004). *Final Report on the August 14, 2003 Blackout in the United States and Canada: Causes and Recommendations*. April 2004. Online at https://www.energy.gov/sites/prod/files/oeprod/DocumentsandMedia/BlackoutFinal-Web.pdf. Accessed 10 November 2026.

Vahdat, A. (2022). “Jupiter Evolving: Reflecting on Google’s Data Center Network Transformation”. *Google Cloud Blog*, 23 August 2022. Online at https://cloud.google.com/blog/topics/systems/the-evolution-of-googles-jupiter-data-center-network/. Accessed 10 November 2026.

Vlaeyen, J. W. S., and Linton, S. J. (2000). “Fear-Avoidance and Its Consequences in Chronic Musculoskeletal Pain: A State of the Art”. *Pain*, 85(3), pp. 317 to 332. DOI https://doi.org/10.1016/S0304-3959(99)00242-0

von Neumann, J. (1956). “Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components”. In C. E. Shannon and J. McCarthy (Eds.), *Automata Studies*, pp. 43 to 98. Princeton: Princeton University Press.

Wachowski, L., and Wachowski, L. (1999). *The Matrix*. Film. Burbank, California: Warner Bros.

Wallner, C. M., and Kitts, J. A. “Hutterites”. In D. J. Wishart (Ed.), *Encyclopedia of the Great Plains*. Lincoln: University of Nebraska-Lincoln. Online at https://plainshumanities.unl.edu/encyclopedia/doc/egp.rel.026.html. Accessed 10 November 2026.

Weinbren, D. (2014). “The World in 1913: Friendly Societies”. *The Historian*, Historical Association, 6 February 2014. Online at https://www.history.org.uk/student/categories/107/resource/7082. Accessed 10 November 2026.

Wiener, N. (1948). *Cybernetics: Or Control and Communication in the Animal and the Machine*. Cambridge, Massachusetts: The Technology Press; New York: John Wiley and Sons; Paris: Hermann.

Work That Reconnects Network. “From Despair to Empowerment”. Online at https://workthatreconnects.org/?p=10523. Accessed 10 November 2026.

World Nuclear News (2024). “Trial Run at Finnish Repository Starts”. 30 August 2024. Online at https://world-nuclear-news.org/articles/trial-run-at-finnish-repository-starts. Accessed 10 November 2026.

WWF (2024). *Living Planet Report 2024; A System in Peril*. Gland, Switzerland: WWF.

Xu, T. (2002). “Airflow Design for Cleanrooms and Its Economic Implications”. Berkeley, California: Lawrence Berkeley National Laboratory. Online at https://www.osti.gov/servlets/purl/832942. Accessed 10 November 2026.

Young Center for Anabaptist and Pietist Studies, Elizabethtown College. “Amish Population Profile 2026”. Amish Studies. Online at https://groups.etown.edu/amishstudies/?p=2517. Accessed 10 November 2026.

Young Center for Anabaptist and Pietist Studies, Elizabethtown College. “Church Discipline”. Amish Studies. Online at https://groups.etown.edu/amishstudies/?p=78. Accessed 10 November 2026.

Young Center for Anabaptist and Pietist Studies, Elizabethtown College. "Technology". Amish Studies. Online at https://groups.etown.edu/amishstudies/cultural-practices/technology/. Accessed 10 November 2026.

Zheng, J., and Meister, M. (2025). "The Unbearable Slowness of Being: Why Do We Live at 10 Bits/s?". *Neuron*, 113(2), pp. 192 to 204. DOI https://doi.org/10.1016/j.neuron.2024.11.008

Zhou, W.-X., Sornette, D., Hill, R. A., and Dunbar, R. I. M. (2005). "Discrete Hierarchical Organization of Social Group Sizes". *Proceedings of the Royal Society B: Biological Sciences*, 272(1561), pp. 439 to 444. DOI https://doi.org/10.1098/rspb.2004.2970

