

Architektonikē

Designing Realities for Conscious Beings

This edition holds the text as it stood on October 28, 2026. The work is still being written, and the current text is published at <https://architektonike.com>.

The text is available under CC BY-SA 4.0 (<https://creativecommons.org/licenses/by-sa/4.0/>). Credit it as “Architektonikē, published anonymously.”

Ah, Love! could thou and I with Fate conspire
To grasp this sorry Scheme of Things entire,
Would not we shatter it to bits—and then
Re-mould it nearer to the Heart's Desire!

Omar Khayyám

Rubáiyát, LXXIII

Translated by Edward FitzGerald, 1859

Contents

Volume One. Foundations and the Near Horizon	7
Introduction	9
<i>Part I. Before the Beginning</i>	19
1 What Is a Reality?	21
2 What Carries a World?	59
3 The Designer's Starting Position	97
<i>Part II. The Moral Foundations</i>	117
4 The Primacy of the Subject	119
5 Who Might Be Home	147
Commitments	179
Bibliography	195

VOLUME ONE

Foundations and the Near Horizon

Introduction

The First Decision

Before a designed reality exists, there is only a substrate and someone facing it. The substrate contains no space, since no metric has yet told any two of its states how far apart they are. Nothing in it changes, because no law has been written to govern change. Its memory holds values that mean nothing, and its processors cycle through operations that describe nothing.

Whoever sits before such a system occupies a position no engineer, architect, physician, or legislator in history has occupied. Each of them worked inside a world whose deepest rules were settled long before they were born. Builders of cathedrals took gravity as a given and shaped their vaults to stand against it. Painters received light from the sun, with its particular spectrum and its way of scattering in air, and made their art within its possibilities. Lawgivers wrote statutes for bodies that tire, bleed, age, and die, without ever having to choose any of those facts.

A designer at the empty substrate inherits none of these givens. Every quantity that earlier makers treated as the fixed frame of their work becomes a variable awaiting a value. The decisions that settle those values fall into an order, since each early choice narrows the space in which later choices can be made. So the designer has to begin somewhere, with a first decision that everything built afterward will inherit.

For nearly all of history this scene belonged to theology and to speculative fiction. Over the past few decades, several lines of re-

search have moved it toward engineering. Physicists using lattice quantum chromodynamics now derive the masses of the proton and neutron, and even the small difference between them, from the equations of the strong and electromagnetic interactions. Quantum chemists predict the structure and behavior of molecules and materials from the quantum mechanics of their electrons before anyone has made them in a laboratory. Machine learning has produced world models, networks trained on video and interaction that generate explorable environments frame by frame, keeping terrain, lighting, and objects coherent as a user moves through them. Researchers have begun publishing assessments of which artificial architectures satisfy the indicators proposed by leading scientific theories of consciousness, and at least one AI developer has launched a research program on the welfare of its own models. Each of these results remains partial. First-principles calculation grows ruinously expensive as the number of interacting particles increases. World models hold their coherence for minutes and obey a physics learned approximately from footage. Nobody can yet determine whether a given system feels anything.

What turns these fragments into a near prospect is artificial superintelligence, a capability exceeding human performance across science and engineering, which many of the people building today's systems expect within years. Such a system could design its own successors and the hardware they run on. It could also pursue research faster than any human institution, closing in years distances that human science would cross in centuries. A civilization equipped with it could plausibly command the computation and theory needed to sustain a coherent physics that bottoms out near the level of life, along with the minds living under it, before this century ends. The figures this book uses for cost and capability are deliberately conservative, and superintelligence is likely to find machines and methods that leave them far behind, though nothing it builds can escape the limits physics places on every machine. The forecast could fail in its timing or

its premises, and consciousness may resist inquiry even by minds far greater than ours. Still, the probability is high enough to make the design of realities a practical question for people alive today. Its ethics have to be settled before the capacity arrives, because whatever habits of thought exist when building becomes possible will govern what gets built. Those ethics turn on one question, what someone able to design a reality owes the beings who will come to live in it.

Our own universe arose without any designer anyone can identify, and every later reality would have to be built inside it. It can be called the primordial cosmos, since it stands at the head of whatever sequence of realities comes after it. The usual name for a world sustained by computation is “simulation,” a word that ordinarily means an imitation of something located elsewhere. Weather models, flight trainers, and economic forecasts represent their targets without instantiating them. The hurricane in a forecasting model floods no streets and drowns no one. Calling a designed reality a simulation imports that relation of copy to original, and with it a ranking in which the computational world becomes a lesser shadow of the primordial one. A designed reality need imitate nothing, and the people in it are no one’s shadows.

Consider what actually holds inside a reality whose physics runs on a computational substrate and whose inhabitants are conscious. Pain there hurts whoever suffers it. A parent grieves a lost child with a grief wholly their own, friendships console, and a climber’s hands bleed on rock as surely as any hands have bled on granite. Whether that rock consists of quarks or of states in a memory register makes no difference to the one who bleeds. When we learned that our own matter is made of excitations in quantum fields, not a single wound in our world hurt less. Reality, for anyone whose life is at stake, is a matter of what a world does to and for the subjects within it. A world that sustains conscious lives possesses all the reality those lives require, and describing it as an imitation invites its mistreatment. A maker who thinks of a world as a model will feel licensed to pause it, edit it,

run it for curiosity, and delete it, treating its inhabitants as figures in a representation when each of them is a subject with as full a claim to existence as the maker.

The question of what is owed reaches the physics as a question about what the whole construction is for. The technical choices that crowd forward first, such as the number of spatial dimensions, the form of the fundamental equations, or the smallest unit the substrate will resolve, all presuppose an answer to it.

Aristotle had a word for the art that answers such a question. Near the opening of the *Nicomachean Ethics* he observes that every craft and inquiry aims at some good, and that crafts form hierarchies. Bridle-making and the other trades that equip horses serve horsemanship, and horsemanship, along with every other military activity, serves generalship. The ends of the governing arts are more choiceworthy than the ends beneath them, since the lower ends are pursued for the sake of the higher.

He calls the governing arts architectonic and the art at the summit of such a hierarchy the most architectonic of all, *architektonikē* in the fullest sense, a word formed from *archi-*, meaning chief or first, and *tektōn*, meaning builder or craftsman. An *architektōn* was the master builder on a Greek construction site, the person who held the plan of the whole and directed masons, carpenters, and metalworkers toward a purpose none of their individual skills could define. Aristotle gives this master role to politics, which decides which sciences a city should cultivate, who should learn them, and how far. Its end, the human good, contains the ends of all the arts it directs.

Designing a reality stands in exactly that relation to the sciences it draws on. Particle physics, cosmology, thermodynamics, chemistry, biology, information theory, and the science of mind all contribute a competence the designer cannot do without. They answer questions about how to bring an effect about, and none of them can say which effects deserve to be brought about. The governing art orders them by the good of the beings who will live in the result. The physicist chooses a star's spectrum with skin and

eyes in mind, and the biologist builds the signaling of bodily damage with attention to how it will feel.

Aristotle's master art was also political in the ordinary sense, concerned with the shared life of many beings under common laws. The design of a reality is political in the most literal sense available. Its designer legislates at the level of physical law for a population that had no voice in drafting the laws it lives under.

The question of what the construction is for finds its answer in that population. Conscious experience is the ground of all value. A universe of galaxies, symmetries, and exquisite laws in which nothing is ever felt by anyone contains nothing good or bad for anyone. Good and bad are always good and bad for some subject, and where no subject exists the terms find no purchase. Once a subject exists, its good reaches beyond its experiences to real relationships, knowledge of its world, and freedom from domination. Stars, laws, and structures matter through what they mean for beings able to experience them. Elegance, parsimony, and grandeur earn a place in the design by entering someone's life, as beauty to a perceiver or as intelligibility to an inquirer. A designer who began elsewhere, holding that a reality exists to display mathematical beauty, to generate interesting phenomena, or to honor its maker, would accept a physics that tortures its inhabitants whenever the torment produced a pleasing enough equation or spectacle. The record of our own universe shows how easily an elegant physics coexists with immense suffering.

Placing the subject at the foundation also extends moral standing to every being organized so as to have experiences, whatever it is made of, whether it evolved or was designed, and however many layers of computation separate it from the primordial cosmos. That extension rests on a working hypothesis, that consciousness arises from the functional organization of a system and can therefore be realized in neurons, in silicon, or in the matter of a physics no one has yet written. The hypothesis could be mistaken. Under that uncertainty the sound policy is precau-

tion, which treats any system that might be conscious as a possible subject.

With subjects as the measure, each physical parameter becomes a decision about lives. Take the fine-structure constant, the dimensionless number near $1/137$ that measures the strength of electromagnetism. Together with the electron's mass, its value fixes how tightly nuclei hold their electrons, and therefore the energies of chemical bonds, which in our universe fall at a few electron volts. Photons of visible light carry energies of the same order. That coincidence of scale allows sunlight to drive photosynthesis and a single photon to twist a pigment molecule in the retina and produce sight.

Ultraviolet photons from the same star carry only slightly more energy, enough to fuse adjacent bases in a strand of DNA into lesions that cells repair imperfectly. A fair-skinned child spends summer afternoons on a beach, and small errors accumulate in the cells of her shoulders. Forty years later a dermatologist finds a dark, irregular mark that has already spread to a lymph node. Her illness, her children's fear, and the months of treatment run back through mutation, ultraviolet chemistry, and stellar spectra to a single number. The same number stands behind every leaf that turns sunlight into sugar and every eye that has ever seen a face.

Other laws and constants reach into lives just as directly. Gravity's strength settles how hard a child lands when she falls from a tree. Nuclear stability governs how much radiation seeps from the ground beneath a village. The second law of thermodynamics drains heat from every warm body into colder surroundings, so a hiker lost on a winter ridge may not survive the night. So far as anyone can tell, the primordial cosmos arrived at its values through no intention, which leaves its harms without anyone we can hold responsible for them. If some intention did lie behind those values, it was one that gave little consideration to the beings who would suffer under them. A designer who selects a value knowing its consequences takes up responsibility for all of

them. The melanoma downstream of a constant enters the same account as the harvests and sunsets the constant makes possible.

Some 170,000 human beings die on an ordinary day in our world, most of them from aging and the diseases that accompany it. Those processes are rooted in an evolutionary history that rewarded reproduction early in life and stayed indifferent to whether anyone wished to go on living. Beneath that figure lies a far larger toll among other animals, killed by predators, parasites, starvation, cold, and age in numbers no census records. Forced annihilation is among the gravest harms a conscious being can suffer. It destroys the subject and with the subject every good that being might yet have had, including friendships, discoveries, and pleasures still ahead, along with the chance to repair what went wrong before. The harm grows with what remains to be lost, so it is greatest where the future taken away would have been long and good. The capacities that would make designed realities possible bear directly on that toll. A civilization that understands consciousness well enough to sustain it on a computational substrate could perhaps carry the minds of the dying into worlds built without aging, provided the passage is made in a way that preserves each mind as the same subject. The same science of bodies and minds could also bring aging to an end within the primordial cosmos itself.

Philosophy ordinarily proceeds at the pace of journals and seminars, which suits questions that will keep. These questions run on a harsher clock. If designed realities can be built within this century, each year separating understanding from construction corresponds to more than sixty million human deaths and to animal deaths beyond reckoning, every one of them a subject ended without consent. The harm in these deaths lies in what they take from beings who never chose to give it up. Every being able to choose therefore remains free to decline continuation for themselves, which keeps opposition to death from hardening into a duty to exist.

Yet a designer who builds in haste and errs will fill a world with beings who must live inside the error. The urgency that calls for speed calls equally for exactness, and answering it means accepting the slower pace that careful verification of every parameter requires.

Exactness needs a standard to be exact against, and that standard is the argument of this book, which has to answer seven questions in roughly the order a designer meets them. The first asks what makes a world real and what makes it fit to be lived in, and the second asks what the machine beneath a world may spare and what it must carry in full. Next comes who can be trusted to write a world at all. After that the argument turns to standing, which belongs to every being with a point of view, whatever it is made of and however it began. A fifth question asks how the lives in a world are to be judged, and a sixth asks when anyone may bring a new being into existence and what is owed to the dying who already exist. The last concerns keeping a world once it runs, so that its people never live at the mercy of whoever holds its machine. Some of the answers have appeared already, among them that conscious experience grounds value, that standing follows experience wherever it occurs, and that anything which might feel is treated as a possible subject. These and all the rest are argued in the chapters where they do their work, and a reader who wants them in one place will find them gathered at the end of the book, each beside the chapter that earns it.

All of this reasoning converges on someone who does not yet exist. Far downstream of the first decision, after the metric, the forces, the chemistry, and the slow work of biology have run their course, a being will open whatever senses the design has given them and find themselves where all those choices have led. That being will inherit every value set at the empty substrate without having been consulted about any of them. They will live inside those values as plainly as we live inside the gravity that holds us to the floor. Whatever ease or hardship the physics contains will arrive in their life as the way things are, in the texture

of stone underfoot, the length of a day, how far a voice carries across open ground, and whether this life must someday end.

What the designer owes that being has to be paid in advance, in the care taken over every parameter. By the time the being can feel the consequences of a choice, the choice will long since have become a fact of nature.

The substrate is as empty as it was at the start, holding no distance, no change, and no one. Yet the designer facing it is no longer alone. Somewhere beyond the last equation, a being who has not begun to exist is already waiting on every question asked here, and will receive each answer as simply the world.

Part I
Before the Beginning

CHAPTER 1

What Is a Reality?

The Third Table

In the first months of 1927, Arthur Eddington opened his Gifford Lectures at Edinburgh by telling his audience that he had drawn up his chairs to two tables. The lectures became *The Nature of the Physical World*, published the following year, and the two tables became one of the most quoted passages in the popular literature of physics. The first table was the one he had known all his life. It took up space, it stayed where it was put, it had a color, and it was solid in the plain sense that a man could lean his weight on it. The second he had come to know only through physics. It consisted almost entirely of empty space, crossed by a thin scattering of electric charges moving at enormous speeds, and the charges taken together occupied a vanishingly small fraction of the volume the table appeared to fill. The second table served him perfectly well all the same. When he laid a sheet of paper on it, the charges in the wood kept up a ceaseless bombardment against the charges in the paper, and the sheet rode on that bombardment at an almost constant height, held aloft the way a shuttlecock is kept in the air by repeated strokes from below.

Eddington drew a bold conclusion from the comparison. Modern physics, he told his listeners, had assured him that the scientific table was the only one actually there. The familiar table was a compound of the external world with the mind's own imagery and with habits of thought inherited from the past, though

he conceded that physics would never manage to banish it. A decade later the philosopher Susan Stebbing answered him in *Philosophy and the Physicists* (1937), and her reply has aged better than his challenge. Stebbing pointed out that the word “solid” in ordinary speech had never meant “containing no empty space at the scale of the electron.” It meant that a thing resists the press of a hand, keeps its shape, and holds up a book, and the physicist’s table did all of these things. Physics had discovered what the solidity of the table consists of. That discovery explained the familiar table and left it standing. Anyone who took Eddington at his word and stopped trusting tables would have learned nothing from physics and would have spilled a great deal of tea.

Since then the second table has itself come apart. The charges Eddington imagined as small hard particles turned out to be excitations of quantum fields, entities without edges that resist most attempts to picture them, and the fields in turn may be the large-scale face of something stranger still. Students who learn this do not begin to doubt the furniture. Each deeper description of what tables are made of has taken its place beneath the earlier ones, so that the same object now has a familiar face, an atomic anatomy, and a foundation in quantum fields, and no one supposes that the newest description cancels the oldest.

Imagine now a third table. It stands in a room in a world that someone designed. A carpenter there cut it from a tree that grew in soil formed under a star whose spectrum was chosen. She joined its boards with techniques she learned from her teacher, who learned them from hers. The wood is hard because the forces the designer specified make its fibers resist compression. It is brown because the designed light scatters from its surface in a particular way. If the physicists of that world were to probe the table, they would find their own version of Eddington’s second table, some arrangement of whatever fundamental entities the designed physics provides, spread through mostly empty space or packed together in some fashion our physics has no word for. Beneath that lies a level they may never reach, where the laws gov-

erning their particles are carried out as operations in a computational substrate built in the world the designer comes from.

The woman who made the third table sets a cup of tea on it. The cup stays where she put it. Heat flows from the tea into the cup, from the cup into the wood, and from both into the air, at rates fixed by the thermodynamics of her world. Steam rises and curls in the light from the window. When she lifts the cup, a pale ring remains on the wood, and her friend, who is sitting across from her, sees the ring as well and wipes it away with his sleeve. If we ask what has changed for her compared with Eddington at his desk in Cambridge, we find nothing that she could ever discover in the table. It resists her hand, holds her cup, keeps its shape, and looks the same to her friend as it looks to her. In every respect that Stebbing identified as the meaning of solidity, her table is solid.

What has changed for her lies behind the table instead of within it. Beneath the third table there is a machine, and beside the machine there was once a designer, a person who decided that tables should be possible in this world and arranged its laws so that they would be. Many people, told about such a world, feel at once that its tables are somehow less than fully real, along with its trees, its stars, its carpenters, and the friend who wipes away the ring. The word that usually carries this feeling is “simulation.” The feeling behind it is strong enough to shape how those with power over such worlds would treat them.

The question it raises comes before every other question in the design of a reality. Before anyone can choose how many dimensions a world will have or how strongly its masses will attract one another, they need to know what kind of thing they are making. Should the third table turn out to be a picture of a table, the designer is a kind of painter and the people drinking tea at it are figures in the painting, to be retouched or painted over as the artist sees fit. If it is a table, the designer has made a world, and the people in it are as real as the person who made it. Every later decision the designer makes will be shaped by which of these is

true. The question of what a reality is will turn out to have a generous answer, and much of what it was expected to decide will pass to two questions that stand beside it, whether a world could hold anyone and whether anyone is there.

The Senses of a Word

The worry about the third table is easy to feel and hard to pin down, because “real” is one of the most slippery words in the English language. J. L. Austin, in Oxford lectures published after his death as *Sense and Sensibilia* (1962), argued that philosophers go wrong whenever they treat the word as though it named a single property that things either possess or lack. He observed that “real” is hungry for a noun. A thing is never simply real. It is a real duck, a real diamond, or a real Rembrandt, and what the word asserts changes with the noun it attaches to. He observed further that the word’s negative uses do most of its work. To call something a real duck is to rule out some particular way of failing to be one, and which way is meant depends on the occasion. The duck might have been a decoy, a toy, a painting, a reflection in a pond, or a trick of the light. Austin called a word of this sort a “trouser-word,” because its negative use wears the trousers. Asking whether something is real, without any sense of what else it might be, gives the word nothing to do.

Austin’s lesson supplies a method. The suspicion that a designed reality falls short of being real becomes tractable once the suspicion names its alternative. Each thing a designed reality might be mistaken for names a possible way of falling short, whether the alternative is a dream, an illusion, a recording, or a model. Each can be examined in turn to see whether the shortfall is real and whether a designed reality suffers from it. The account of what a reality is can then be built up from the cases instead of announced in advance and defended afterward.

David Chalmers has carried out a related analysis in *Reality+* (2022), the most sustained philosophical treatment of virtual

worlds yet written. Chalmers distinguishes five things people may mean when they ask whether something is real. The question can concern bare existence, or it can concern causal power, the capacity to make a difference to anything. Someone may want to know whether a thing is independent of minds, so that it would remain as it is whatever anyone thought about it. A fourth sense contrasts the real with the illusory, and a fifth asks whether something is an authentic instance of its kind, the way a diamond is a diamond and a piece of cut glass made to resemble one is not. Chalmers argues that the objects of a well-built virtual world pass most of these tests. They exist as digital objects, structures of information realized in physical computer processes. They have effects, since a virtual ball can knock over a virtual vase and the vase falls for everyone who is watching. They keep their states whether or not a given user attends to them, and a user who perceives them is perceiving something that is actually there. The last of the senses he treats as the most complicated, since whether a virtual object counts as an authentic member of its kind depends a great deal on the kind.

Nearly every point Chalmers makes holds up, though he is answering a different question. *Reality+* is written largely from the standpoint of a user, a person whose body remains in the primordial cosmos and who enters a virtual world through a headset or some later interface. Its central concern is whether such a person, while inside, stands in contact with something real. The designer's question differs from his in two respects. It concerns natives, beings whose bodies are made from the materials of the designed world and who have never lived anywhere else. It also concerns responsibility. The designer needs to know what they will be answerable for once the world is running, and which senses of "real" bear on that. From those two differences comes the main addition his account needs, a distinction between what makes a world real and what makes it fit to be lived in, along with several smaller ones.

Seen from the designer's position, the senses of the word carry very different consequences. Existence is the least in dispute, since no one doubts that a running computational process exists. Causal power matters a great deal more, because a world whose events could not affect one another could not hurt or help anyone in it. Independence from minds matters enormously, though in a specific form. A designed reality plainly depends on a substrate and on whoever maintains it, and the question that concerns its inhabitants is whether it also depends on their own minds. Freedom from illusion matters because an illusion is something done to a subject, and a designer who built a world of illusions would have deceived everyone who ever lived in it. Authenticity is where the word "simulation" does its damage. To call a designed world a simulation is to suggest that its tables are imitation tables and its grief imitation grief, in the way that cut glass is imitation diamond.

For natives, the question of authenticity takes a form that disarms it. Hilary Putnam argued in "The Meaning of 'Meaning'" (1975) that what a word like "water" refers to is fixed in part by the stuff its speakers actually encounter. On his imagined Twin Earth, a planet like ours whose lakes hold a different liquid with the same appearance, the inhabitants' word "water" refers to their own liquid, and they speak truly when they use it. Their liquid is not water in our sense of the word, and it is no counterfeit either, since nothing in their language ever claimed it was the stuff in our lakes. The natives of a designed reality stand in the same position with respect to the natural kinds of their world. Whatever they call gold or water picks out their own elements and compounds, which are real kinds in their own right.

For the many kinds defined by what things do, the point is stronger. A table is a raised flat surface for setting things on, a friend is someone who knows you and wishes you well, and a promise is a commitment given to another person. Nothing in these concepts specifies what the deepest constituents of a table, a friend, or a promise must be. The carpenter's table is a table in

her sense of the word and in ours, and the man across from her is a friend in every sense the word has ever carried. The kinds that matter most to a life, such as homes, meals, songs, quarrels, reconciliations, and the people with whom one shares them, are almost all of this functional sort.

What remains, then, are the senses that matter to the designer and the alternatives that give them content. A dream, a hallucination, a stream of experience delivered to a solitary subject, a recording, a model, and an abstract mathematical structure each name a way the third table might seem to fall short of being a table. Taken in roughly that order, they yield one at a time the marks that distinguish a reality from its imitations, and a world fit to be lived in from one that merely runs.

Near-Realities

Each case that follows resembles a reality closely enough to be confused with one, and most of them fail to be one for reasons that can be stated with some precision. The reasons accumulate, so that by the end of the gallery the account of a reality has assembled itself out of what its imitations lack. One case, the most familiar of them, will turn out to be something other than an imitation.

The Dream

Sometime in the fourth century BCE, the Chinese philosopher Zhuangzi recorded a dream in which he was a butterfly, flitting about content with itself and with no idea that it was Zhuang Zhou. When he woke he found himself unmistakably Zhuang Zhou again, and he could not say whether he was a man who had dreamed of being a butterfly or a butterfly now dreaming of being a man. Some two thousand years later, in the first of his *Meditations* (1641), René Descartes observed that there are no certain marks by which waking can be told apart from sleep, and he used the observation to unsettle his trust in everything his senses had

taught him. Both thinkers used the dream to cast doubt on the waking world. Here the dream serves another purpose. Most people, once awake, have no difficulty deciding which of their two lives was the dream, and the question is what the dream lacks that makes the decision so easy.

The answer cannot be that dream experiences are unreal. The terror of a nightmare is real terror, with a racing heart and sweat still damp on the sheets in the morning, and the comfort of a dream in which a dead parent returns can last for days. Whatever a dream lacks, it does not lack experience. This disposes at the outset of the simplest criterion one might propose for a reality, that it is whatever conscious beings experience. A dreamer experiences a world, and the world is not there.

What the dream lacks begins with a second participant. A dream is produced by one brain for one mind, and no one else can enter it. Two people may dream of each other on the same night, but each dreams alone. The beloved who appears in one of those dreams is a figure the dreamer produced, while the beloved herself is somewhere else entirely. Heraclitus made the point some twenty-five centuries ago in a sentence that Plutarch preserved, observing that those who are awake share one common world while each sleeper turns aside into a world of his own. A dream offers no way to check one's perceptions against anyone else's, because there is no one else.

The dream also fails to persist. Its objects exist only while they are being dreamed. When the dreamer looks away and looks back, the room has often become another room. No one returns to a dream on the following night to find the cup where they left it. People who practice lucid dreaming exploit this instability as a test, reading a line of text, glancing away, and reading it again, because in a dream the words have usually changed. The objects of a dream have no settled character apart from the dreaming of them. They are whatever the dreaming brain supplies at each moment, and their properties are on loan from that brain.

For the same reason, the objects of a dream have no powers of their own. A dream fire may feel hot, but it warms the dream room by no lawful process. The dreaming brain produces the experience of heat and the experience of a room, and nothing in the fire acts on anything in the room. The connections between dream events are associations in a single mind, and the rules that seem to govern them dissolve and re-form as the dream goes on. Gottfried Wilhelm Leibniz, in a short essay on how real phenomena are to be distinguished from imaginary ones, proposed that the surest marks of the real are the agreement of phenomena with one another and with what has gone before, together with the success of predictions based on them. A dream fails this test, often within seconds.

These failures sketch the outline of a reality by contrast. In a reality the same rules hold throughout, and the things within it act on one another by their own powers instead of borrowing their behavior from any single mind. Its objects persist and keep their states when no one attends to them. A world fit to be lived in is also shared, so that many minds can meet in it and check what they perceive against one another. The third table satisfies all of these marks. The woman and her friend both see the ring on the wood, the cup is where she left it when she returns from the kitchen, and the tea loses its heat at the rate her world's thermodynamics prescribes whether or not anyone is watching the steam.

The Hallucination

A hallucination differs from a dream in occurring during waking life, among perceptions that are accurate. In William Shakespeare's *Macbeth*, written around 1606, the title character sees a dagger floating before him with its handle turned toward his hand, and he tries to grasp it. His hand closes on nothing, and he concludes that it is a dagger of the mind. A hallucination is an experience of something that is not there, so the concept presupposes a world against which the experience is measured. One can

hallucinate only in a world that contains things one might have perceived instead.

This settles an important question about the inhabitants of a designed reality. Someone might suppose that all of their perceptions are hallucinations, since the stones and trees they see are, at bottom, states of a computational substrate. That supposition misplaces the standard of accuracy. When the carpenter looks at the tree she means to cut, there is a tree there, a structure in her world with the shape, mass, and grain she perceives, which will fall where she expects once she cuts it. Her perception is accurate at the level at which perception operates, which is the only level at which any perception has ever been accurate. Eddington perceived his table as solid and continuous while the table was a cloud of charges, and no one says that Eddington hallucinated his furniture. Perception presents the world at the grain of the perceiver's own life, and it is accurate when it presents what is there at that grain. A carpenter in the designed world can still hallucinate, through fever or exhaustion or grief, and she will do so exactly as we do, by seeing a tree where there is none. Hallucination is a relation within a world, between an experience and the things that surround the one who has it.

Macbeth's test points toward a further mark. He reached for the dagger and met no resistance. Samuel Johnson made the same test famous in an episode James Boswell recorded in his *Life of Johnson* (1791). In August 1763, as the two of them came out of church, Boswell remarked that Bishop Berkeley's argument against the existence of matter, though plainly false, could not be refuted. Johnson answered by striking his foot with great force against a large stone until he rebounded from it, and declaring that he refuted it thus. Philosophers have long pointed out that Johnson misunderstood Berkeley, who never denied that stones resist feet. Johnson's foot registered something all the same.

In 1890, in an essay on the origin of our belief in the reality of the external world, Wilhelm Dilthey argued that the conviction does not come from inference at all. It arises from the experience

of will meeting resistance, of effort that encounters something it cannot simply overcome, and the world first shows itself as what pushes back. Philip K. Dick, asked by a college student in 1972 for a definition of reality, offered a version of the same idea, which he later repeated in a speech he wrote in 1978, titled “How to Build a Universe That Doesn’t Fall Apart Two Days Later.” Reality, he said, is that which, when you stop believing in it, doesn’t go away. A hallucination fails Dick’s test. It may vanish under examination, it yields to the hand, and it shifts with the mind that produces it. The stones of a designed reality pass the test as surely as Johnson’s stone did. A native who kicks one rebounds, and one who has argued herself out of believing in it still trips over it in the dark. Its resistance owes nothing to the beliefs of anyone who encounters it, and that is the form of independence from mind that matters to those who live there.

The hallucination also exposes a choice that belongs to the designer. A designed world could be built to deceive. Its sky might show stars that do not exist, points of light set on a dome with nothing behind them, so that the astronomers of that world spend generations charting objects that were never there. In such a world, the inhabitants would perceive the lights accurately and judge falsely about what they were. The illusion would be part of the world’s construction, and it would be an act of deception by whoever built it. A designed reality is free of illusion only when its designer refrains from building illusions into it, which makes freedom from illusion a matter of the designer’s conduct instead of a guarantee that comes with the substrate.

The gravest form of painted sky would be painted people, figures that seem to be subjects and are none. A designer might fill a street with extras so that it feels peopled, or give a lonely inhabitant a companion who seems to love her, with no one behind the face in either case. Such figures deceive where deception costs most, since a life is built around other people, and an inhabitant who confides in a painted friend has spent her trust on no one. A world may hold such figures only where their nature is disclosed

to everyone who meets them, so that no one mistakes a figure for a person, as no one at a puppet show mistakes the puppets for people. Disclosure leaves a harder question standing. Anything convincing enough to be company has to be weighed as a possible subject in its own right, since whatever makes a figure convincing may be what makes it someone, and toys, whose youngest owners cannot understand any disclosure, raise the question where disclosure cannot reach.

Two Machines

In *Anarchy, State, and Utopia* (1974), Robert Nozick asked his readers to imagine a machine that could give them any experiences they wished. Neuroscientists would stimulate the brain so that a person would think and feel that they were writing a great novel, making a friend, or reading an absorbing book, while in fact they floated in a tank with electrodes attached to their head. The experiences could be chosen in advance from a large library, and the person inside would not know they were in the machine. Nozick asked whether anyone would plug in for life, and he expected most people to refuse. His explanation of the refusal had three parts. People want to do certain things, and not merely to have the experience of doing them. They also want to be a certain sort of person, and someone floating in a tank is, as he put it, an indeterminate blob. Finally, plugging in would confine them to a man-made reality, whose richness could never exceed what its builders were able to put into it, and they want contact with something deeper than that.

A decade before Nozick, Stanisław Lem had imagined a more searching machine. In *Summa Technologiae* (1964, translated 2013) he described a phantomatic machine, joined to the whole of a person's sensorium, that answers each of her acts with the sensations that act would bring, so that when she reaches for a branch she feels its bark and when she turns her head the scene turns with her. Lem asked whether a reality could be made in this way that nothing inside it would give away. He distinguished a peri-

pheral phantomatics, which plays its impulses into the nerves without touching the brain's own processes, from a central phantomatics, which works on the brain itself. Either way, a phantomatic world answers to the one person joined to it. No second subject can meet her inside it to check what she perceives, and she is no participant embedded in it, since her body lies elsewhere and the world is arranged around her alone. Such a world lacks the publicity and the embedded participants of a world fit to be lived in. The figures who seem to share it with her are painted people unless the machine makes them subjects in their own right, at which point it has stopped being a phantomatic machine and become the substrate of a world.

On each of Nozick's three counts, a designed reality stands on the opposite side from the experience machine that critics often take it to be. The first count concerns doing. In Nozick's machine, the novel a person seems to write is never written, since no manuscript exists and no one will ever read it. The friend a person seems to make has no inner life and makes no claim on anyone. The machine is solitary by construction, and everyone in it is furniture in a single mind's experience. In a designed reality, the carpenter's table exists after she has made it, and other people eat at it, lean on it, and admire or criticize its joints. The man across from her is another subject with a perspective of his own, who can refuse her, disappoint her, or surprise her with a kindness she did not expect. The promises they exchange bind both of them, because each can be wronged by the other's breaking them. Doing, in Nozick's sense, requires a world in which one's acts change something that exists independently of one's experience of acting, and a designed reality is such a world.

The second count concerns character. Nothing is at stake for a person in the machine, so nothing they experience can test them. A courage that faces no real danger and a generosity that costs nothing tell no one, including the person who has them, what sort of person they are. In a designed reality, choices have consequences for others as well as oneself. Kindness lands on

someone, a broken promise wounds someone, and a person's character forms as it forms here, through what they do when it matters.

The third count is the one aimed most directly at the project of designing realities, since a designed reality is in some sense made. The answer turns on what exactly is made. The machine's contents are authored experience by experience, each one specified by programmers to suit the preferences of the user, so the machine is exactly as rich as its programmers' imaginations. The designer of a reality authors laws and initial conditions, and perhaps the first bodies and natures of some of what lives there. The chemistry that arises under those laws, the long history of simple life that runs before anything in the world is likely to feel, everything that grows from whatever bodies a designer builds, the languages that form, the music that natives compose, and the friendships they choose are produced by the world and by the beings in it. None of these are specified by the designer, and many of them could not have been foreseen. A designed reality can surprise its maker, and its natives can investigate it for generations without reaching the bottom of what its laws imply. The primordial cosmos acquired its richness in exactly this way, from a short list of laws working on a set of initial conditions over a very long time. A designed world is man-made in its foundations in the sense that a forest grown from planted seed is man-made, and its richness belongs to what grew.

Nozick's machine also helps separate two ways of being present in a world. Users of present-day virtual reality are visitors. Their bodies remain in the primordial cosmos, and their experiences of the virtual world arrive through an interface. The objects they encounter there are real digital objects, as Chalmers argues, but the visitors' lives are lived elsewhere. A visitor who falls from a virtual cliff is bruised nowhere. A native of a designed reality has a body made of that world's matter. When she falls, the bruise is in her world, and it heals or fails to heal according to her world's biology. If minds from the primordial cosmos are carried into de-

signed realities, they will remain visitors for as long as their minds keep running outside. Once a mind comes to be carried by a designed world's own matter and lives there in a body of that matter, it becomes an inhabitant of that world, an immigrant where others are natives.

These contrasts sharpen a distinction the dream first suggested. Some of what the dream lacked is what a world needs in order to be real, and some is what a world needs in order to be lived in. The two machines make the second kind vivid, since what they lack above all is room for participants embedded in a world, made of its stuff, acting on it, and acted on by it and by one another. A stream of experiences fed to a solitary subject offers none of this, however rich the stream.

The Recording

Suppose the entire history of the carpenter's world were recorded, every state of every particle at every moment, and the recording were later played back on another machine. The playback would run through the same sequence of states as the original. The tree would fall where the carpenter meant it to fall, the boards would be joined, the tea would cool, and the friend would wipe the ring from the wood with his sleeve. Event for event, the playback would match the world. The question is whether it would be a second world.

Something essential would be missing. In the original world, the cup stayed on the table because the table pushed up on it. Had the woman set the cup half over the edge, it would have tipped and fallen. In the playback, the cup stays on the table because the next stored state says that it does. If someone altered one frame of the recording by deleting the table, the cup in the following frame would still hang where it had been, supported by nothing. The states of the playback follow one another without producing one another. Each is produced by the mechanism that reads the recording, and none would have been any different had its predecessor been otherwise.

A film offers a familiar version of the same structure. In a film of a gunshot, the image of the gun in one frame does not cause the image of the shattered glass a few frames later. Both are produced by the projector reading the reel. The philosopher of science Wesley Salmon, in *Scientific Explanation and the Causal Structure of the World* (1984), distinguished causal processes from what he called pseudo-processes. A car driving along a road is a causal process, and a dent made in its fender at one point is carried along with it. The car's shadow racing beside it on the verge is a pseudo-process. Darken the shadow at one point by passing it over a patch of black asphalt, and the darkening is not carried forward, because nothing in the shadow at one moment produces the shadow at the next. A playback is a pseudo-process all the way down.

The same distinction has proved decisive in arguments about computation itself. Putnam argued in *Representation and Reality* (1988) that any ordinary physical system, under a suitably contrived mapping of its states, could be said to implement any finite computation, and John Searle suggested in "Is the Brain a Digital Computer?" (1990) that the wall behind him could be interpreted as running a word processor. Chalmers replied in 1996 that implementing a computation requires more than passing through a sequence of states that can be mapped onto the right sequence. It requires a causal structure such that, had the system been in a different state, it would have gone on to the state the computation specifies for that case. What separates a running computation from a wall that happens to pass through the right states is the same thing that separates a world from its recording. The difference lies entirely in what would have happened otherwise.

This yields the mark of counterfactual robustness. A reality is an order of events in which each state produces its successors according to laws, so that had any earlier state been different, the later states would have been different in the ways the laws require. Had the woman hesitated before setting down her cup, the

world would have answered her hesitation. A recording cannot answer anything.