Index

abstraction differential, 117

accretion, 277

afraid of losing, what am I, 248

AI, pain and joy in, 86

air changes, in clean rooms, 148

Alanya, fishers of, 186

Alexander, Scott, "Meditations on Moloch", 179; 182

aliens, why not invite the, 47

alignment, vi; 2; 15; 19; 21; 41; 56; 64; 131; 173

alignment among people, possible and not established, 173

alignment, two senses of the word, 173; 174

Allcott and colleagues, Facebook deactivation experiment, 215

allure (**of category three topics**), 74

alp, shared pasture at Törbel, 255

alpha particles and memory chips, 35

AlphaChip, 109; 114; 122

AlphaGo, 11

Americas after 1492, epidemics in the, 159

Amish, 254; 256; 257; 258; 268

analytically consistent, 30

***Ape Trap*, in Part V**, 108; 115; 116; 121; 127; 144; 158; 208; 211; 213; 215; 223; 225; 228; 230; 232; 233; 242; 295; 297; 326

Apollo lunar quarantine, 154

architecture line (**closed loop, open ended, autonomous**), 108; 120; 317; 319

argument in one sentence, v; 326

arithmetic of many pathogens at once, 159

arms control, shared risk and balance in, 302

arms race framing, 321

arrow of time, 28; 36
art and sexuality, 76
artificial companion, 228
artificial general intelligence (AGI), 3; 66; 107; 119
artificial womb (the lamb study), 230
Asch experiments, 80
Ashby, W. Ross, law of requisite variety, 125
Asilomar conference (1975), 300
Asimov's laws, 52
Asse II mine, 155
assemblies, citizens', capture through organizers, 205
asymmetry of pain, 83; 101; 250
ATP (pinch point analogy), 67
AT&T and the iPhone (natural experiment), 225
attunement, 91
aviation fuel, leaded, 175
balance point (winds on a globe), 60
bandwidth limitation, 63; 93
baptism, voluntary adult, 257
barn raising, 278; 324
barrier, as a maintained thing, 149
barrier, who would pay for it, 150
base reality, 30
Beecher, Henry, wounded soldiers, 239
begging the question, 12
bell curve and long tail, 185
Bell's theorem, 35
Benedict, Rule of Saint, 257
Bennett, Charles, 165

Biological Weapons Convention, 301
biology constraint, 4; 19
Bird Count, Christmas, 256
birth rates, and the arrival of the smartphone, 225
birth rates, sorted by deployment of technology, 223; 225
blinding laser weapons, ban on, 301; 305
boat metaphor, 14; 64; 137
body as epistemological instrument, 95; 103; 230; 232; 285
body, a whole composed like a, 279
body, knowing in a group, 96
body, meaning known through the, 232
Bohm, David, dialogue groups, 103
boot-loader argument, 47
bootloader, society as, 115
boundary between environments, 6; 148
bridge, prediction of a, 59
broadcast, 70
builder, what the proof teaches a, 317
builders and the level, 199
Busy Beaver number, 53
camera on a wall (every controller is inside), 124
cancer (a cell that grows for itself), 172
cap on growth, no way out, 310; 317
capturable node, 205
capture, cost of, 208
capture-cost parity, 208
carbon atom, cost of simulating, 26
care (as enabling, not compelling), 78; 88; 101; 230; 249
care and pressure, a friend who has lost her work, 78

care, cost that every trap takes, 179; 190
care, kept only outside competition, 181
care, no score kept in, 189
care, not implied by pain and joy, 87
catalytic converter, 175
catastrophizing, 240
category error, 88
category three topics, 74; 246
ceiling and floor (of control), 133; 135; 137
***Ceiling Below the Floor, The*, in Part V**, 119; 124; 129; 133; 135; 137; 139; 144; 295; 298; 306; 311; 316; 317; 326
cells of one body (one metabolism, one supply, one fate), 172
Centaur (model of human choices), 211
certificate, carried by paid positions, 203
chain of production, a chip's, 114
check on the world, what stands in front of it, 199; 217
checklist, aviation (Model 299), 191
chess, predicting a player, 10; 133
children, a job, and whether a life means something, 221; 232
chip fabrication, few hands in, 296; 306
chip factory, choosing a site for, 3
chlorofluorocarbons, 156; 157; 176
choice that binds every builder at once, 295
choice, basis of, vi; 4; 77
choice, the one thing left to, 310; 326
chronic pain, 240
clamp, on compounding gain, 185
Clarkson, Thomas, 304
class of three conditions, 108; 120