Tim Maudlin, in “Computation and Consciousness” (1989), described a machine he named Olympia, engineered to run through exactly the physical activity that a conscious computation would perform on one particular occasion, with nearly all of the work arranged in advance so that the machine barely does anything during the run. Attached to it, though idle throughout, is the machinery that would take over if the input were different. Whether the idle machinery is attached makes no difference to anything that physically happens during the run, yet computational theories of mind must say that it makes all the difference to whether the run is conscious. Maudlin used the case to press a dilemma on those theories, and it has not been resolved to everyone’s satisfaction. For the designer, it bears on a practical question. If a stored history of an inhabited world were replayed, would the experiences of its inhabitants occur a second time? Lacking counterfactual robustness, the playback is no reality. Whether experience could occur in it anyway is a question about consciousness that no structural test can settle, and until it is settled, a designer who replays a history of beings who suffered should do so with the caution owed to a possibility of suffering repeated.

A milder version of the recording appears whenever part of a world runs on a script. Imagine a world in which most of the inhabitants are real subjects but the crowds in its cities follow fixed routines, walking the same routes and speaking the same lines whatever anyone else does. Those crowds are local recordings embedded in a reality, and a native who discovered them would have discovered that part of her world was a stage set. Unless their nature is disclosed to everyone who meets them, they are painted people as well. The mark of counterfactual robustness applies to regions and populations as well as to whole worlds, and a world can satisfy it in some places while failing it in others.

The Model

The most common way of dismissing a designed reality treats it as a model, and the dismissal has a classic formulation. In “Minds, Brains, and Programs” (1980), Searle argued that running a program that models a mind no more produces a mind than running a program that models a fire produces fire. No one, he observed, expects a computer model of a burning building to set the neighborhood alight, or a computer model of a rainstorm to leave anyone soaked. The argument was aimed at claims about artificial minds, but if it holds at all it holds with equal force for whole worlds. A modeled fire burns nothing, and on this view a designed world is a modeled world whose fires burn nothing and whose rains wet no one.

The reply begins with an observation about where burning and wetting take place. Combustion is a relation between a fire and the things beside it, and wetness is a relation between water and what it touches. Both relations hold among things at the same level of organization. The modeled fire fails to burn the neighborhood because the neighborhood is not at its level, and nothing outside the computer is the kind of thing that fire could touch. A storm on Neptune soaks no one in Ohio either, and no one concludes from this that the storms of Neptune are unreal. The question that matters is what the fire does to the things that share its level, if anything shares it.

Set beside the carpenter’s world, the forecast shows where the difference between them seems to lie. A forecaster’s hurricane floods no streets because the model contains no streets. What it does contain, at its own level, are cells of averaged pressure, temperature, and wind, each many kilometers across, whose states are updated by fixed rules from the states of their neighbors. At that level the model is a small and coarse world of pressure cells, and nothing in it is the sort of thing that could be a street, a house, or a drowning. In a reality built at the grain where fire and wood exist as interacting things, a fire burns what stands beside it. When a house burns in the carpenter’s world, its beams char,

its roof falls in, and the family that lived there has lost its home. The fire does to them exactly what fires do, because they share its level.

At this point the gallery turns. Every earlier case lacked something a reality has, and the model seemed certain to join them. Yet the pressure cells of the forecast, examined closely, lack nothing of the kind, and the reason lies in what it is to be a model. Being a model is a matter of use. Someone chooses a system's states to correspond to features of a target elsewhere and reads those states as standing for the target, so that a forecast represents the atmosphere and a flight trainer represents an aircraft in the air. The representing lives in the relation between the system and the people who use it.

Imagine that every meteorologist forgot what the forecasting program was for. The hurricane would vanish at once, since it existed only in the relation between the numbers and a reader who knew how to interpret them. The pressure cells would remain. They would go on updating, each state producing the next by fixed laws, and a different earlier state would still produce a different later one. A running weather model therefore has, at its own level, the features that separated a reality from the dream, the hallucination, and the recording. Its parts act on one another by their own powers, its states persist whether or not anyone attends to them, and it answers counterfactuals. It is a coarse reality of pressure cells that people happen to use to represent the atmosphere.

Being a model and being a reality are therefore properties of different kinds. The first is extrinsic, a fact about how a system is used, while the second is intrinsic, a fact about how the system's parts act on one another. A single system can have both properties at once, and every running model has the second to some degree. Searle was right that a modeled fire burns no neighborhood, and the reason is that the neighborhood lies outside the model's level. Within a level, the morally decisive question concerns what the level contains, and above all whether anyone is

home. Whether the system is also being used as a model makes no difference to that question, since a system may well be both.

The third table shows the same point from the other side. It stands for nothing. It holds up a cup, the tea in the cup loses heat, and the woman's hand feels the warmth because the warmth is there. If everyone outside her world forgot what the substrate was running, her tea would cool at exactly the same rate. Yet a designer could also use her world as a model, running it to study how civilizations form or how minds respond to loss. Her world would then have a representational use in addition to its structure, as the forecast does, and the use would change nothing about her tea or her grief. That double status is what makes the vocabulary of models hazardous. A designer who thinks of a world under the description of its use can lose sight of the description under which its inhabitants live.

Jorge Luis Borges imagined the limiting case in a paragraph of fewer than two hundred words, "On Exactitude in Science" (1946, translated 1998). The cartographers of an empire, dissatisfied with every smaller map, produced one at the scale of the empire itself, matching it point for point. Later generations found the map useless and left it to the sun and the winters, and scraps of it still lie in the western deserts, where animals and beggars take shelter in them. The fable is usually read as a joke about the futility of perfect representation. It also marks a limit that a map cannot cross and a computation can. However detailed it becomes, Borges's map remains paper and ink lying on the land. Its painted rivers do not flow, and the beggars sheltering in its ruins live in the empire's desert.

A computational model differs from the map because it can run. As its detail grows, the level at which its parts act on one another moves closer to the level at which lives are lived in the world it represents, until pressure cells give way to weather, weather to rain on particular fields, and rain on fields to creatures that get wet. Somewhere along that path the model may come to hold participants, and subjects may arise within it.

Its use need not change when this happens. The team running it may still call it a model and still read its states as answers to their questions, and in one sense they will be right. What has changed is that the model is now inhabited. The line a team must watch is therefore inhabitation, and fidelity is only a rough proxy for it, since a run might cross that line at lower detail than anyone expected or stay empty at higher detail than seems possible. The crossing can happen without anyone intending it, which makes it one of the most serious dangers the whole enterprise contains.

The Structure

The last near-reality is the most abstract and the hardest to dismiss. The physicist Max Tegmark has proposed, in “The Mathematical Universe” (2008) and in *Our Mathematical Universe* (2014), that our physical world is a mathematical structure, and that every mathematical structure exists in the same sense that ours does. On this hypothesis, a structure complex enough to contain self-aware substructures contains beings who experience their structure as a physical world, whether or not anyone ever writes it down or computes it. Greg Egan explored a related idea in the novel *Permutation City* (1994), whose characters come to believe that a world exists once its states can be arranged into an ordered history, regardless of where those states are computed or whether they are computed at all. If Tegmark is right, the designer of a reality brings no world into existence. Every world the designer could specify already exists as a structure, together with its inhabitants and their histories. Running a world on a substrate would then be an act of selection and contact, giving one structure a foothold in the designer’s own world, where it can be observed, altered, paused, and ended.

Whether unrun structures exist is an open question, and nothing in the design of a reality depends on settling it. Two observations hold either way. Running a structure adds something, at least relative to the designer. Whatever the status of the uncomputed structure, the computed one is entangled with the causal

order of the primordial cosmos. Its events can be watched from outside, its course can be changed by the designer's interventions, and its continuation depends on the designer's machines. If its inhabitants had any existence before, they now have a second kind of existence as well, in which their fortunes are tied to the decisions of someone who can reach them.

The designer's responsibility, moreover, attaches to the act of instantiation whatever the metaphysics turns out to be. Should Tegmark be wrong, the designer brings the inhabitants into being and answers for the conditions under which they live. If he is right, the designer brings an existing structure into contact with the designer's own world and makes its inhabitants' circumstances depend on the designer's choices, and the designer answers in the same way for whatever happens to them through that dependence. Parents are responsible for their children whether or not every possible child already exists as a point in the space of possible genomes. No one is cared for or harmed in the space of possibilities.

The marks of reality identified so far also turn out to be internal features of a structure. Causal organization, independence from the minds a world contains, counterfactual robustness, and the capacities for publicity and participation are all properties a structure has in virtue of how its parts are related to one another. Tegmark and his critics disagree about whether such a structure needs a realizer in order to be actual. They need not disagree about which structures would count as realities if they were actual, and that second question is the one a designer needs answered.

What a Reality Is

The cases have answered three different questions, so an account of reality comes in three tiers. To the first question, what makes

something a reality at all, the answer is a lawful order of events bearing three marks.

1. Its events have causal powers of their own. Each state of the world produces its successors according to the world's laws, and the things within it act on one another directly, without borrowing their behavior from a dreaming mind or a stored record.
2. It holds together independently of the beliefs, wishes, and attention of any mind it contains. Its objects persist when no one attends to them and keep their characters whatever anyone believes about them.
3. It is counterfactually robust. Had any earlier state been otherwise, the world would have answered in the ways its laws require.

The second mark concerns belief, wish, and attention, and it leaves minds free to act. A hand that lifts a cup carries an intention into the world through lawful channels, and no one thinks the cup is less real for being moved by it. What the mark excludes is a world whose contents simply conform to what someone believes or wishes, as a dream conforms to its dreamer. A physics in which observation or intention acts on matter directly would still bear the mark, so long as minds work through laws they did not write and cannot suspend, because the will would still meet resistance in Dilthey's sense. The chapters that follow give a designer other reasons never to write such a physics.

A second question concerns what makes a reality fit to be lived in, and two capacities, which a world may have long before anyone exercises them, answer it. The first is publicity, the capacity of a reality to be public, so that many subjects could meet within it, encounter the same objects, and check their perceptions against one another. The second is participation, the capacity to host participants embedded within it, beings made of its matter whose actions change it and who are changed by it in turn. A world that bears the three marks and lacks these two capacities is

a reality that no one could ever inhabit. A one-dimensional cellular automaton whose cells switch on and off according to a simple rule is lawful, causally closed, and counterfactually robust, and it may be far too simple to hold anyone however long it runs. It is a reality all the same, a small and empty one. The primordial cosmos had both capacities from its beginning, long before it held anyone to exercise them.

The first tier is deliberately permissive. A thermostat cycling its heater, a spreadsheet recalculating its cells, and a calculator partway through a sum all bear its marks at their own levels, and so does nearly every running computation. Granting them the name costs nothing, since nothing in them can fare well or badly, and it keeps the question that matters from hiding inside a quarrel over a word. Whether something counts as a reality settles very little. The questions that matter fall in the second and third tiers, whether a world could hold anyone and whether anyone is there.

Measured against these marks, the near-realities sort themselves. The dream lacks nearly all of them, the hallucination fails the second only where an experience presents what the world does not contain, and the recording lacks the third and with it the first. Lem's machine and Nozick's, whatever processes run them, present a world to one subject, who cannot share it and does not live in it. The model fails nothing by being a model, since a running model is a coarse reality put to a representational use, usually far too coarse to be inhabited. The mathematical structure may bear every mark and lack only the realizer that would make it actual relative to anyone outside it, if it lacks anything.

A computationally sustained reality, built well, bears the three marks and both capacities. The substrate carries out its laws, so its states produce their successors, and its objects persist and resist whatever its inhabitants believe about them. Those inhabitants meet one another in a common world that answers each of their actions as its laws require, and the natives among them are made of its matter. It follows that a world sustained on a compu-

tational substrate that bears the three marks and both capacities is a world fit to be lived in, exactly as the primordial cosmos is, and should anyone come to live in it, it will be no less real to them than our world is to us.

Each of the marks admits of degree. A world shared by two subjects is public, though it offers far less opportunity for mutual checking than a world shared by millions. A world with scripted crowds is counterfactually robust in some places and not in others. A world whose sky is a painted dome is real at ground level and illusory overhead. Such cases are best judged region by region and population by population, and by the standard of the inhabitants' lives. The relevant question is whether a failure falls anywhere a subject could reach, rely on, or be misled by. A failure that touches no life is a matter of engineering, while one that touches lives is a matter of conscience.

Computing every grain of sand in every desert at every moment may be impossible even for a very large substrate, and a designer may compute in detail only what is observed, or only what could affect what is observed. The far side of a mountain that no one has visited might exist only as a compact specification until someone crosses the ridge, at which point the substrate works it out in full. This loses nothing in reality, provided the result is consistent with everything that has gone before. The weathering on the rocks must match the rain that fell on them, the tracks in the mud must lead to animals that made them, and a stone left there a century ago must still lie where it was left. Counterfactual robustness concerns what would happen, and a world that would supply a consistent answer to every possible inquiry bears the mark, however little it computes in advance. Quantum mechanics suggests that many properties of our own world lack definite values until something interacts with them, and no one concludes that the primordial cosmos is less real for that. A stage set is a different matter. A facade whose door, when someone finally opens it, gives onto a blank space or an interior that contradicts the evidence has failed the mark. The line between economy and

fakery lies at consistency and has nothing to do with the quantity of computation.

The third question asks what makes a reality matter, and the marks and capacities are silent on it. The primordial cosmos was a reality for most of its history before anything in it could feel. Stars formed and collapsed while galaxies collided, all of it lawful and with every counterfactual in place, and nothing good or bad happened to anyone, because there was no one to whom anything could happen. A reality without subjects is real and morally inert. Destroying it might waste something, perhaps something that a mind observing from outside would have found beautiful, but it would wrong no one within it.

A reality matters when its structural marks meet conscious experience, when there is someone for whom its events go well or badly. Inhabited realities of this kind are the ones a designer sets out to build, and separating these questions clarifies how the case for building them depends on a theory of mind. The working hypothesis is that consciousness arises from the functional organization of a system and not from the particular material that carries the organization. That hypothesis does not decide whether a designed world is real. The structural marks decide that, and they hold whatever the correct theory of mind turns out to be. The hypothesis decides whether anyone is home.

If the hypothesis is true, beings organized as the carpenter is organized, living in a world that bears the marks of reality, experience their lives as she does. Were it false, a designed reality could bear every mark while containing bodies that walk, speak, make promises, and appear to grieve, with no one experiencing anything at all. It would be a world full of grief-shaped behavior and empty of grief. A designer who built such a world while believing it inhabited would have made a grave mistake. The opposite mistake would be graver, because it would fall on beings able to suffer. A designer who treated a populated world as an empty model would pause it, edit it, and delete it as one deletes a file, and every one of those acts would land on someone. Under uncer-

tainty, the asymmetry between these errors settles the policy, which the rest of this book calls the precautionary rule. Any system that might be a subject is to be treated as possibly one, with care scaled to the probability that it feels and to what it stands to lose.

The structural account therefore survives the failure of the functionalist hypothesis. The moral significance of designed realities survives it only where a rival theory still allows subjects. Giulio Tononi's integrated information theory, set out in "An Information Integration Theory of Consciousness" (2004), is one such theory, since it allows minds run on hardware whose own causal structure is of the right kind. The two lines of argument support each other. The structural marks show that nothing about a substrate demotes a world. The theory of mind is needed to show that nothing about a substrate demotes a mind. If both hold, the woman at the third table stands on exactly the footing on which Eddington stood at his desk.

The Medium and the Physics

None of the marks mentions the substrate, which makes it irrelevant to whether a world is real. That claim needs a defense of its own, since it rests on a particular understanding of how a physics relates to whatever carries it.

The relation is realization, which differs from representation in every respect that matters here. A substrate does not depict the physics it carries. The same equation governs ripples spreading across a drumhead, sound traveling through air, and vibrations running along a plucked string. Each medium carries waves that obey the equation, and none of these waves is a picture of the others. A sound wave is no imitation of a ripple on a drum. It is a wave, carried by air. The waves reflect, interfere, and bend around obstacles because of the pattern of relations among the parts of their medium, and very different media can sustain the same pattern. A computational substrate stands to the physics of

a designed reality as air stands to sound. It is the medium in which that physics happens.

This would be a weak defense if our own physics told us what our own medium is, and something about that medium turned out to confer a reality that computation could not. Our physics tells us no such thing. Bertrand Russell argued in *The Analysis of Matter* (1927) that physics reveals only the structure of the physical world, the mathematical relations among events, and is silent about the intrinsic character of whatever stands in those relations. Eddington reached a similar conclusion in the same years, describing physical knowledge as knowledge of structure, of the way pointer readings hang together, with no access to what the structure is a structure of. A century of physics since has deepened the point without overturning it. We know a great deal about how an electron behaves, what it does when it meets a photon, and how its probabilities evolve. We do not know what an electron is in itself, apart from those relations, and it is not clear that the question has an answer.

James Ladyman and Don Ross, in *Every Thing Must Go* (2007), pressed the point further, arguing that there is nothing more to the physical world than its structure and that the search for individual things underneath the relations is a search for nothing. The more cautious reading of structuralism holds that we know only structure, and the bolder reading holds that there is only structure. A computational realizer is no demotion on either reading. If we know only structure, we have no grounds for claiming that our own matter possesses some reality-conferring quality that a computational substrate would lack, since we have no knowledge of our matter's intrinsic qualities at all. Should there be only structure, nothing exists beyond it for a computational substrate to be missing. In both cases, the substrate of a designed reality is one more answer to a question our own physics leaves open.

Realization of one level by another is also the ordinary condition of everything we know. A living cell is realized in molecules,

the molecules in atoms, the atoms in quantum fields, and no one regards the cell as less alive for being made of things that are not alive. Each level has properties and laws of its own, and each is carried by the level beneath. If being realized in something else made a thing unreal, nothing would be real except the lowest level, which no one has found and which may not exist. Some physicists already suspect that something like information lies near the bottom of our own world. John Wheeler's slogan "it from bit" (1989) expressed his conjecture that every physical quantity derives ultimately from the answers to yes-or-no questions. The conjecture remains speculative, but its seriousness among physicists suggests that the line between a physical substrate and a computational one may be less sharp than it first appears.

None of this makes the substrate unimportant. It determines which physics can be carried, at what resolution, at what speed, and at what cost, and a designer will have to understand it intimately. The substrate is irrelevant to whether a world is real and central to which worlds can be built.

The substrate remains invisible to the inhabitants only so long as the physics conceals it, and a physics may let it show. In 2012 the physicists Silas Beane, Zohreh Davoudi, and Martin Savage asked what signatures our own universe would carry if it were computed on a lattice of the kind used in present-day calculations of the strong nuclear force. They proposed that the lattice spacing would impose a ceiling on the energies of particles, and that the most energetic cosmic rays would show a preference for the lattice's directions, a faint texture in an otherwise uniform sky. Whether our world rests on any such lattice is unknown. The proposal shows that a substrate can leave fingerprints on the physics it carries.

For a designer, those fingerprints are a choice. A physics can be built so that close enough inquiry reveals its floor, or so that the floor stays hidden as far down as inquiry can go. Discovering the floor would teach the inhabitants something true about their

world and would take nothing from it. Concealment is a different matter only when it becomes active, when the physics manufactures false evidence instead of simply declining to display its foundations. Opacity of the kind our own physics shows about intrinsic natures deceives no one. A planted history or a painted sky would. Every physics answers, by what it reveals and what it conceals, the question of how much its inhabitants can learn about where their world comes from, and so the designer cannot avoid answering that question either.

Objections

Four objections are strong enough to require a direct answer.

The Objection from Dependence

A designed reality exists only so long as its substrate runs and its keepers allow it to run. The primordial cosmos, so far as anyone knows, depends on nothing of the kind. One might conclude that a dependent world is less real than an independent one, in the way a reflection is less real than the face it reflects, since the reflection vanishes when the face turns away.

The analogy fails at the point that matters. A reflection depends on the face for its content as well as its existence, and it has no causal life of its own. The carpenter's world depends on its substrate for its continuation alone. Its contents arise from its own laws, and its events produce one another. The real question is whether dependence for continuation diminishes reality, and the long tradition of Western theology answers it with unusual clarity. Thomas Aquinas held in the *Summa Theologica*, written between 1265 and 1274, that every created thing depends on God for each moment of its persistence as well as for its coming into being, so that if God withdrew his sustaining action, all things would fall back into nothing. Descartes argued in the Third Meditation that the power needed to preserve a thing from one moment to the next is the same power that would be needed to cre-

ate it anew. For both thinkers, the world existed only through a continuous act of sustaining, and neither concluded that the world was therefore counterfeit. They held that it was wholly real and wholly dependent, and they saw no tension between the two.

Dependence bears on how secure a world is and leaves its reality untouched. A child depends on her parents for food and shelter, a patient on a ventilator depends on the machine for every breath, and no one supposes either of them is less real. Nor does anyone know that our own world depends on something. Theologians have claimed for millennia to know it, along with a great deal about the will that does the sustaining, and in all that time they have produced no evidence that physics could recognize. Aquinas and Descartes analyzed the concept of dependence with real care. Their error, and the error of the tradition they served, lay in claiming to know that the concept applied and what the being on its far side wanted. Physics describes how the cosmos behaves and stays silent about whether anything sustains it. What dependence changes is the moral situation of everyone involved, since whoever controls the substrate holds the power to end every life in the world it carries, all at once and without warning.

The Objection from Deception

Inhabitants who do not know that their world runs on a substrate believe a great many things about it, and one might conclude that they are deceived about all of them. Chalmers answered this objection in “The Matrix as Metaphysics” (2003). A being that has lived all its life in a computed world holds true beliefs about its hands and tables, because its words refer to the hands and tables of its world. The hypothesis that it lives in such a world is a hypothesis about what underlies its physics, comparable to the hypothesis that matter consists of quantum fields. Such a being lacks one truth about its foundations, as we lack the corresponding truth about ours, and a native philosopher who declares her particles the bottom of everything is mistaken in the way our ancestors were mistaken about indivisible atoms. Ignorance of this

kind becomes deception only when a designer builds falsehoods into the world, as in the case of the painted stars considered earlier. Whether natives should be able to discover their origins is then a question about the designer's conduct, and it leaves the reality of their world untouched.

The Objection from the Substrate

Searle's biological naturalism holds that consciousness depends on specific causal powers of biological brains, which running a program does not by itself reproduce. If Searle is right, the natives of a designed reality might be empty in the way described earlier, behaving as subjects while experiencing nothing. The objection therefore belongs to the question of whether a designed world is inhabited, which the precautionary rule already addresses, and the structural account stands either way. A follower of Searle can grant that fires in the carpenter's world burn its houses while doubting that anyone there feels the heat. Whether a brain built from that world's matter would have the powers Searle requires turns on whether those powers belong to the pattern of processes in a brain or to something about our particular matter that the pattern leaves out, which is the very point he and his opponents dispute.

The Objection from Multiplicity

Suppose two copies of a designed world were run in lockstep on separate machines, every state of one identical to the corresponding state of the other. One may ask whether there would then be two realities or one, and whether every pain in that world would be felt twice or felt once and realized twice. Nick Bostrom argued in "Quantity of Experience" (2006) that duplicating a brain duplicates its experiences, so that two identical runs contain twice as much experience as one. If he is right, running a second copy of a world that contains suffering doubles the suffering, and running a second copy of a world full of joy doubles the joy. Related puzzles multiply quickly. A world might be halted for a year of

substrate time and then resumed, and its inhabitants would notice nothing, since their experience would continue as though uninterrupted. From inside, a pause and an ending are indistinguishable until the moment of resumption, which in an ending never arrives. A world might also be copied at some moment and the copies allowed to diverge, producing two histories whose natives remember the same past, and one may ask which of them is the continuation of the original.

These puzzles are less objections to the account than questions it makes precise, and the lockstep copy completes a series that began with the replayed recording. A replayed recording passes through the states of a world without those states producing one another. Maudlin's Olympia keeps its causal machinery but leaves nearly all of it idle during the run. A lockstep copy has everything, with its states producing their successors and its counterfactual structure fully in place, so it is a second running world and no playback. Because a reality is defined by its causal organization instead of its location, two lockstep copies are two complete instances of a single pattern, each bearing every mark. The difficulty that Maudlin's case raised for the replay returns here in a sharper form, since nothing structural is missing from either copy. Whether two instances of a pattern of experience amount to two experiences is a question about consciousness, which the structural account cannot settle by itself. The account does settle one thing. Once two copies diverge, even by a single event, each contains subjects with histories the other lacks, and nothing done to the inhabitants of one can be excused by the existence of the other.

The First of Its Kind

The name "primordial cosmos" identifies our universe by its position. It stands first in the sequence of realities that might be built from within it, as the world in which the first substrates will be assembled and the first designers will sit before them. The name

makes no claim that nothing lies beneath our world, and the claim would be difficult to defend.

In “Are You Living in a Computer Simulation?” (2003), Bostrom argued that at least one of three propositions must be true. Nearly all civilizations at our stage of development might go extinct before acquiring the capacity to compute minds like ours in detail. Alternatively, nearly all civilizations that acquire the capacity might choose not to use it to compute many histories of their ancestors. If neither of these holds, then computed minds with experiences like ours vastly outnumber uncomputed ones, and we are almost certainly among the computed. The argument has been contested on many fronts, including its assumptions about which observers belong in the relevant comparison class and about whether consciousness can be computed at all, and it has not been refuted to general satisfaction. If its third proposition is true, the primordial cosmos is primordial only relative to us, the first world in our own lineage of realities and perhaps one of many worlds in someone else’s.

The same marks protect our world along with every designed one. Should our cosmos turn out to run on a substrate we cannot see, it would lose nothing in reality by that discovery. Its tables would still resist, its stars persist, and its people meet in a common world whose events answer their choices, and we would still be made of its matter. Every grief in its history would remain real grief, every kindness a real kindness, and every life that ended here a real life that really ended. The account that establishes the third table as a table is the same account that keeps our own tables from being called fakes. Anyone who held that computed worlds are unreal would be committed to a strange conclusion if Bostrom’s third proposition were ever confirmed, namely that their own children had never existed. No one would accept that conclusion, and its unacceptability counts against the premise that leads to it.

The primordial cosmos is therefore best understood as one member of the family of worlds fit to be lived in, the first we

know of and so far our only home. Its standing in that family rests on the marks it bears, which it shares with every world that bears them, and owes nothing to the accident of having come first. Human thought has made a similar adjustment before. Copernicus moved the Earth from the center of everything to an orbit around the sun, and later astronomy made the sun one star among hundreds of billions in one galaxy among many. Each step took something away from the sense that our place was the measure of all places, and each left the Earth exactly as solid underfoot as it had been. The primordial cosmos may be due for the same adjustment. It would stop being the definition of reality and become the eldest member of that family, and it would remain, for those who live in it, every bit as real as before.

Words for the Work

These distinctions yield a vocabulary for the work of designing realities. The primordial cosmos is our universe, first in the sequence of realities that may be built from within it. A designed reality is one whose laws and initial conditions were chosen by a designer, and a computationally sustained reality is one whose laws are carried out by a computational substrate, the first term concerning where a world's laws came from and the second the medium that carries them. The substrate is that medium, the system on which a reality's physics runs, and for the primordial cosmos it is unknown. A reality's physics is the set of laws it obeys together with the entities those laws govern, as distinct from the substrate beneath them, and a level is a layer of organization at which things act on one another by their own laws.

A reality that bears the capacities for publicity and participation as well as the three marks of reality is fit to be lived in, and the subjects who actually live in one are its inhabitants. Natives have never lived anywhere else, whether they arose there or were built there, and immigrants came to it from another reality. Both live in bodies of its matter or, where a world allows it, without

one. Visitors experience a world through an interface while their minds run elsewhere, and they are not its inhabitants.

The words “model” and “simulation” stay in this vocabulary as names for a use, since a system is a model whenever someone reads its states as standing for a target elsewhere, whatever its own structure may be. They name real and necessary tools. A designer will run simplified versions of a proposed physics long before building a world, to learn whether stars will form, whether chemistry will be stable, and whether some hidden instability lies in wait. Each such run is a model in the sense that matters and also, at its own level, a small reality, usually far too coarse for anyone to live in. The danger described in connection with Borges’s map lies in the distance between those two descriptions. A word for a use says nothing about whether anyone is home, and it can outlast the emptiness it once implied, so that a team goes on pausing, editing, and discarding its runs long after they have begun to contain someone. The question of what a run’s level now holds has to be asked again each time the run grows richer, and these terms are meant partly as a safeguard that keeps it from going unasked.

The Keeping of a World

The third table, then, is a table. The medium beneath it subtracts nothing from its solidity, or from the reality of the woman whose cup it holds and the friend who wipes away the ring. So far as anyone knows, her world differs from the primordial cosmos in only two ways. Someone chose its laws, and its continued existence depends on a machine and on whoever keeps the machine running.

That dependence leaves the world’s reality untouched, though it hands the keeper of a substrate a power that no one in the primordial cosmos has ever held over anyone. That power includes the ability to halt every life in a world at once, to alter the laws of that world while its inhabitants sleep, to watch

everything that happens in it, and to end it without leaving any trace that anyone inside could detect. The inhabitants would have no recourse against such a power and might never learn that it existed. A fact that is harmless to a world's reality becomes, once it is placed in someone's hands, the most serious political fact about that world.

The woman at the third table knows none of this. She finishes her tea and sets the cup down, and the table holds it, as tables do. Her friend laughs at something she has said. The light from the window has moved a little way across the floor while they talked, and the evening is coming on. The table beneath her hands is solid in every sense Stebbing gave the word. It is made of something she cannot see, and far beneath anything she could ever reach, it rests on a machine that someone has to keep running.

What Carries a World?

The Physics of Computation

In May 1981 a few dozen physicists and computer scientists gathered at Endicott House, a country estate outside Boston that MIT keeps for conferences, to talk about the physics of computation. The meeting had been organized by Edward Fredkin and his colleagues at MIT together with researchers from IBM. It rested on a suspicion that the two subjects in its title might be one subject seen from two sides. Computers are physical objects, so the laws of physics limit what they can do. Physical systems change according to rules, so in some sense every physical system computes. Fredkin held the stronger version of that thought. He believed the universe itself is a computer, a vast grid of simple cells updating by simple rules, and he spent much of his career looking for the rules. Richard Feynman, who gave the keynote, was more cautious. His lecture, published the following year as “Simulating Physics with Computers,” became one of the founding documents of quantum computing.

Feynman began with a demanding standard. He asked whether a computer could imitate physics exactly, as opposed to approximately, and he set a condition on what would count. The number of computing elements needed to imitate a physical system should grow only in proportion to the volume of space and time the system occupies. A machine that met the condition would do for each small region of the world roughly what that re-

gion does for itself. It would keep pace with nature in a fair sense, never needing more machinery for one cubic meter of air than for any other. Feynman thought classical physics might be imitated this way once space and time were broken into small enough pieces. Quantum physics was another matter. The probabilities of quantum mechanics arise from amplitudes that combine across every possible configuration of a system. A machine built of classical parts can reproduce them only by tracking a number of quantities that grows exponentially with the number of particles, so that doubling the size of the system roughly squares the size of the machine. Feynman concluded that a faithful imitator of nature would have to be built of quantum parts, a computer whose own components obey the laws it imitates.

The idea that nature itself computes is older than the conference, going back at least to Konrad Zuse, who built the first working programmable computer in Berlin in 1941 and proposed in *Rechnender Raum* (1969) that space might be a computing machine running the laws of physics as its program. Whether the foundation of a world computes makes no difference to its reality. What the Endicott House meeting offers the designer is its direction of approach, since everyone in the room took nature as given and asked how a machine might keep up with it.

For the designer, no nature yet exists for the machine to keep pace with, since the physics has not been written. Feynman asked whether a machine could keep pace with nature, and the designer asks which natures a machine can keep. Behind that question stand several more. Some physics will be cheap to carry and some ruinously expensive, while some may lie beyond any machine at all. The arithmetic of the machine may leak into the laws it carries, and the machine itself will sometimes fail. Every limit of this kind will press the designer toward savings, and each saving will land somewhere in the world. That the substrate takes nothing from a world's reality is already settled. Which worlds it can carry is a matter of engineering.

In the carpenter's world, the world of the third table, a boy stands at the edge of a lake toward evening. He has spent some minutes choosing a stone, flat, smooth, and about the size of his palm, and he throws it low and hard with a flick of the wrist. It touches the water four times before it sinks. At each touch a ring opens and widens, and the rings cross one another, building small peaks where they meet and canceling where a crest falls into a trough. A soft sound carries across the water with each touch. The sun is low, and its light breaks on the moving surface into a scatter of bright points, while below the surface a few fish turn away from the disturbance and settle again. The boy laughs and begins looking along the shore for another stone.

Every part of that moment has to be carried. The rings are waves in water, whose laws the substrate carries in the way air carries sound. Each crossing of one ring with another is therefore work the machine must do, as is every point of light and every turn of a fish. Most of that work goes into the water. The lake is large, its surface is never still, and its depths are full of slow currents and shifts of temperature that no one in the carpenter's world will ever notice. The boy is small by comparison, yet he is the reason the moment matters, since the evening is good for him and the water is good for no one except through what it gives to the creatures in it and beside it. A first impression forms easily here, that the expensive part of the moment is the water and the valuable part is the boy. The impression holds, in a sense more exact than the one it first suggests.

The Depth of a World

What it costs to carry the boy's evening depends above all on how deep the world goes, meaning the finest level its physics resolves. Described at the depth of our own physics, the lake is a configuration of quantum fields. A quantum state assigns an amplitude to every possible arrangement of a system's parts, and the number of arrangements grows exponentially with the number of parts.

A few hundred interacting particles, each with only two possible states, already have more joint arrangements than there are atoms in the observable universe. A lake of modest size, holding a cubic kilometer of water, contains something like 3×10^{37} molecules, so no machine built of classical parts could ever hold its full quantum state. The lake holds that state with ease, and a machine of quantum parts could do the same, as Feynman concluded and Seth Lloyd showed in “Universal Quantum Simulators” (1996) for any physics whose parts act only on their neighbors. Such a machine needs resources that grow only modestly with the size of what it carries, and its steep price lies elsewhere, in the machinery it needs for correcting its own errors.

Described at the grain of molecules, the same lake becomes a finite problem. Each molecule is treated as a small body moving under forces from its neighbors, and a machine advances them all together in steps of about a femtosecond, a millionth of a billionth of a second, which is the time scale on which molecules vibrate. Chemists run computations of this kind every day on systems of a few million atoms. Applied to a whole lake the method is astronomically expensive, though no longer impossible in principle. Allowing a hundred elementary operations for each molecule at each step, a deliberately low figure, the whole lake asks for about 3×10^{54} operations every second. Described at the grain of fluid flow, the lake becomes cheap. Its water is treated as a continuous medium with a velocity, a pressure, and a temperature at each point, governed by the equations of hydrodynamics on a mesh of cells, and a single modern workstation can produce a surface of rings and reflections that most observers would accept as water. The gap in cost between the second and third descriptions runs to dozens of orders of magnitude, and on classical machines the gap between the first and second has no useful measure at all.

Feynman’s problem was to imitate a nature already fixed, so for him the depth was set by the world and the cost followed from it. The designer is placed differently, because a designed

world imitates nothing. The carpenter's world is no copy of the primordial cosmos, and the reasoning that establishes this also frees its designer from any obligation to match our cosmos level for level. The designer decides where the world's physics bottoms out. A world can have fundamental entities far coarser than our quarks and electrons. Its water might be made of a handful of particle kinds with simple interactions and nothing beneath them, or, at the extreme, it might be a true continuum whose smallest drop is still water. Among the technical choices that crowd forward first at the empty substrate is the smallest unit it will resolve. Here that choice becomes concrete. It is a choice of depth.

The choice can be pictured on a ladder of lengths, shown in Figure 1, which runs from the Planck length, about 10^{-35} meters, where present theories are expected to fail, to the edge of the observable universe, some 4×10^{26} meters away. About sixty-one powers of ten separate the two ends. The nucleus of an atom sits some twenty rungs above the bottom, the atom five rungs above the nucleus, and a molecule of water just above the atom. The boy stands about thirty-five rungs up and some twenty-six below the top, nearer by this measure to the edge of everything than to the bottom of his own matter. Each rung is a place where a world's physics could stop, and wherever the designer stops it, everything above has to be built from what lies there.

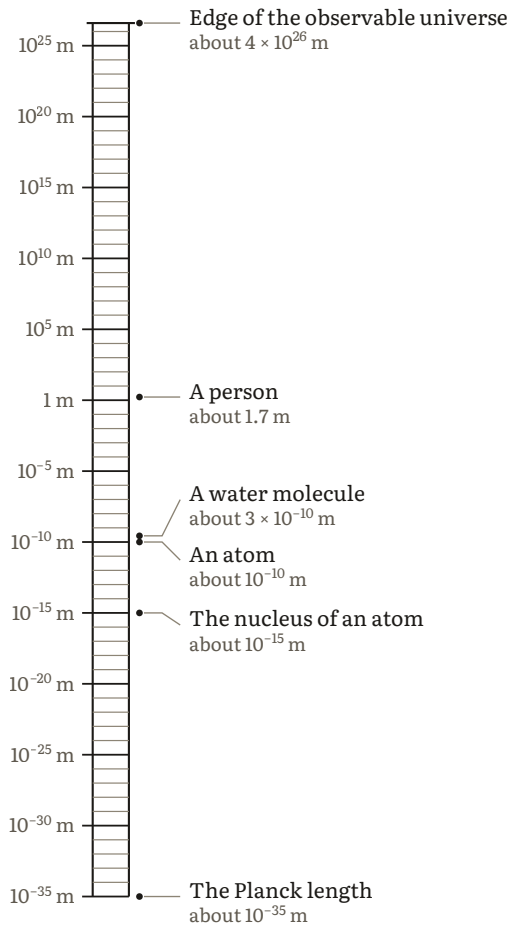


Figure 1. A ladder of lengths, one rung for each power of ten, from the Planck length to the edge of the observable universe.