clean room, as a wall with a price, 148; 149
closures, the two, 66; 123; 144; 163; 170; 233; 241; 310; 315; 326
cloth metaphor, 90
cloth, in a life, 91
coherence, vi; 71; 100
coherent influence, conditions for, 69; 98
cohort of about twelve, 281
colleague relationship, 12
commitment mechanisms (Kanter), 253
commitment, in a plan's author, 247
committee of committees, 200; 277
committee of twelve and groups of six, cost of capture, 208
common knowledge, 299
community of goods, 257; 260; 267
community (two to three hundred people), 273; 274
companion and feed, hours taken by, 228
competitor, everyone in an economic system is a, 180
completeness of knowledge of operation, 319
composition (of small groups), 276; 277; 279
computation, limits of, 26; 36; 43
conceptualization (of pain), 237
conditions for coherence, further, 289
conditions for coherence, the four, 100; 250; 258; 267; 268; 286
conduct, change of, seen first, 296
confiding, the hours of, 228
consent, in a virtual reality, 50; 51
containment, record of, 156
content, judging a message on its, 200
contingency as prior, 39; 43; 84

continuity ethics, 321
continuity of relationship, 254
control system (five components), 62; 127
control, impossibility of adequate, 67; 144
control, six steps of, 63; 127
convergence, why the word, 8
coordination, last restraint on a trap, 183
cost of staying, 208
council of ten, 200
count of code, from a cell to the Earth, 168
coupled channels, floor rising with, 137
created by and enabled by, 28; 36
crossing, the, 119; 295; 310; 317
cyanide (acting at a pinch point), 67
cybernetic constraints, 63
cyborg, human-machine, 52
data center, bandwidth of one, 133
decider, harm after the death of the, 187
de-personally, 78
definition, consistency of, 30
design statement, 312
deterministic compute substrate, 50
die, program and atom, 35
diethylstilbestrol, 157
dinner table (group of five), 270
displacement of relationship, first, 223
displacement, tightening on itself, 232
distance, as a barrier, 152
domesticated (orientation), 79

domesticated to wild, why not directly, 81
domesticated, by a feed, 211
driver on a winding road (six steps of control), 127
driverless cars, 109; 112; 122
drunk driving, the turn against, 304
Dunbar number, 272
economic decoupling, 176
economy, pressure at a community's boundary with, 265
ecosystem dynamics, 7; 15; 21
efficiency and fragility, 137
Embedded Observer Lemma, 125
embodied complexity, 165
emerald metaphor, 89
enabled by and created by, see *created by and enabled by*
energetic entanglement, 69
energy extraction, automation of, 121; 122
entanglement (quantum), 32
environmental modification, convention against (ENMOD), 301
ephemeral group process, 282
erasing a person from every mind, 31
error correction, as control with three elements taken out, 127
evolution, drivers of, 76
excuse, the, and what removes it, 298
existential risk, 49
experience and information, 9
extensivity of choice, 86
exterior and interior universes, 29; 32
extremal material, 245
FANUC plant (robots building robots), 109; 114

Faroe Islands, measles in 1846, 159
father and daughter, one hour against seven, 223
feed, prediction of its readers, 211
feral (orientation), 80; 82; 244
feral cat, 82; 227
feral to wild, the way back, 215
fertility rate, general, 225
fertility rate, total, 223
Feynman diagrams, analogy with, 58
Fiji, measles in 1875, 159
film reel (order and time), 36
first person, of a group, 100; 283
first-mover disadvantage, 194; 196
flash crash of May 2010, 137
floor, of shared chemistry, 7
floors, the figures as, 169; 170
founder, what survives a, 264; 267
four conditions, and the eight conditions of The Multipolar Trap, 196
four marks (of a narrow system), 107; 117; 319
friendly societies, 264
funders, relieving them of the belief in profit, 203
fusion (crossing the valley), 275
gain-of-function research, funding pause on, 301
game and world, 11
game, where it begins and stops, 189
gardener (Alexander), 183
gate, where the price of a distant wall falls, 153
gatekeeper, 97; 184; 266; 268
gene flow and genetic drift, 76