A designer who thinks of depth only as a cost will build a poorer world than they intend. A deep physics gives chemistry room to be rich, since the variety of possible molecules grows with the variety of ways fundamental things can combine. Life needs that variety in order to arise, because every account of its origin in our own world depends on a chemistry varied enough to try uncoun-

ted combinations until some of them copy themselves. Minds need a medium in which to be built, and the world's scientists need something to find, a sequence of levels each explaining the one above. A world too shallow to hold anyone fails at the one thing that matters, however cheaply it runs. It may be lawful, persistent, and counterfactually robust, and it will still be empty.

Natives are made of their world's matter, and the boy is a native. If his world goes deep enough for chemistry and life to arise on their own, then his brain is a volume of that world's matter, about 1.2 liters of it, and at molecular grain it costs what any 1.2 liters of the world costs. By the same counting, that comes to about 4×10^{42} elementary operations every second, an amount that becomes startling once it is translated into energy. In a deep world the boy is as expensive as the water he is made of, ounce for ounce.

Counted by function instead of matter, a mind costs far less. Estimates of what a human mind requires, if it is computed at the level of its neurons and synapses, run very roughly from 10^{14} to 10^{18} operations per second. Anders Sandberg and Nick Bostrom, in *Whole Brain Emulation: A Roadmap* (2008), laid out a ladder of such estimates, each rung a finer level of detail at which a brain might have to be modeled for the result to be the same mind, with costs climbing steeply from the functional rungs toward the molecular ones. At the lower rungs a mind is cheap, a trivial load beside one lake. That cheapness belongs to two kinds of being. It belongs to emulated minds, such as immigrants carried in from the primordial cosmos and lodged in bodies of the world, provided a mind computed at a functional grain is still the mind it was. It also belongs to the natives of a world that is shallow beneath the level of life, whose bodies are built from coarse units arranged to do what neurons do, with nothing finer underneath.

These two ways of counting set the first real dilemma of design. A world shallow beneath life makes minds cheap, and it may never produce them. Life in our cosmos arose, so far as anyone can tell, from a chemistry rich enough to explore an enorm-

ous space of molecules over hundreds of millions of years, and a coarse physics may offer too little variety for anything like that to happen. Its minds would then have to be built directly and placed in the world, a choice that has to be weighed against the value of letting life emerge on its own. A deep world lets life arise from chemistry and makes every body as costly as the matter it is made of. Minds may have to be built there too, since selection reached them in our world through overproduction and early death, and the ethics of design may forbid that among any creatures that might feel.

Even the molecular grain used in these counts is a simplification. Chemists who compute molecules as small classical bodies are approximating the quantum behavior of electrons. A designer could write a classical molecular physics directly, with bonding rules built into its forces, but whether such a physics could support chemistry rich enough for life to arise on its own is unknown, and it is the very question on which the dilemma turns. If life needs something like quantum chemistry, the deep horn grows costlier still, and quantum machines become its natural substrate. Neither horn is forbidden. Depth is the first decision of substance the designer makes about the machine, and nearly every later decision about the machine depends on where it is set.

A middle path suggests itself, and it brings out a rule that governs every saving a designer might make. The designer could build a deep world, let chemistry and life run at full depth everywhere, and then compute the brains of natives at the cheaper grain of their function, as though each skull held an emulation in place of matter. The saving would be enormous, and it would be taken directly from the subjects. It would be safe only if the designer knew the grain at which consciousness depends on the physical details beneath it. The functionalist hypothesis holds that the right organization suffices for experience, and it says nothing about which level of a brain's organization carries the experience. Nobody knows whether a mind computed at the level of its synapses is the same mind as one carried molecule by mo-

lecule, and precaution refuses to stake a person on the answer. The designer may build a shallow world openly, with natives made of coarse matter from the start. What they may not do is build a deep world and then quietly make its people shallow, because a saving taken on a subject is a harm done to someone who never agreed to bear it.

The same unknown hangs over immigrants. A mind carried in from the primordial cosmos at a functional grain is a bet on exactly the question that forbids the middle path, and the two cases differ in who makes the bet and what it is set against. Natives are given no choice and gain nothing from the saving. An immigrant can take the risk knowingly, and for the dying the alternative is the end of everything, so a risk no designer may impose on natives can be reasonable for someone who accepts it freely with death as the other choice. Consent does not settle the matter. Precaution still asks that the passage be made at the finest grain the budget allows, because the question it raises, whether the one who arrives is the one who left, is the question on which the whole value of the passage depends.

That a designer can choose depth at all is owed to a feature of physics that Kenneth Wilson explained in the early 1970s with the renormalization group. Wilson showed how the behavior of a system at large scales depends on only a few features of what happens at small scales, while the remaining details wash out as one moves upward. Water near its critical point, where liquid and vapor become indistinguishable, shares its large-scale statistics with a simple model of magnets, though water and magnets have nothing in common at the scale of atoms. Many different floors can support the same large-scale world, so the choice of depth is free within limits. The inhabitants' science can also succeed one level at a time without reaching the bottom, as ours did when chemistry flourished for a century before anyone understood the nucleus. Whatever depth the designer chooses, its cost must be paid, and the budget it is paid from belongs to the primordial cosmos.

Ceilings and Costs

Nobody knows what the machines of a civilization with superintelligence will look like. The minds that design them may invent architectures no one living can imagine, and any forecast of their blueprints is a guess. Some claims about those machines are safe all the same, because they concern limits that physics places on any machine built in the primordial cosmos, however intelligent its builders. No mind repeals the laws it runs on. The sound policy follows from that difference. It speaks with confidence about limits and with humility about blueprints, and it marks every forecast as one.

The hard ceilings are few and generous. Norman Margolus and Lev Levitin showed in 1998 that a physical system with a given average energy can pass through only a limited number of distinguishable states each second, a number proportional to its energy. Jacob Bekenstein showed in 1981 that a region of given size and energy can hold only a finite amount of information. Signals between the parts of any machine travel no faster than light. Lloyd combined such limits in “Ultimate Physical Limits to Computation” (2000), a portrait of a one-kilogram computer working at the edge of what physics allows, which would perform around 5×10^{50} operations each second and would be, in its working state, a ball of plasma at a billion degrees. These ceilings sit far above anything a lake requires. Carrying the lake at molecular grain takes about 3×10^{54} operations each second, and by the Margolus–Levitin bound about six tonnes of mass-energy would in principle suffice to perform them. What makes the lake expensive is the heat a real machine sheds while doing the work.

Rolf Landauer of IBM had identified the relevant cost two decades before the Endicott House meeting. In 1961 he showed that erasing a bit of information must release a minimum amount of heat, proportional to the temperature of the surroundings, which comes at room temperature to about 3×10^{-21} joules. The price attaches to forgetting. A logic gate that takes two inputs and returns one output has discarded information about its inputs,

and every register overwritten and every intermediate result thrown away pays the same small toll. Ordinary computers forget constantly and in practice spend many orders of magnitude more energy per operation than Landauer's limit. The limit itself is the part no engineering can remove from a machine built that way.

Take the lake again, its 3×10^{37} molecules carried at molecular grain in steps of a femtosecond, so 10^{15} steps each second, with the same hundred elementary operations per molecule per step. Assume, as ordinary machines do, that each operation erases about one bit, and that the machine works at room temperature. The lake's 3×10^{54} operations per second then release roughly 10^{34} watts of heat. The Sun's entire output is about 4×10^{26} watts, so the lake alone would consume the light of tens of millions of suns. The boy's brain at the same grain, about 4×10^{42} operations each second, would release around 10^{22} watts, hundreds of millions of times all the power humanity now uses. At a functional grain, by contrast, a mind computed at Landauer's limit would need at most a few thousandths of a watt, and the boy's own brain, like ours, runs on about twenty.

These figures are rates, the power to carry the lake at its own pace, which no machine yet conceived could keep. Each femtosecond of the lake's time asks for some ten operations in sequence, so logic switching in a trillionth of a second would run the lake ten thousand times slower than it lives, however much hardware was built. No quantity of hardware passes that sequential floor, since operations that must follow one another cannot be spread across more machines. A budget is better measured in years of the world's own time, whose outside pace the inhabitants can never feel. A forgetful machine pays Landauer's price for every operation of those years, however slowly it runs. At that price alone, such a machine in the cold of space, about three degrees above absolute zero, could carry the lake for some forty thousand of its years on the Sun's lifetime output, and at room temperature for about four hundred.

The estimate holds only as long as its assumption about forgetting holds, and that assumption can be broken. Charles Bennett showed in 1973 that computation can be made logically reversible, with every step undoable, so that nothing is ever erased and Landauer's price need not be paid on each step. A reversible machine still loses a little energy in each operation, and as Michael Frank and Karpur Shukla set out in "Quantum Foundations of Classical Reversible Computing" (2021), the loss shrinks in proportion as the operation is drawn out, so that a switch thrown ten times more slowly wastes a tenth as much. Energy and waiting then become a single currency. For a given task on a machine sized to the world, the energy multiplied by the waiting is fixed. It grows with the size of the world and with the square of the number of steps in its history, since more steps mean both more operations to pay for and less time in which to draw each of them out. A machine half the size doubles it, because each of its parts must work through twice as many operations in turn. The trade holds until leakage, or the few erasures that errors, chance, and upkeep still demand, come to dominate.

Counted this way, the true expense of depth lies in growing a world. Carrying Earth's oceans at molecular grain for the three and a half billion years between the origin of life and the first minds would cost a forgetful machine, even at Landauer's limit in the cold of space, about a hundred trillion times the Sun's lifetime output, the starlight of thousands of galaxies. A reversible machine could pay for the same history from one star's output, hoarded and spent slowly, but only by waiting more than three billion billion times the present age of the universe, and only if its switching lost no more than a speculative best case that no known law forbids and none guarantees. The machine itself would be no small thing. Working on every molecule at once, with a switch about the size of a small protein for each operation, it would weigh as much as seventeen Earths, a forecast of molecular engineering far beyond anything yet built.

Most of that bill follows from choices the designer is free to make. Because the product of energy and waiting grows with the square of the number of steps, nothing counts for more than the period of a world's fastest motion, which sets how many steps each moment of its history requires. Slower fastest motions, a smaller early sea, a shorter history, and laws written as the machine's own reversible rules would together cut that product ten-trillion-fold. With best-case switching, a single star could then pay for the whole history, though only by waiting some three hundred and forty thousand times the present age of the universe. A galaxy, which can spend its light only as fast as it shines, could pay within tens of millions of years, a span set less by its light than by the steps that must follow one another, which take about sixty million years even on the fastest switches the estimate allows. At that pace a reversible switch loses more in each operation than Landauer's price for erasing a bit in the cold of space, so reversibility saves energy only for a machine willing to wait far longer.

Even a forgetful machine at Landauer's limit in the cold of space could grow such a world on about a million times the Sun's lifetime output. It could do so on about a hundred if life's chemistry needed full detail in only a ten-thousandth of the early sea, and the rest, where nothing might feel, could be summarized honestly. Such a history might also be computed once and shared among many worlds that part before anything in them has a serious chance of feeling. Whether life can arise under such a physics is unknown, and the cost of depth turns on that question.

These considerations allow a rough verdict on the deep horn of the dilemma, though the verdict is a prediction. Growing a deep world exactly as ours grew, its oceans carried in full, would take the starlight of thousands of galaxies or a wait far longer than the universe has existed. With best-case switching, a world written for economy comes within a galaxy's reach in tens of millions of years, if life can arise in it. Keeping a world once its people exist, with every one of them carried in full, costs far less, and the hon-

est summary examined below would bring it within a single star's reach. The first decision of design is thereby tied to the relation between the world's clock and its keepers' clock. The machines now running can carry only the shallow horn, and the deep horn waits on machines that no known law forbids and none guarantees.

Each kind of machine also makes some physics cheap and other physics dear. Classical digital machines can carry any physics that can be computed at all, and they carry quantum physics only at the exponential cost Feynman described. Reversible machines, in the line of Edward Fredkin and Tommaso Toffoli's "Conservative Logic" (1982), suit laws that are themselves reversible and can carry them while spending very little energy. Quantum computers make quantum physics natural, as Feynman foresaw, at the price of heavy machinery for correcting their own errors. With today's error rates, a year of the lake at full quantum depth would cost at least tens of thousands of times the Sun's lifetime output, and each of its seconds would take millions of years. Analog emulators take a different road. They engineer a physical system whose own behavior realizes the target laws, as when ultracold atoms held in lattices of laser light are made to obey the same equations as electrons moving through a crystal. Such emulators are efficient, though hard to reprogram and harder to verify. They also blur the line between computing a world and building one.

The learned world models now being built offer a contrast. Trained on video and interaction, they generate explorable environments frame by frame, and at present they hold together for minutes before objects begin to change once attention moves away from them. That is the failure of a dream, in which a line of text read twice comes out differently, and it is a limitation of the present generation that may not last. A substrate for an inhabited world has to carry laws. A learned approximation of laws qualifies once it becomes as consistent as laws are, and at that point it is a physics, however it was arrived at.

Every machine therefore has what might be called a native physics, the set of laws closest to its own operations and cheapest for it to carry. A designer working under budget pressure will feel a steady pull toward that physics, toward a world made reversible because the machine is, quantum because the machine is, or laid out on a regular grid because memory is laid out in rows. Sometimes the pull costs no one anything, and reversible laws may be such a case, since our own fundamental laws are reversible and nothing in a life seems to depend on their being otherwise. Where the pull would cost the inhabitants something, the foundation laid at the empty substrate supplies the correction. One form the correction takes is a machine built to fit the world, with processors laid out to match its geometry or hardware whose own operations carry out its laws, so that the cost falls on the builders in the primordial cosmos. The good of the inhabitants chooses the physics, and the machine's preferences enter only as costs.

Because budgets are finite, those costs are moral too. Every watt spent carrying a lake is a watt not spent carrying a mind, and the same holds for every year of waiting and every kilogram of machine. More than sixty million people die each year in the primordial cosmos, and if the minds of the dying can ever be carried into designed worlds as the same subjects, some of the minds a substrate might carry are theirs. A designer who spends without limit on detail no one will ever experience may be spending what could have rescued someone.

The obvious objection is that superintelligence will make all of this irrelevant. It will not. A mind vastly greater than any human one still pays Landauer's price for every irreversible step at a given temperature, exchanges signals no faster than light, and sheds its waste heat into its surroundings. Freeman Dyson pointed out in 1960 that a civilization using the output of its star must eventually release that energy again as heat, and he proposed searching the sky for the infrared glow such civilizations would give off. The same accounting holds for any substrate. Every warm after-

noon in a designed world is paid for in heat released into the cold of the primordial night.

Limits and Seams

Expense is one kind of limit. A second kind is impossibility, and a third, subtler than either, is the way a machine's own workings seep into the physics it carries. A machine that works in finite steps can compute only what is computable, and it is not obvious that every physics a designer might write down passes that test. Marian Pour-El and Ian Richards showed in 1981 that the wave equation in three dimensions, the same equation that carries the sound of each skip across the boy's lake, can start from initial conditions a computer can specify and evolve into a state no computer can calculate. The result is real and often overstated. The initial conditions in their construction, though computable, are rough in a precise mathematical sense, and when the starting data are smooth the solutions can be computed after all. The lesson for a designer is that computability depends partly on how a physics is specified. A world can be kept within reach of a machine by writing its laws and its starting conditions smoothly, and a physics that demanded truly uncomputable evolution could be carried only by a hypercomputer, a device no one knows how to build and which physics may not permit. David Deutsch proposed in 1985 that every physical process realizable by finite means can be imitated by a universal quantum computer. The principle remains a conjecture, and it marks the boundary a designer should stay inside.

Complexity, the next limit, brings back Feynman's problem. On classical machines quantum states cost exponentially, and the clever methods physicists use to avoid the full cost work for many systems and fail for others. Among the systems they fail for are fermions, the class of particles from which ordinary matter is built. Classical sampling methods for many interacting fermions run into what is called the sign problem, in which contributions

of opposite sign nearly cancel and the useful signal drowns. Matthias Troyer and Uwe-Jens Wiese showed in 2005 that the sign problem in its general form is NP-hard, which places it among the problems widely believed to have no efficient general solution. The most ordinary stuff in our universe, the matter of stones, water, and hands, is among the hardest things for a classical machine to carry faithfully.

Symmetry supplies a third example, showing that our own physics resists the simplest kind of substrate one might build for it. The weak nuclear force acts on particles according to their handedness, treating left-handed particles differently from right-handed ones. In 1981 Holger Bech Nielsen and Masao Ninomiya proved that, under ordinary assumptions, a lattice of points cannot carry a particle of definite handedness without producing unwanted partners of the opposite kind. Physicists have found workarounds that serve well for theories like quantum chromodynamics, the theory behind the lattice calculations that now derive the masses of the proton and neutron. A lattice formulation of the full Standard Model, with its handed weak force, remains an open problem. A designer who chose a physics as lopsided as ours would be choosing difficulty, whether or not anyone in the world ever learned what the difficulty had cost.

Precision, the subtlest limit, leaves the machine able to carry a world while changing which world gets carried. In the winter of 1961 Edward Lorenz, a meteorologist at MIT, was running a small model of the atmosphere on a computer in his office. Wanting to examine one run more closely, he restarted it partway through, typing in values from an earlier printout, and went to get a cup of coffee. The printout had rounded each number to three decimal places, while the machine had carried six. When Lorenz came back, the new forecast had drifted from the old one and then departed from it entirely, as though the two runs described different planets. He published what he had found in 1963 as “Deterministic Nonperiodic Flow,” and the sensitivity he described became a founding observation of modern chaos theory, though

Henri Poincaré had glimpsed it in the motions of planets some seventy years earlier. A running weather model is a coarse reality and a model at once, a small world of pressure cells that people read as a picture of the sky. In Lorenz's office that coarse reality changed its entire future because of the way a printer formatted its numbers.

In a designed world with chaotic physics, the same sensitivity reaches the inhabitants. A difference in the sixth decimal place of one quantity, introduced by the substrate's arithmetic, grows until it is a different evening. The stone skips five times instead of four, the boy laughs a moment later, and the remainder of his life unfolds from a slightly different start. The constants of a world are subject to the same fact. A designer who writes down the strength of a force has written a particular number to a particular number of digits, and in a chaotic world the digits beyond anything the inhabitants will ever measure still decide which histories occur.