germline editing, 301
gestalt consciousness, 100; 103; 286
Goodhart's law, 190
governance without system, 85
governor, the power to stop makes a, 205
gradient, of a score, 191
grains, talcum and sand (pictures of a number), 171
Great Oxidation Event, 161
group consciousness, v; 100; 102; 103; 233; 283; 291; 326
group methods, scale of, 306
group sizes, 271; 291
guilds, 255; 259; 266; 281
hand on a board, 95
harbor, boats fishing one bank, 294
hard randomness, 28; 34; 84; 142; 166
Haudenosaunee (the Iroquois Confederacy), 253
haul trucks, driverless, 114; 122
Hawaii, honeycreepers and avian malaria, 161
He Jiankui, 301
healing, facilitation of, 83; 84; 101
hidden state, 133
hidden variables, 35
high-frequency trading, 23; 113; 137
holding (a coupled system in a safe range), 151
hopeless and choice, the two halves of the sentence, 326
hopelessness, weaponization of, 310
hours together, what they are for, 233
house, three conversations about a, 75
Huffaker's mites, 160

hull integrity, see *boat metaphor*
human in the loop, measure for narrow systems, 319
Hutterites, 254; 257; 260; 267; 272
hydrofluorocarbons, 176
ice, two falls on the, 236
immanent, 94
improbability of this world, 170; 171
Incommensuration Theorem, 36; 131; 166; 321
increase in capacities, principle of, 56; 100
industry held near the needs of living things, 175; 176
information and experience, see *experience and information*
instance process and global process, 4
instrument under the user's control, what it teaches, 215
intelligence, insufficient without care, 98
interaction as fundamental, 41
intermediate causation, 77
intermediate causation, in front of a group's check, 199
intermediation of choice, 78
interpretations of quantum mechanics, 37
invasive species analogy, 15; 161
Ise, Grand Shrine of, 257; 260; 266
isomorphism, 31; 37
Italian blackout of September 2003, 137
Japan, fertility in, 223
joy, shared in a group, 284
Kanter, Rosabeth Moss, 253
keeping, the (of the refusal), 295; 306
kibbutzim, 264
Kigali Amendment, 176

knowing, from somewhere, 285
Kolmogorov complexity, 165
LambdaMOO, 205
language models, where they stand, 112
large language model, 9; 12
leaded gasoline, 175
leak, patching a, 15
least chance of harm, at every step, 142
least hazard, 131; 142; 310; 312
leaving the feed, 215
Lightner, Candy, 304
limit of resources and limit of structure, 139
limits carry upward, methods do not, 139
lines of code, as a measure, 167
litmus test and reference range, 192
logical depth, 165
long-termism, 48
looking away, from what cannot be borne, 245; 246; 247
loop, running and closed, 107; 113; 116
machine as our child, 48
Macy, Joanna, despair and empowerment work, 248
marketplace seller, reputation spent once, 208
markets, merging, 17
Marshaling the Misalignment, **in Part V**, 108; 116; 120; 131; 139; 144; 173; 183; 318; 319; 322; 325; 326
Master Theorem, 66
Matrix, The (film), 31
meaning, sources of (surveys), 221
meaningfulness and embodiment, 93

means and outcome, 85
measurement event, single, 38
measurement, a single, walked through, 38
meeting places, kept many and small, 213; 215
member who leaves, and the federation of communities, 267
merged and separate intelligences, 16; 284
messages, where they go (hub and circle), 282
meta-theory of interpretation, 37
Minesweeper Madness, **in Part V**, 131; 139; 142; 144; 311; 313; 326
mistrust and partnership, 227
moderators, effect on deliberation, 205
monasteries, dissolution of the English, 266
Mondragon cooperatives, 264
Montgomery bus boycott, 72; 280
Montreal Protocol, 156; 176
Moon, separation on the, 152
moratorium, 300; 303
mother's body, provision of the, 230
multipolar trap, v; 97; 294; 319; 326
nanotech and the power equation, 49
navigation service, prediction and habits, 211
neighbor, a sentence for a, 324
nesting illusion, 278
nesting of scales, 278
nociception, 239
non-computational elements, 28; 36
non-dualist, 84
non-falsifiable, 38
non-other-than, the, 77

norms, speed of spread, 304; 305
Northeast blackout of August 2003, 137
nuclear waste, walls for, 155
numbers on the risk (per act, per year, per century), 312
numbness, as a way of living, 249
oak and its carved copy, 165
obedience, as a measure of control, 129
Oneida Community, 264
one-way mirror (platform), 205
Onkalo repository, 155
OODA loop, 64; 127
operator and supplicant, 129
orbit, a base in, 152; 153
organic and artificial, line between, 4
orthogonality thesis, 2
Ostrom, Elinor, 252; 255
oubliette, 53
ownership and competition, 180
ozone layer, 156
P and NP, 139
pain and pleasure, as perceived change in potentiality, 84; 86; 236; 239; 284
pain and suffering, 86; 237
pain, asymmetry of, 83
pandemic, and the community, 278
parking brake, whether anyone releases it, 310
particle glider, 27
partnership, fall in, 227
Pascal's mugging, 48
patent office, perpetual motion and the, 299; 312

perception, sanction and declining to finance, 303
percolation point, 137
perfect prison, 51
persistent branching, 142
phantom limb, 239
phase parallax, 284
phase transition (change of kind), 194
pilot-wave theory, 35; 36; 37
pinch points, 28; 67
piper, he who pays the, 22; 203; 245
planetary protection, 155
point of no return, 119
pool metaphor, 69; 72; 98; 228
power law, 185
precursors, 306
predictability, limits of, 59; 66; 133
Premise 0 and Conclusion G, 310; 312; 326
present at a step, necessary to a step, 113
pressure that runs one way (apes and ants against a bulldozer), 7
principal-agent problem, 19; 129; 289
principle of identity, 42
process as primary, 41
projection (of pain), 237
proof, and common knowledge of it, 298; 299
protective adjustment, 245; 246
pseudo-randomness, 29; 34
psychopathic, sociopathic, 87
Q, the substance, 42
Quaker meeting for business, 103; 252; 254; 258

question, meeting the question a message carries, 200
rabbits in Australia, 15
race to build AGI, as a multipolar trap, vii; 99; 294
race to build AGI, as a valley crossing, 196
Ramachandran, V. S., mirror box, 239
rate equation, 319
rationalization, 245
rats, island of, 179; 183
recombinant DNA moratorium (1974), 300
recombination, 76
recommendation, field experiment on diversity of listening, 211
record, written, and a broken chain of practice, 260
recursion, practical limits on, 67
red button test, 321
reductionism, 27
referee and rulebook (symmetry and continuity), 131
refraining, acting, finding someone, 324
refuge, 160; 161
refusal, why people can hold it, 297; 298
regulator's measure, 108; 118
regulator's sign, for a set of systems, 118
rehumanizing and renaturalizing, 85
relational bandwidth, 228
relational value, 230
replacement cost, 167; 170
replicating process (and inert influx), 15
representation, 279
restraints on a trap, the four, 182
restrictions, the five, 118

Rice's theorem, 120
rival scores, 294; 302
rotation, as a clamp, 186
Royal Society, 255
rubber-stamping (**oversight in name**), 117
safe side of the line, 319
safety, as real thriving, 52
sailing, from books and on the water, 10
salary, in front of the check, 203
sangha, 103
scales of control, table of, 144
Schrödinger equation, 26
score, better, does not take care's place, 190
screwworm (**sterile insect technique**), 223
second condition, and the four things in front of it, 217
seed (**of a random number program**), 35
selection, acting on a substrate, 5; 23; 150
self-control, controllability as, 66
separation through time, 159
series, how its arguments join, 326
SETI, 47
sexuality and meaningfulness, 93
Shakers, 257; 266
shared environment, 6
shares of a product, no fixed (**field of wheat**), 227
sheer duration of life, 53
shops, counted by a guild, 259
shunning, 258
shunning, prohibitions held without, 303