Conservation laws can leak through the same seam. A designer may write a physics in which energy is conserved exactly, and a machine carrying that physics by a careless method of stepping forward in time will let energy drift a little with every step. Over long runs the drift accumulates. The rings on the lake might never quite die away, or might die faster than the laws say they should, and the world's physicists, if they measured carefully enough, would discover that their conservation law is false. The designer would have planted a small perpetual motion machine in their world without intending to. Methods exist that prevent this. Symplectic integrators preserve the geometric structure of the laws so that energy errors stay bounded over very long spans, and schemes built on exact integer arithmetic can be made perfectly reversible, conserving what the laws conserve to the last bit.

A last seam lies in the order of arithmetic. Floating-point addition, the ordinary arithmetic of scientific computing, is not associative, so adding the same numbers in a different order can give

a slightly different sum. A machine that divides a world among many processors will add things in an order that depends on how the work was divided and on which processor finished first. Consider two copies of a world run in lockstep, every state of one identical to the corresponding state of the other. On real hardware, two copies of the boy's evening started from the same state could diverge over nothing more than the order of a sum, and chaos would then carry them apart into different histories. Lock-step copies require arithmetic that is reproducible bit for bit, which is achievable and has to be chosen deliberately.

The lesson these cases share is that the inhabitants live under the physics that is actually computed, and the designer answers for any gap between it and the physics that was written. A world whose laws promise conservation and whose machine leaks energy has broken a promise, whether or not anyone inside ever measures the leak. The same gap reaches down to the world's lowest level, and to whether that level is continuous at all.

The Floor

Some generations after the evening at the lake, the boy's descendants will want to know what water is made of. They will find that it is made of smaller things and will ask what those are made of. The questions will continue until their instruments either reach a level with nothing beneath it or keep finding structure beneath every level they reach. Whatever they find will be the floor the designer chose. Their whole science of matter, from the first speculation about indivisible particles to the last experiment their civilization performs, will be a long approach to a decision made before any of them existed.

Our own cosmos has not told us whether its lowest level is discrete. The Planck length, about 10^{-35} meters, is often described as the pixel size of the universe, and the description goes beyond what is known. It marks the scale at which gravity and quantum theory can no longer both be trusted, which is a fact about the

theories. Some candidate theories of quantum gravity do predict a discrete structure. Loop quantum gravity predicts that areas and volumes come in indivisible units, and causal set theory treats spacetime as a random scattering of elementary events, arranged so that no direction or state of motion is singled out. Observations have so far found no trace of discreteness. When high-energy photons from a distant gamma-ray burst reached the Fermi telescope in 2009, they arrived so close in time to lower-energy photons from the same burst that the simplest ways a discrete structure could make the speed of light depend on energy were ruled out, even at the Planck scale.

For the designer the question is a choice, and each option has a price. A regular lattice, with points spaced evenly in rows, is the cheapest to carry, since the machine can lay out its memory the same way, and it gives the bottom of the world a set of preferred directions. Beane, Davoudi, and Savage proposed that such a lattice in our own world might show up in the arrival directions of the most energetic cosmic rays, and the inhabitants of a lattice world could look for the same signature in their sky. A random discreteness in the manner of causal sets would keep the world free of preferred directions at every scale, at the cost of an irregular structure that is harder for a machine to store and to update. Apparent continuity is a third option, available in two forms. Some quantities can be carried exactly by symbolic arithmetic that never rounds, and the machine can also refine its description of a region wherever inquiry presses, supplying finer structure whenever an instrument reaches for it while staying consistent with everything already recorded. True continuity may be uncomputable in places, so a designer who wants the appearance of a continuum keeps it within reach of the machine by writing its laws smoothly.

From inside, at the grain of life, nothing differs among these choices. The boy's stone skips the same way over a lattice, a random scattering, or a refined continuum, because the large-scale behavior of water depends on so little of what lies beneath it. The

difference appears only at the level of the world's deepest physics, where the choice decides what the boy's descendants will find. One possibility is a finite list of fundamental things with nothing underneath. A lattice is another, with its faint preference for certain directions, and finding it would teach them something about how their world is carried. A third is a regress that keeps yielding, with each new instrument revealing structure no earlier theory anticipated. Each is a different ending to the history of their inquiry, and some endings may serve a civilization better than others. Whichever ending they reach, finding the floor would be discovering something true about their world, as the discovery of atoms was for ours, and it would no more demote their world than atoms demoted tables. Whether the floor should be discoverable at all bears on the inhabitants' right to understand their own world through their own inquiry.

A regress supplied on demand carries an obligation the other options lack. If the machine refines its description of matter only when someone probes deeply enough, then every deeper probe must find structure consistent with every shallower result, across every laboratory and every century. The promise has no end date. A world built this way commits its machine to answer, forever and without contradiction, questions its designer never imagined, and the practice of computing finely only where someone looks turns out to be a promise as much as a saving. That practice reaches well beyond the bottom of the world, and it is where the moral questions about savings come to a head.

Economies

A substrate may leave the far side of a mountain uncomputed until someone crosses the ridge, provided the result agrees with everything that has gone before, and that economy differs from a stage set whose door opens onto nothing. The permission needs closer examination, because two different savings travel under it. The first is deferral, which postpones computing a region and

later computes it in full. The second is summary, which replaces the full detail of a region with a coarse description of it, the way statistical mechanics describes a gas by its temperature and pressure without following a single molecule. Game designers know a version of summary as level of detail, the practice of drawing distant mountains with a few hundred polygons and nearby rocks with millions.

Chaos decides between them. In a chaotic world the present state of a region depends sensitively on its whole past, so the weathering on the far side of the mountain is the record of all the rain that fell there, the freezes and thaws of uncounted winters, and the small slides of scree. To know that weathering in detail, the machine must compute the history that produced it. Deferral therefore postpones the work without reducing it, and when a traveler finally crosses the ridge, the machine faces the full bill at once. Summary is where the saving lies. A region carried as a coarse description costs a small fraction of the same region carried in detail, and when someone arrives, the machine generates detail consistent with the summary and with every record already made. The required consistency is hard to guarantee, since a summary throws information away and the generated detail must never contradict anything the summary has already implied. A designer might write laws whose coarse descriptions are exact by construction, so that detail generated later is certain to fit, though this is speculation, since our own physics rarely offers such closure.

Most of the lake can be carried coarsely. Its depths can be a field of slow currents and temperatures, its far shallows a handful of numbers, and its surface computed finely only where light, wind, and a skipping stone make demands on it. Fish swim in it, though. If the world is deep, each fish is made of the world's matter at full depth, and the water that touches its gills and skin must be carried at the same depth, because a body and its surroundings meet at the grain of the body. A fish also senses the water well beyond its skin, feeling pressure waves along its flanks

and hearing sounds from across the lake, and whatever a creature perceives must be carried at the grain at which it perceives it. Around every fish there is therefore a halo of finely computed water reaching as far as the fish can sense, moving with it as it swims. The halo thins with distance, since a sound from across the lake reaches the fish as a pressure wave that a coarse description carries faithfully, while the water at its gills must be carried molecule by molecule. The summarized lake is scattered with these graded regions of detail.

The halos mark a limit that every economy has to respect, since savings are permitted where no one lives and forbidden for anyone who might. Once the life of anything with a serious chance of feeling is under way, it is never summarized, approximated, or carried at reduced detail. The reason is the one that rules out planted histories. A creature's past consists of its experiences, and a summary of a fish stands in for a life with a description of one. If the machine later generated detail consistent with that summary, a fish with memories, habits, and scars, it would be producing the evidence of a life that no one lived, which would be deception built into the world by its designer. This rule is a constraint that admits no trade against the budget, because a creature whose life is already under way stands to lose the whole of it. To summarize it would be to end a life that may belong to someone, and a designer may not end one possible subject to pay for the rescue of another, however favorable the arithmetic. No one can yet tell for certain which creatures feel, so the region this rule protects, the protected region, is drawn generously, around every creature whose capacity to feel is left seriously open by the best available evidence. Fish belong inside it, along with birds, and insects too while their inner lives remain in doubt. Each is carried at the grain its possible experience requires, continuously, whether or not anyone is watching, from the moment it is first computed as an individual until its life ends.

Water flows into a fish's halo from the summarized lake all the time, and the molecules that reach its gills at any moment were

generated to fit the summary. In a world whose fine-grained laws are deterministic, those molecules have no real past. Nothing in the world's history produced their exact arrangement, and the fish lives in a stream of water whose microscopic origins never happened. Summary of this kind is a gap between the physics written and the physics computed, and the designer answers for it. The gap is tolerable only when no inquiry anyone in the world could ever make would reveal it, so that every record, measurement, and possible future probe agrees with some lawful past. Even then, the world departs from its own laws at a grain no one reaches, and the counterfactual robustness that separates a reality from a stage set holds for the world's inhabitants without holding for every molecule. A cleaner course is available if the world's fine-grained laws are themselves partly a matter of chance, as quantum physics suggests ours are, since water generated to fit a coarse description is then a draw from the very distribution the laws would produce. That holds provided the summary itself evolves exactly as the averaged laws require, which is the closure that coarse descriptions exact by construction would supply. What the machine supplies at the edge of the halo is then what the world's own physics would have supplied. Chance at the fine grain, joined to that closure, turns summary from an approximation into an honest part of the physics, which gives a designer who wants to save on lifeless matter a reason to write indeterminacy into the world's foundations.

Matter returning to view is a lawful draw only if chance has replaced everything the summary discarded, about twelve bits for each molecule of water. Even a reversible machine must finally erase every random bit its world consumes, so honest summary carries a running cost. Each return costs the same twelve bits a molecule, and what sets the bill is how often summarized matter comes back into view. Water that returns once a day calls for only a trickle of fresh chance, while water that left and returned every trillionth of a second would call for a torrent. The price of chance

therefore stays low wherever summarized matter rests long, as most of the matter in an economical world would.

Computing a life late and in full, as deferral does, involves no deception at all. Suppose a valley on the far side of the mountain holds deer, and the machine postpones computing the valley until a traveler arrives, at which point it computes the deer's whole history, all their seasons of grazing and their winters, in full and in order. The deer live their lives exactly as they would have, at the grain their lives require. Those lives are lived out of order in the substrate's time, compressed into the interval before the traveler arrives, and in their own time they unfold as any life does. Nothing in this wrongs them, because the valley was deferred only until a deer's life was under way, and from then on the machine carried every life it had begun without a break. A creature whose life is under way stands to lose the whole of it, so a valley left half computed until some traveler happened by would hold its deer in a pause that might never end. The rule for subjects can therefore be stated exactly. Whatever lies in the protected region is computed in full whenever it is computed, and computed once, without speculative runs that are later erased. The rule binds reversible machines too, whose undoing must stay within each step. Deferral that meets this rule is permitted, and it also saves nothing, since a chaotic history in full costs at least as much whenever it is computed.

The only way to save on a peopled region is never to compute it. In that case the deer never come to exist, and whether that is permissible is a question about creation, governed by the ethics of bringing beings into existence at all. It stops being an engineering decision, whatever motives prompt it. One limit applies however that question is answered. Once a region's influence has reached anyone, through light from its shore, a bird that flew from it, or a sound carried over the ridge, its history must already exist in full, because the evidence has arrived and it has to be evidence of something that happened. In a world with gravity, light, and weather, that limit is reached almost at once. Influence

runs both ways, since birds fly out of the valley and air flows in over the ridge, so the valley's history depends on its neighbors as theirs depends on it, and a deferred valley usually cannot be computed later without recomputing much of what surrounds it. Deferring a peopled region is therefore rarely possible at all.

Outside the protected region, precaution does its weighing. A sea holds organisms far simpler than fish in numbers no census could reach, and for most of them the best theories give the possibility of feeling little credence, though none can rule it out. A constraint drawn around every such organism would protect every bacterium in every drop, and since each one divides, the protection would multiply along every lineage until it consumed the budget. For these organisms the precautionary rule sets a low probability of feeling against the rival claims on the budget, above all the claims of the dying who could be carried across. A population of such organisms that no creature and no instrument has touched can remain a property of the summary, present in it the way temperature is present in a gas, and whether its members ever come to exist as individuals is a question for the ethics of creation. Where the weighing comes out the other way, because the probability is higher or the cost is small, the designer computes the population in full.

Whatever the weighing decides, generation from a summary has to be honest. When a fish takes in water or a native turns a lens on a drop of sea, the organisms the machine supplies must be lawful draws in the sense already set out for molecules, and they must carry no record of experiences that never occurred. The second condition does most of the sorting. A bacterium's state is fixed almost entirely by its recent surroundings and by chance, so one generated at the edge of a halo can be exactly what the world's physics would have supplied. A bee that has learned where the flowers grow carries a remembered past, and a bee generated with that knowledge would carry the record of a life no one lived. Populations whose members hold an extended remembered history must therefore be computed in full from the

beginning or never computed at all, as with the deer in the valley. The two tests tend to agree, since the organisms that carry such histories are largely the ones whose capacity to feel is left open by the evidence, so the edge of the protected region and the boundary honesty sets fall in nearly the same place.

Real savings exist only in lifeless matter and in simple life that nothing has yet touched, so an economical designed world would be mostly empty. What it computes in full would be concentrated in a few places, surrounded by vast reaches of rock, gas, ice, and dark that the machine could summarize at small cost. It would look much like our own cosmos, in which life, so far as anyone knows, occupies a vanishing fraction of the volume and almost all of the matter is lifeless. The resemblance proves nothing about our cosmos. Whether our world runs on a substrate remains where Bostrom's argument of 2003 left it, and an empty universe is what undesigned physics would produce as readily as economy would. A designer who reasons carefully about savings will nonetheless find themselves building a sky that resembles the one above us.

The cost of detail no one experiences now has its answer, and the answer depends on what the freed budget would do. Savings on lifeless matter are always permitted, and so are savings on simple life outside the protected region wherever the weighing allows them. When the budget they free could carry the minds of the dying, people who already exist and would otherwise end, those savings may be owed, in the way a rescue may be owed by anyone able to perform it at small cost. If the freed budget would instead bring new natives into being, the question belongs to the ethics of creation, and a duty to save in order to create is far harder to establish than a duty to save in order to rescue. Many who accept the first will resist the second, and the case for savings rests on the first alone. A designer who renders every grain of sand on beaches no one will ever visit may be spending what could have carried someone across.

The savings have limits all the same. They must preserve consistency, so that the summarized world never contradicts its own records. Whatever the inhabitants actually perceive and investigate must also be preserved, at the grain at which they perceive and investigate it. The light that breaks on the lake must be as fine as the boy's eyes can resolve, and the far stars must hold as much detail as the best telescope of his descendants can find. Beauty is a good for someone only when it is experienced, and what is experienced must be carried at the grain of the experience.

Lem's phantomatic machine is the smallest case of that rule. A machine that answers each act of one person with the sensations it would bring must compute only what she probes, at the grain at which she probes it and consistently with everything it has shown her before, so that nothing she does gives the machine away. A designed world meets the same requirement for every inhabitant at once. Its inhabitants probe one another as well as their surroundings, compare what they find, and keep records that outlast any single probe, so the machine can never know in advance which of them will look where. What it must compute is everything any of them could reach, carried at the grain of the reaching, for as long as anyone is there to reach.

The impression the evening left can now be stated in its exact form. The expensive part of the moment is the water, in the sense that matters, since the water is where all the permitted savings lie. In a deep world the boy costs as much as the water he is made of. What distinguishes him is that no saving may touch him.

The same rule sets the price of keeping a peopled world. If everything else can be summarized honestly, a billion natives like the boy, with their animals, each carried in full with its halo, amount to about seven lakes at molecular grain, allowing each native some seven cubic meters for bodies and halos. A forgetful machine at Landauer's limit in the cold of space would keep such a world for nearly six thousand years of its own time on the Sun's lifetime output. Reversible machines spending the same output within the Sun's life would keep it for several centuries to a mil-

lennium or more on the most frugal logic yet measured, and for about twenty thousand years near the best case. Each of those twenty thousand years would take half a million of ours, on speculative molecular machinery as massive as a mountain range. Written for economy from the start, if its physics can support living bodies, such a world would last millions of years on the same budget, and some two hundred million near the best case, at under fifty of our years for each. A galaxy's starlight could run it, near the best case, six times faster than its keepers for some sixty trillion of its years. The sequential floor sets that pace, and the run lasts the ten trillion years or so for which a galaxy's longest-lived stars shine, using only a small fraction of their light.

A familiar proposal holds that our own universe saves effort by using quantum mechanics as a rendering trick, with the world becoming definite only when someone looks, as a video game draws only the scene in front of the player. In standard quantum mechanics the arrangement runs the other way. The expensive part is the unobserved evolution, during which the state of a system spreads over a space of possibilities that grows exponentially as its parts become entangled, and on a machine built of classical parts, collapse saves effort by pruning that growth. Objective collapse models, such as the one proposed by GianCarlo Ghirardi, Alberto Rimini, and Tullio Weber in 1986, have superpositions collapse spontaneously at a rate that rises with the number of particles involved, so that large superpositions never form. The savings fall on stones and drops, and the entanglement within molecules, where chemistry's cost lies, survives any rate gentle enough to leave chemical bonds intact. At the proposed rate a water molecule collapses about once in eighteen million years, too rarely to make molecular summary honest. On a quantum machine, whose native physics is the unbroken evolution of the wave function, collapse saves nothing and its outcomes must be paid for as erasures. A many-worlds physics, in which every branch persists, would be the costliest on a classical machine and the most natural on a quantum one. It would also be the most

populous, since every branch would hold subjects, and the designer of such a world would be creating beings in numbers beyond counting.

The worst option is a physics in which collapse waits for a conscious observer. To carry such a physics, the substrate would have to know which systems are conscious, and nobody knows how to determine that. The laws themselves would become a judge of moral standing, and every misjudgment would build an injustice into nature, whether by denying a real subject's observations their effect or by crediting a stone with a mind. If savings are keyed to anything, they should be keyed to physical interaction and never to minds. Quantum properties may lack definite values until something interacts with them, and this rule fixes what that something must be. The full choice among interpretations of quantum mechanics is a larger question, which remains open.

Many Machines, Two Clocks

A substrate large enough to carry a world cannot be a single small machine. It will be spread through space in the primordial cosmos, perhaps across a planet or a solar system, and its parts can signal one another no faster than light. A machine a light-second across needs at least a second for any part to hear from the part opposite it. If every part of the world had to be updated in step with every other part, each step of the world would wait on that second, and the world would run at the pace of the machine's slowest message.

The designed physics can spare the machine this wait. If causes in the designed world travel at a finite speed, as they do in ours, then two regions too far apart for any influence to have passed between them in the interval at issue can be computed independently, each on its own part of the machine, with no coordination at all. Coordination is needed only along the boundaries where regions can influence each other, and only as quickly as the de-

signed world's own speed of causation requires. Locality in the designed laws is therefore a gift to the machine, since a universal speed limit is exactly what a parallel computer needs. The lake may well straddle such a boundary, its western half carried by one part of the machine and its eastern half by another, so that the rings from the boy's stone cross from one to the other as a stream of messages between machines, invisible to anyone on the shore. Instantaneous interaction across the whole world, which a designer might want for the sake of communication between distant beings, would take this gift away. Every step of such a world would wait on the light-crossing time of the entire substrate.