signal that is the capacity itself, 208
simulating layer, 29
simulation boundary, 28
simulation hypothesis, 26
simulation, hole in, 29; 30; 31
singleton, vii; 16; 98; 183
skills of sense-making, the four, 213
Slammer worm, 137
slave trade, abolition of the British, 304
smallest group whose agreement would suffice, 306; 307
smallest working case, 280
smartphone, arrival of the, 225
social species, 87
soft takeoff, 118
somatic awareness, 93
song, text and code, 281
speech, information rate of, 17
spell checker (methods stay at their level), 139
Spence, signaling, 208
standard toolkit, v; 97
step to check first, 315
sterile male, 223; 228
Stoic school, on suffering, 237
stone, before anyone looked, 42
stop, the (of the building), 295; 306
store of questions, 275
structural psychopathy, 228
structure, not enough alone, 99
subject and object, changing places, 43

subject and object, from interaction, 42

subjective causation and objective transition, 78

subjectivity, first-person, 28; 36; 43

***Substrate and the Ecosystem, The*, in Part V**, 116; 124; 135; 144; 148; 156; 158; 161; 163; 172; 192; 203; 223; 228; 230; 233; 241; 261; 284; 286; 297; 302; 306; 319; 325; 326

***Substrate Needs and Evolution*, in Part V**, 108; 115; 116; 119; 122; 153; 160; 326

substrate needs convergence, v; 8; 170

substrate-independence (Bostrom), 28

suffering, 237; 238

super-determinism, rejection of, 35

superintelligence, 3; 59

superposition state, 32

surgeon and patient, 20; 129

sustainability, principle of, 56; 100

symmetry ethics, 321

tagging an assumption, 13

Taiwan, drought and chip factories, 60

teachers, arithmetic of, 259

ten bits a second (human throughput), 133

ten to the thirty-third (the figure in several places), 170; 171

test ban treaty (1963), 301

thermostat, the loop of control in a, 64; 127

Thich Nhat Hanh, 103

third market (reproductive and care labor), 230

three geometries of growth, 57; 97; 184

three transhumanisms, 48

thresholds, how set, 192

topology (of wildness and domestication), v; 79

Törbel, 255
town meeting, New England, 254
toxicity default, 158
training data and real-world experience, 9
traitorism, total, 47
transactionalism, 85
transcendent, 76
transhumanism, 46; 48
trauma, 241; 242; 243
trauma, healing of, 244; 249
trauma, in a superintelligent creature, 243
treasury dynamics, 97
truck on a grade, 310
turbine analogy, 70
Turing machine, 36
two principles of ethics, 321
uncanny valley (of group size), 273
unicity, 167; 171
uploads, 52
valley crossing, 194
value of the Earth, 170
value of the Earth, the odds it demands, 312
vanity, 91
venue, whoever keeps it, 205
Viking landers, sterilization of, 155
village water program, 270; 281
virtual reality, consent in, 50; 51
von Neumann, John, reliable computer from unreliable parts, 139
vouching for a proof, 300

vulnerability, shared, 19

wall between ecologies, cost of, 149; 151

wall between ecologies, why it cannot be kept, 162

war veteran and the backfire, 241

Waste Isolation Pilot Plant, 155

water table (threshold of change), 194

waterfall metaphor, 70; 98

"we made this", 284

weapon process (and its inverse), 77; 250; 258; 321

weapon, cases of a, 77

what must be held, and by what, 124

wheat, field of (two causes that multiply), 227

***Why AGI Cannot Be Made Safe*, in Part V**, 107; 113; 116; 119; 121; 144; 150; 153; 158; 163; 168; 170; 174; 176; 180; 183; 187; 194; 203; 221; 223; 232; 242; 245; 265; 272; 279; 286; 297; 302; 310; 313; 315; 319; 323; 324; 326

***Why No Control Suffices*, in Part V**, 125; 127; 133; 135; 139; 142; 144; 152; 161; 170; 172; 187; 279; 284; 311; 316; 326

wild (orientation), 79; 82; 101

wild, and stubbornness, 80

window for action, what to watch for, 121; 123

wireheading, 48; 228

wisdom, 85; 86

work and artificial intelligence (surveys), 221

workshop (clamp), 185

workshop of eleven, 270

workshop (sustainability and capacity), 57; 58; 135; 280

written record in four parts (the question, what matters, why it matters, the questions that remain), 200

yes-or-no list in place of a score, 191