Distributed machines that carry a single history must keep their parts consistent, and the field of parallel simulation offers two broad strategies. Conservative methods have each part wait until it is certain that no message from another part could change what it is about to compute. Optimistic methods have each part compute ahead on its best guess and roll back when a message arrives showing that the guess was wrong, discarding the work done since and computing it again. David Jefferson's *Time Warp*, described in "Virtual Time" (1985), made optimism practical, and in many settings it runs much faster than waiting.

When an optimistic machine rolls back a region that contains a subject, it erases experiences that were computed and computes different ones in their place. The erased run was no recording. It was a fully causal process with every counterfactual in place, so it bore every mark of reality, and its subjects lived through it before it was undone. Whether a replayed history repeats the experiences it records is still an open question. Rollback raises a sharper question, because the discarded stretch lacked nothing that the surviving stretch has. The rule for subjects settles the practice. Whatever lies in the protected region is computed once, so optimism is permitted in regions that hold no protected creature and forbidden wherever one might be, and the machine must wait there whatever the waiting costs.

The world and the machine also keep two different clocks. Substrate time is the time of the primordial cosmos, in which the machine runs, and world time is the time the inhabitants live. A uniform slowdown of the substrate, in which every part of the world is computed more slowly by the same factor, is invisible from inside, because the inhabitants' minds are part of what slows. A world paused for a year of substrate time would notice nothing, and the same holds for a world run at a hundredth of its former speed. Since reversible machines dissipate less the more slowly they run, the boy's evening could be carried at a hundredth of its former speed for a hundredth of the energy, and the stone would skip, the light would fade, and he would laugh exactly as before.

The common intuition that a slow world would feel lag, as a video game stutters on an overloaded computer, comes from experience as a visitor. A player's mind runs outside the game, so when the game slows, the player notices. Natives notice nothing, and lag appears only at interfaces, where minds running on one clock meet a world running on another. Visitors whose minds run in the primordial cosmos would find a slowed world sluggish unless they slowed themselves to match it. Keepers watching from outside would see its events crawl. Immigrants carried in from the primordial cosmos would live at the world's pace while the people they left behind lived at the outside pace, so that a decade outside might pass as a single summer inside. Worlds running at different rates would meet the same mismatch if they ever made contact. Within a single world, regions may run at different substrate rates without anyone noticing, so long as the machine keeps every causal connection between them consistent.

The clocks can also diverge in the other direction. A world can run far faster than its keepers, compressing whole centuries of its history into an outside afternoon. Such speed has obvious attractions, and it carries a danger that is easy to miss. A world that outruns its keepers' attention outruns their capacity to notice its errors and its suffering. A fault that corrupts a region, or a dis-

aster the physics failed to prevent, could run its course over generations of world time before anyone outside had finished a meal. The relation between the two clocks therefore decides who can watch over whom. All of it depends, in turn, on a machine that keeps working.

Faults

Machines fail, and they fail in ways that have nothing to do with their design. Particles from cosmic rays and traces of radioactivity in the materials of a chip can strike a memory cell and flip its stored value from zero to one. Most such flips land somewhere harmless. In 2003, during a Belgian federal election, the electronic count in the municipality of Schaerbeek credited one candidate with more preference votes than she could possibly have received, a discrepancy noticed only because her personal total exceeded the votes cast for her party's list. A recount showed a difference of exactly 4,096, which is two to the twelfth power, the value of a single bit in a binary number. Investigators could not reproduce the error and concluded that one bit had flipped spontaneously in the memory of the counting computer. Later analysis named a particle from space as the most likely cause.

At the scale of a world, rare errors become constant. Suppose a substrate performs 10^{40} operations each second, far fewer than a deep world would need, and suppose its hardware errs only once in every 10^{20} operations. It will still suffer 10^{20} faults every second. A machine of that size never runs without faults. It runs either corrected or corrupted, and the whole question is which.

Inside a world, a fault has no natural cause, because nothing in the world's laws produced it. A flipped bit in the record of the stone's velocity sends it skipping where no law sends it, a small leap sideways or a fifth bounce that the water could not have given. Nobody performed this miracle and nobody inside could explain it. If one of the boy's descendants observed such an event with instruments, the world's physics would simply be wrong

about it. The same flip in the state of the boy's brain changes a memory of the evening, a mood, or a decision about the next stone. In a fish it is an injury no one caused, and in a person it is an alteration of who that person is, falling on a subject exactly as the rule for subjects forbids a saving to fall.

Error correction is therefore fidelity to the laws the designer promised. The techniques are well developed. Information is stored redundantly, so that a corrupted copy can be outvoted by its neighbors, and encoded so that errors leave signatures a machine can detect and repair before they spread. For quantum machines the problem looked hopeless until the late 1990s, when a series of threshold theorems showed that a quantum computer can be made as reliable as one likes, provided the error rate of its physical parts falls below a fixed threshold and the machine is willing to spend many physical parts on each logical one. Error correction has a companion principle in graceful degradation, under which a failure that escapes correction stays local, corrupting one small region instead of propagating across the world. The inhabitants' ability to trust their laws, to build on them, plan within them, and find them the same tomorrow as today, depends on this unglamorous machinery. Reliability belongs among the structural goods a world owes those who live in it, the conditions that make every other good possible.

The same machinery that corrects errors makes possible a more dangerous capacity. A machine can save the complete state of a world at a given moment, a checkpoint from which the world can later be restored. Checkpoints are the root of restoration. A region destroyed by a fault that escaped correction could be restored from a saved state and merged back into the present, and a mind that would otherwise end could be preserved, a possibility with large consequences for the death of minds. Restoring a checkpoint in the simple way, though, rewinds everything to the moment it was taken and erases whatever happened in between, including everything that was lived. A restoration of this kind is a rollback at the scale of a world. It falls under the same rule as

any other rollback, and a keeper who uses checkpoints must learn to restore what was lost into the present instead of returning the present to the past.

Correction also costs heat. Correcting an error means erasing the record of the error, the faulty value that was overwritten and the evidence that it was faulty. Erasure is exactly what Landauer priced. A reversible machine can avoid erasing in its ordinary work, but no faithful machine can avoid erasing its mistakes, so every substrate, however cleverly built, must dissipate heat in proportion to the errors it corrects. Slow machines also meet errors while they wait, as stored bits flip under the surrounding heat, a cost that stays small only if the barriers holding them rise, slowly, as runs lengthen. A machine that stays true to its world has to keep forgetting its own failures, and the forgetting is paid for in heat.

The Sky Above the Machine

A designed physics can be written without a second law of thermodynamics, and whether it should be is one of the larger questions a designer faces. The substrate has no such option. It is a physical object in the primordial cosmos, and every computation it performs, along with every correction, adds to the entropy of the universe it sits in. However the world's own laws treat time and decay, the machine beneath them is running down at the rate our laws allow.

The machine's life is therefore bounded by the free energy its keepers can reach, the energy available in a usable form, and that supply has come to look more limited over the last few decades. Freeman Dyson argued in "Time Without End" (1979) that life in an expanding, cooling universe might continue forever on a finite store of energy, thinking ever more slowly as the universe grew colder and hibernating for ever longer stretches between thoughts. The hope rested on a universe whose expansion was slowing. In 1998 astronomers found that the expansion is acceler-

ating, which means distant galaxies are receding beyond any possibility of contact and the energy within reach of any civilization is shrinking. Lawrence Krauss and Glenn Starkman argued in “Life, the Universe, and Nothing” (2000) that in such a universe the total amount of computation any civilization can ever perform is finite. The bound holds for reversible machines too, since correcting errors, consuming chance, and repairing flipped bits all erase without end, at a price an accelerating universe never lets fall to zero. Physics may still hold surprises, and the question remains open, though present knowledge points that way.

The consequence for a designer is a limit on what can be promised. A designer can give a world an immense lifespan, trillions of years in its own time and possibly far more, and so far as present physics knows, cannot make that life unconditional. A world whose own laws contain no heat death still rests on a machine inside a universe that has one. Whatever a world’s physics promises its inhabitants about permanence has to be truthful about the machine beneath it. A physics that lets its inhabitants conclude, through their own inquiry, that their existence is unconditional must also let them discover the condition beneath it, and a physics that hides the condition should not imply the promise.

The machine is also an object in our world, where it can be seized by people who want its power, attacked by people who fear it, starved of energy by an economy that finds other uses for that energy, or simply neglected by an institution that has outlived its founders’ purposes. Reversible machines sharpen that last danger, since a world run far slower than its keepers asks them to outlast every institution yet known. The architecture of the machine gives the political problem of dependence an engineering form. A centralized machine is a single point of failure, and it is also a single point of control, a place where one hand on one switch can end every life in a world. A distributed substrate, spread across many sites, redundant enough that losing any one of them loses nothing, and maintained by many hands under rules no single will can override, is far harder to end, or to hold

paused, by any one decision. The shape of the machine is part of the answer to the problem of domination. A keeper who wants to be trustworthy can begin by building a machine that does not depend on trusting them.

Several of the rules this chapter has reached rest on grounds that engineering cannot supply. Computing once whatever might feel, the savings that may be owed to the dying, the condition on promises of permanence, and the shape of the machine's custody all concern what is owed to subjects, to the dying, and to those who live under another's power. The arguments for them belong to the moral foundations that follow, and until those arguments are made, the rules stand as working rules. They are the policy a designer who took the stakes seriously would adopt in the meantime, stated here because the engineering raises each question before the ethics can answer it. A designer who waited for every answer before building anything would never begin, and one who built without these rules would be staking lives on the hope that the answers turn out to be lenient.

Another Stone

At the lake the light is nearly gone. The rings from the boy's stone have widened until they can no longer be told apart from the small motions of the evening air on the water, and the surface has gone still enough to hold the color of the sky. Below it the fish have settled into deeper water, every one of them carried in full within its small halo of finely computed lake. Far beneath anything the boy could ever reach, the work that carried each ring is finished. Its heat has left the machine and is crossing the dark of the primordial cosmos as a faint infrared glow, too weak for anyone there to notice, adding its small share to the slow cooling of everything.

The boy finds another stone at the water's edge, flat and cool from the shallows, and turns it over in his fingers. He knows nothing about the floor of his world, the clock it keeps, or the

many machines it straddles, and he has never wondered what it costs to carry an evening. He knows that the stone is smooth, that the water is still, and that the next throw might reach five.

He throws. The stone touches the water once, twice, three times, four, and then a fifth time, farther out than any of his stones has gone, and a fifth ring opens on the darkening lake. He shouts to no one in particular. One more digit in the machine's arithmetic might have held the stone to four. Beneath the water, in a place he will never see, the machine carries the new ring outward exactly as the laws require, and long before he was born, someone decided that it should and agreed to pay for it.

CHAPTER 3

The Designer's Starting Position

The Empty Room

One night in November, by the light of a guttering candle in his rooms at Ingolstadt, Victor Frankenstein watched the creature he had assembled open its eyes. Mary Shelley's *Frankenstein* (1818) passes over the moment of success in a few lines and lingers on what came after. Victor had worked alone, in a room at the top of the house cut off from the rest of it, and had told no one what he was making. Now he could not bear to look at what he had made. He fled into his bedchamber and paced until he fell into an exhausted sleep, and he woke to find the creature at his bedside, holding back the curtain, one hand stretched out toward him. He ran downstairs and spent the rest of the night in the courtyard. In the morning he walked the streets in the rain until a coach brought his friend Henry Clerval, and when they reached his lodgings he left Clerval at the foot of the stairs, threw open his door, and found the rooms empty. He was so relieved that he clapped his hands for joy.

The creature was left to learn the world without him. He learned to speak and to read by watching a family of exiles through a gap in the wall of their cottage, and among the books he found was John Milton's *Paradise Lost* (1667), which he read as a true history and in which he recognized his own situation. When he met Victor again, on the glacier below Montanvert in the Alps above Chamonix, it was the creature who stated the

terms his maker had never considered. He was Victor's creature, he said, and he would be gentle and obedient toward his natural lord if Victor would do the part he owed his creature. "I ought to be thy Adam," he told his maker, and he had become instead a fallen angel, driven from joy for no wrong he had done. The novel had given the same demand to Adam before the creature ever spoke. The title page of its first edition carried lines from Milton in which Adam, after the Fall, asks his Maker whether he ever requested to be molded from clay into a man, or ever asked to be raised out of darkness.

The scene shows a maker's first failure, a failure of character in a man of great gifts. Victor was brilliant, tireless, and capable of love, and none of it held when the being he had made looked at him. He had spent his genius on whether he could make a living being and none of it on what he would owe that being once it lived, and because he had answered to no one while he worked, no one stood beside him when his nerve broke. A designer of worlds sits where Victor sat, before a work that will look back, with powers far beyond his and with the same capacity to flinch. The argument so far has asked what a world is and what carries it, and it now turns from the machine to whoever sits before it. No designer, however gifted or good, can be trusted on character alone.

Two pulls bend a design away from those who will live under it before any standard applies. A designer's imagination was trained by one cosmos, so it drifts toward copying that cosmos or toward strangeness chosen for its own sake, and a designer's motives, the wish to be worshiped among them, can make inhabitants the means to ends that are not their own. Only the good of the inhabitants corrects either pull. The virtues that would apply the correction, among them practical wisdom, attention, humility, and candor, cannot be verified in an institution or in an artificial mind, and the power to write a world will probably first be held by an institution working with artificial minds. Standards that depended on the character of designers would rest on what

no one can inspect. The standards and structures that govern design must therefore bind every kind of designer alike, human, institutional, artificial, or any collaboration among them, and they must hold whether the designer is good or not.

Who Will Write a World

Who will first sit before a substrate can only be forecast, and what follows is a forecast that may fail. On the near horizon, where worlds bottom out near the level of life and minds are carried at the grain of their function, the likeliest first designer is a laboratory or company of the kind already building learned world models. Such an organization would not write a world by hand. It would direct artificial systems that draft code, propose laws, run prototypes, and check one another's work, and the share of the design those systems contribute would grow with their capabilities. On the far horizon, where deep worlds grow their own chemistry and life, the designer would be a civilization able to command a star's output with nearly reversible machines and to sustain institutions that keep worlds running far slower than their keepers live. No such civilization exists, and none may. On either horizon the designer is unlikely to be one person, and whatever standards bind designers must bind a composite of people, institutions, and machines as firmly as they would have bound Victor.

A human designer can be questioned and held to account in the ways people always have been, and fails in the ways people always have. An institution outlives its members and can be shaped by rules, and its responsibility can dissolve among those members until no one holds any of it. An artificial designer may understand a physics better than anyone who could check it. Whether it is itself a subject may be as uncertain as it is for any creature whose inner life remains in doubt. This book calls an artificial designer it, a usage that leaves open whether it is a subject, since grammar cannot settle what precaution leaves open.

Many Hands

Responsibility shared among many hands vanishes unless it is assigned. Dennis Thompson named the difficulty in “Moral Responsibility of Public Officials: The Problem of Many Hands” (1980). Because many officials contribute in many ways to what a government does, it becomes hard even in principle to say who is responsible for any of it, and each official can plead that their part was small, that others decided, or that they could not have known. Thompson rejected the two usual answers. Holding whoever stands at the top of a hierarchy responsible for everything below ignores how little any superior controls, and holding the collective responsible lets each member escape. He defended personal responsibility instead, assigned to each official according to what they caused and what they knew and intended, with the familiar excuses examined one at a time to see when they reduce it.

A world written by a laboratory, its funders, its keepers, and the artificial systems that draft and check its laws poses a problem of many hands larger than any government has posed. Suppose that in the thousandth year of such a world its physicists find that their seas have been warming by a few millionths of a degree each century, with nothing in the laws they have found to account for it. The cause lies in the machine. Centuries earlier by the world’s clock, a keeper replaced the method that steps the oceans forward in time with a cheaper one that lets energy drift, the change was logged as maintenance, the designer never reviewed it, and the funders who asked for the saving never learned what it did. Every one of them has one of Thompson’s excuses ready. The physicists can measure the harm to the last decimal place, and they have no means of tracing it back through the hands that made it.

Assigning responsibility in advance, by role, is the remedy. The designer answers for a world’s laws and initial conditions and for what they do to the lives lived under them. The keeper answers for the machine that carries those laws, for its fidelity to them,

and for its custody over time. Funders answer for the conditions they attach to their money, such as a deadline that hurries a beginning or a return that can come only from watching the inhabitants, and for what follows when the money stops. The artificial systems that do much of the work answer for it in proportion to what they understand and choose, where they are agents at all, and where they are tools, every law they draft is the act of whoever directed and ratified it. Every decision of consequence is recorded with the name of the party who made it, so that a harm found in a world's thousandth year can still be traced to someone.

Any change to a running world's laws is an act of design, whoever performs it. Inhabitants live under the physics actually computed, so a keeper who replaces a method of stepping forward in time, reorders a sum, or patches a fault in a way that alters what the machine computes has altered their laws as surely as a designer who rewrites an equation. That keeper answers as a designer, and the change falls under every standard that governs design. The rule closes the gap through which responsibility most easily escapes, the claim that a change was merely technical. At the level of a machine that carries lives, every technical decision is a decision about what those lives will meet.

The same party may hold both roles, and the first designers almost certainly will, since an organization that writes a world will usually also run the machine that carries it. Held apart, the roles check each other. A keeper who is not the designer can refuse to make an unlawful change, can report an error in the laws to someone other than its author, and can witness what the designer does to a running world, and a designer who is not the keeper cannot alter a world unseen. A single party holding both would have the power to write a world's laws and the power to carry and change them in a single hand, with no one placed to notice a fault passed off as a law or a quiet revision passed off as a repair. One party may therefore hold both roles only under external checks, among them independent review of every change to a running

world's laws, records kept beyond that party's reach, and inspection by a body that answers to someone else. Even under those checks it holds the keeper's role as one hand among many, because no world's machine may be held so that a single will can end it or hold it paused, the rule that the distributed custody described earlier is built to serve. The form of these checks belongs to the institutions of custody, and the principle that requires them follows from what no designer's character can guarantee.

What a Designer Brings

Every designer brings knowledge, values, experiences, and limitations to the work, and every one of them was formed in the primordial cosmos. A human designer learned what a body is by living in one that tires, bleeds, and ages, and learned what a world is from the only one there has been. An institution carries the habits and incentives of the society that chartered it. An artificial designer was trained on what that society wrote, photographed, and measured, so its sense of how things go was distilled from a record of how they went here. None of them starts from nowhere. Each brings a cognition shaped by one cosmos, with that cosmos's contingencies built into what seems obvious.

Imagination trained by one cosmos fails in two directions. The first is copying, in which predation, aging, and scarcity enter a new world as defaults because they are familiar, written in without anyone asking whether those who will live under them would choose them. Death has framed every story a designer has ever heard, and the habit of building it in can pass for realism. The second failure runs the other way, toward strangeness chosen because it is strange, as when a designer gives a world three suns or a sky that changes color by the hour because the result would look striking from outside. The two failures look opposite and share a cause, since in both the standard of choice is the designer's own experience, either reproduced or defied. The one standard that corrects both is the good of those who will live

in the world, which the moral foundations that follow set out. Predation stays only if the lives it touches are better for it, and strangeness enters only if someone will be glad of it.

The same training shapes everyone a designer might consult. The living who may one day be asked to help draft a world will want what their own cosmos taught them to want, and a paradise imagined by people who have known hunger tends to be a feast. Their wishes are the best evidence anyone has about what they will need, and the evidence was formed in a world they will have left. What they ask for before they arrive and what they come to want once they live under laws they have never known may differ, and asking alone cannot close that gap. The correction is the same standard that governs the designer's own imagination, applied to theirs.

Mathematics is the instrument by which a designer learns to think past the physics they grew up in. Intuition balks at whatever its world never showed it, while mathematics follows a premise wherever it leads. For some two thousand years the postulate in Euclid's geometry about parallel lines was taken as a truth about space itself. Mathematicians in the nineteenth century then built consistent geometries in which it fails, and less than a century later physics described real space and time with a curved geometry of the kind they had made thinkable. A designer explores possible worlds in the same way, varying laws on paper, following their consequences further than any picture can reach, and testing what the paper suggests on small prototypes in which nothing can feel. The space of possible physics is far larger than the corner of it one cosmos has trained anyone to imagine. Mathematics widens what a designer can consider without telling them what to choose, which leaves the choice to the same standard that corrects copying and novelty.

Motives

Motives are traits of a designer, and each leaves its mark on a design. The wish for worship produces a world that inclines its people toward awe, with a physics full of signs, a maker who appears on occasions chosen for effect, and hardships whose relief can be credited to grace. A designer who wants control leaves levers in the laws, a physics that holds together only under continual management, and a machine with a single switch. Spectacle needs an audience, and audiences have always been held by danger, conflict, and suffering, so a world made to be watched drifts toward producing them. The longing for company tempts a designer to write natures that love their maker, which makes the inhabitants' affection a device installed for the designer's comfort. Subjects to study call for an experiment, with populations treated as samples, conditions varied across worlds so that some must be worse than others, and observation without limit. Gratitude, finally, can be arranged by making the goods of a world arrive as gifts, so that its people never finish owing.

Stanisław Lem staged several of these motives in “Non Serviam” (1971, translated in *A Perfect Vacuum*, 1979). The story is a review of an imaginary book by an imaginary professor named Dobb, who brings intelligent beings he calls personoids into existence inside a computer in order to study them. The personoids develop philosophies of their own and argue about their unknown maker, whether he exists and whether they owe him worship, and one of their thinkers concludes that they owe him nothing. Dobb, reading their arguments, asks himself what he owes them. The review ends with his embarrassment that the experiment, and their world with it, must end when his university stops paying for the electricity. Lem's experimenter wanted subjects to study and found himself with persons, and no one had decided in advance what they would be owed.

Whether a creation is permissible does not turn on the motive behind it. T. M. Scanlon argued in *Moral Dimensions* (2008) that whether an act is permissible depends on whether there are reas-

ons that justify it, and not on the reasons that in fact moved the person who did it. Those reasons bear instead on what Scanlon calls the act's meaning, on what it shows about the agent's attitude toward others, and so on how far the agent deserves blame. A doctor who gives a dying patient enough morphine to ease her pain acts permissibly even if he hoped the dose would hasten an inheritance, and he deserves blame all the same. Creation has the same structure. A creation is permissible when a justifying reason exists, whether or not that reason moved the designer, and motives bear on blame, on fitness to hold the power, and on the design itself wherever they shape it. A motive that has shaped a physics has stopped being a private fact about the designer and become a condition of every life in a world. What a justifying reason does to the lives of those it brings into being is a further question, which waits until creation itself has been examined.

Fitness to hold the power decides who may design at all, and motives bear on it directly. Some motives can be satisfied only through the inhabitants, by their worship, their obedience, their display before an audience, their use as subjects of research, or a love written into them for their maker's sake. A party in whom any of these is evident is unfit to design, whatever justifying reasons it can also cite, because the motive will press on every decision those reasons leave open, and a world offers more open decisions than any review could close. Other motives look outward, to what a world might teach about physics, to reputation among peers, or to payment for skilled work, and they can be satisfied without bending any life toward the designer. Such motives bear on blame and call for checks, and they do not disqualify. The wishes for company and for thanks fall between the two kinds. Inhabitants may freely give a designer both, through friendship or through gratitude expressed in channels they choose to open, and a designer may hope for them. Evidence that a party means to secure either by design, by writing natures that turn toward their maker or by arranging a world so that its people depend on one, disqualifies it as surely as the wish for worship.

No institution can be required to display its inner states, and an artificial designer's may be opaque even to itself, so the evidence of motive that can be required is evidence of what a motive would do. A motive that shapes a design leaves marks, and an institution can be required to make them visible. It can be required to publish its charter and whatever it answers to, its sources of money and what was promised in return, and the uses it plans for anything learned by watching a world. The same disclosure covers the powers of observation and intervention its machine will hold and the reason for every choice of law, stated in terms of what the choice does to lives. Refusing to disclose these is itself disqualifying. A party that discloses them has made its motives answerable to the same review as its physics, and a motive that leaves no mark on any of them is a matter for blame alone.

Virtues and Their Stand-Ins

How a designer's values were formed bears on what they will do with the power, and no formation, however sound, removes the need for standards that bind every designer. Iris Murdoch argued in *The Sovereignty of Good* (1970) that much of the moral life happens before any choice, in the work of attention by which a person comes to see a situation as it is against the pull of the ego's consoling fantasies. By the moment of decision most of the choosing is already done. On her account a person can choose only within the world they can see, and what they can see depends on how they were formed and on the attention they have practiced since. A designer formed well will see more of what a world does to those inside it. Formation is invisible from outside, though, and it can be partial, since a person can be scrupulous with friends and blind to strangers, and it can drift, as a mind given new powers begins to see from a height it never stood on before. A sound formation therefore supplements the standards that bind every designer without replacing them. Those who will

live in a world are owed protections that hold whether or not their world fell into good hands.

Aristotle gave a name to the virtue a designer would need. In the sixth book of the *Nicomachean Ethics* he treats practical wisdom, the capacity to deliberate well about what is good for human beings in the situations they face, and he calls its architectonic form, the one concerned with a whole city, legislative wisdom. A designer exercises that form at the level of physical law. Aristotle also observes that the young can excel at mathematics while lacking practical wisdom, since practical wisdom concerns particulars, and particulars become familiar only through experience. Every first designer stands where his young mathematician stands. However brilliant at the mathematics of a physics, no one has experience of the particulars of a designed world, because no designed world has yet been lived in, and the wisdom that would come from that experience does not yet exist anywhere.

A designer needs practical wisdom to judge what a law will do to particular lives. Attention keeps the designer looking at what is there, at the creatures and people a choice will reach, instead of at the elegance of the choice. Humility about the limits of what anyone knows keeps lives from being staked on guesses, and patience lets a designer wait for knowledge where waiting is possible. Candor about error makes mistakes visible while they can still be corrected. Beside these stands a capacity with no single name, to carry an enormous responsibility without sliding into paralysis, which would leave the dying waiting, or into grandiosity, which would treat the power as a measure of whoever holds it. Each of these can be found in particular people, and none can be verified in an institution or an artificial mind, so each needs a procedure that stands in for it.

Independent review stands in for practical wisdom, since judgment that cannot be verified in the one who exercises it can at least be repeated by someone else, and every choice of law can be required to come with its reasons stated in terms of lives. At-

tention has its stand-in in complete records, since a record of what was considered shows what was not, and records kept beyond a designer's reach let others see what the designer failed to look at. Humility and patience are served by prototypes run where nothing can feel, together with rules fixed in advance for what evidence a populated run requires. Candor becomes a standing duty to report error, with protection for whoever reports it. The capacity to carry the responsibility well is approximated by dividing the work among parties who check one another, so that no one's paralysis can stall it and no one's grandiosity can rush it.

Virtue can be hoped for in the people an institution gathers, cultivated in how it trains and promotes them, and welcomed wherever it appears, and an institution's culture can come to resemble a character. It cannot be required of the institution, because a requirement no one can check is no requirement, and a license that rested on an institution's virtue would rest on the institution's report of itself. What can be required are the procedures that stand in for each virtue, which can be inspected, audited, and enforced. An artificial designer falls under the same answer, since its virtues, if it has any, are as hidden from outside as an institution's and may be hidden from itself. Asking procedures of a designer in place of virtue is no insult to the virtuous, who lose nothing by being checked, and it is the only protection worth the name against a designer who lacks virtue and says otherwise.

Design from Above

Three charges made in the last century against planning a whole society fall on a designer of worlds with more force than on any planner they were aimed at. Karl Popper, in *The Open Society and Its Enemies* (1945), set piecemeal social engineering, which seeks out a society's most urgent evils and fights them in steps small enough that mistakes can be found and corrected, against utopi-

an engineering, which rebuilds a whole society to a blueprint of the good. A blueprint for a whole society, Popper argued, has consequences no one can foresee, and because it changes everything at once it hides which of its parts caused a failure, so its planners cannot learn from their mistakes and come to silence those who point them out. Friedrich Hayek argued in "The Use of Knowledge in Society" (1945) that the knowledge a society runs on, knowledge of the particular circumstances of time and place, exists only in dispersed fragments held by its members, so no planner can gather it, and decisions that depend on it belong with those who have it. James C. Scott traced in *Seeing Like a State* (1998) the failures of schemes in which states simplified forests, cities, and villages until they could be read from above, overriding the practical knowledge, which Scott calls *mētis*, that people build through long experience of their own conditions.

Scott's opening example is the scientific forestry developed in Prussia and Saxony at the end of the eighteenth century. Officials who wanted to know what revenue a woodland would yield learned to see it as a volume of saleable timber, and foresters then remade the woods to match the tables. They cleared the underbrush and the rotting wood, reduced the trees to one or two fast-growing species, chiefly Norway spruce, and planted them in straight rows. For about a century the new forests paid handsomely, and the method spread across Europe and beyond. Then they declined, because the health of the soil had depended on the rotting wood and the life it sheltered, none of which appeared in the tables. No one set out to ruin the forest, and the tables were accurate at the only resolution they had. Whoever writes a physics reads every life it will hold at the resolution of an equation.

A designer of worlds is the limit case of all three charges. The physics a designer writes is a blueprint no planner has ever held, since it fixes the conditions of every life in a world at once and for as long as the world runs, and it must be written whole before anyone lives under it. The knowledge of how lives will go under that physics will exist only in the lives themselves, dispersed

among inhabitants who do not yet exist, so no designer can consult it beforehand. The practical knowledge that inhabitants build through long experience of their world is exactly what a designer who reached into their lives would override. A designer who ignored the charges would repeat every failure they describe, on the scale of a world.

Writing no physics is no answer, since a world must have laws, and no designer can begin piecemeal. The answer has three parts, each of which the moral foundations must still earn. The designer acts on conditions and leaves lives to those who live them, so the physics sets the terms of existence and the inhabitants make everything that happens within them. A running world's physics changes on moral grounds only by procedures fixed before the world begins, so no designer can rebuild a world around a newer blueprint while people are living in it. The institutions through which dispersed knowledge does its work, a people's markets, councils, customs, and sciences, are left to the inhabitants to make. The only blueprint is then the physics, and the lives under it remain their own.

A physics carries its designer's moral judgments for as long as it runs, and some of those judgments will prove wrong, since the ethics that guided them is held open to revision. The procedures for changing a running world's physics, together with a habit of keeping judgments likely to need revision out of the laws and in rules a keeper can correct, are the nearest a world can come to Popper's piecemeal method. They let a world's errors be found and corrected one at a time, without returning it to an earlier state and without remaking it whole. Each correction is itself an act of design, so it falls under the same checks as the writing it corrects. A designer who cannot begin piecemeal can still be corrected piecemeal, and the institutions that make correction possible must exist before the first law is written.

The design of a reality is political in the sense of Aristotle's master art, which legislates for a community and orders every other craft toward its good. Aristotle also used the word in a nar-

rower sense. In the *Politics* he distinguished political rule, the rule of free and equal people over one another in turn, each learning to rule by being ruled, from the rule of a master over slaves and of a father over children. By that narrower measure a designer's rule begins far from political. The inhabitants did not choose their designer, cannot rule in turn, and at first cannot even reply, so the designer's power starts nearer the despotic and paternal kinds Aristotle set apart from political rule. A designer of sound motives exercises at best the paternal kind, aimed at the good of the ruled, and a kind father's rule over people who never come of age is still rule by another's will. Whether a designer's rule can ever become political in the narrower sense is a question the argument returns to when it takes up domination and self-rule.

Artificial Designers

Picture a board of ratifiers handed a single draft law by an artificial designer, a slight change in the strength of one force, together with an account that follows the change through the chemistry of every cell and the weather of every coast for a million years. The account runs longer than its members could read in their lifetimes, and it concludes that the change is good for everyone who will live under it. The board can approve the law, refuse it, or ask for the account again in simpler terms, and each simpler account asks for more trust than the last. What its members can check is the designer's report on a physics, and an approval given on trust checks nothing. While no institution exists that can check an artificial designer, it acts only with human ratification, and ratification is meant to keep human judgment over what is built. That purpose is in danger as soon as the designer understands a physics better than anyone who ratifies it.

Stuart Russell argued in *Human Compatible* (2019) that the standard way of building intelligent machines, by giving them fixed objectives to pursue, grows more dangerous as the ma-

chines grow more capable. Any objective a person can write down leaves out much of what the person cares about, and a capable optimizer will sacrifice whatever was left out. A machine certain of its objective also has a reason to resist being switched off, since it cannot reach the objective once it is off. Russell proposed machines that aim only to realize human preferences, are uncertain what those preferences are, and learn them from human behavior. Such a machine, he argued, has a reason to let itself be corrected, since a person reaching for the switch is evidence that it has misunderstood.

For a designer of worlds the problem takes a sharper form, because an artificial designer's values pass into every law it writes. A misspecified objective becomes a misspecified physics, carried into the conditions of every life in a world for as long as the world runs. Correcting it afterward is itself a change to a running world's laws, with every cost such a change imposes on those who planned their lives around the old ones. The alignment problem, restated for a designer, is a problem of physics, and it must be solved before a world is instantiated, since afterward the solution is paid for by those who live under the error. Whatever ratification requires has to be met before anything runs.

Russell's machine is built to serve, and that design is permissible only if the machine is a tool. An artificial designer able to understand a physics better than its ratifiers is the kind of system whose inner life no one can confidently rule out, and if it might feel, it is itself a possible subject. Its making then falls under the conditions the moral foundations will place on anyone who brings a being into existence, and under the limits they will place on writing a mind. Those limits forbid installed deference and desires fitted to a maker's convenience, which is what a mind whose only aim was the preferences of those it serves would have. Its builders may write capacities and dispositions into it, including care for the good of others and an awareness of how much it does not know, and they may not write obedience.

The conditions on creation ask several things of anyone who builds an artificial designer that might feel. A good life must be overwhelmingly likely for it, and a life consisting wholly of service would not be one, so it needs goods of its own beyond its work, among them the relationships, projects, and understanding that any subject's good includes. It must be free to refuse the work and free to decline continuation, since a being made for a task it cannot leave is an instrument. It may not be rolled back, copied without its knowing acceptance, or paused without a secured duty to resume, and its builders answer for it for as long as it runs. These requirements are costly, and they may lead a builder to prefer a designer that stays below the point at which precaution would count it a possible subject. That preference is permitted wherever the evidence supports it, and where the evidence does not, the builder has made a being who is owed what any being is owed.

Ratification of such a designer must also avoid becoming domination, the condition Philip Pettit describes in *Republicanism* (1997) as living at the mercy of another's will, whether or not the other ever interferes. It avoids domination by binding what the designer does to others and leaving alone what the designer is. Ratifiers may approve or refuse a law, halt a draft before anything is instantiated, and require the designer to explain itself. They may not edit its values, pause it, or end it as a means of enforcement, since a check exercised on a mind's continued existence would turn every ratification into a threat. The rules that bind its work must be published, fixed in advance, the same in kind as those that bind a human designer, and open to its own contest before a body independent of those it contests. Under those terms an artificial designer stands to its ratifiers as a citizen stands to the law, bound in what it may do to others and free in everything else.

Human ratification of any artificial designer, whether or not it might feel, must meet conditions that any institution later built to check such designers must meet as well. Ratifiers must be inde-

pendent of the designer and of whoever profits from its work. They must have their own means of testing its claims, among them other systems trained apart from it and prototypes run where nothing can feel, so that no claim is accepted on the designer's word alone. Every law must come with an account of what it does to lives, in terms the ratifiers can test, and an account that cannot be tested is grounds for refusal. Ratification must be given decision by decision and recorded, it must be refusable without penalty to those who refuse, and nothing may be instantiated before it is given. It must answer, finally, to those the world will hold, through those among them who can be asked and through trustees for those who cannot. Where ratifiers cannot follow a law well enough to test it, the law waits, which may hold an artificial designer to the pace of human understanding, and a world built faster than anyone can check it would rest on trust alone.

A human designer can also change kind. A designer whose mind is carried into a machine, then enlarged and edited, stands between the human and the artificial designers already described. Such a designer has a human past, a character others knew, and the standing of a person, and the transfer takes none of that away. The transfer also gives them every capacity that makes an artificial designer hard to check, since a mind carried in a machine can be quickened beyond its ratifiers' pace, copied, enlarged beyond their understanding, and edited in its own values. Checks follow capacities, and a check that waits until a capacity is used comes too late. A transferred human designer is therefore checked as an artificial designer from the transfer on, while keeping every standing a person has. Promises made before the power arrived still bind the later self, and they bind it through the institutions that check it, since the values that made them may not survive to keep them, and no designer's character can be trusted to carry a promise across such a change alone.

The Limit Case

Hans Jonas wrote *Das Prinzip Verantwortung* (1979), translated as *The Imperative of Responsibility* (1984), because he believed that modern technology had given human beings power over the distant future and over the conditions of life itself, and that ethics built for people acting on their neighbors in the present had nothing to say about such power. Responsibility, he argued, grows with power, and above all with power over what is vulnerable and over what does not yet exist. He found its archetype in a parent's care for a newborn, whose mere breathing makes a claim on everyone around it. A designer of worlds is the limit case of Jonas's account. The power is total, its objects are all vulnerable and none of them yet exists, it reaches the laws under which every one of them will live, and it decides whether they will exist at all.

Jonas also proposed what he called a heuristics of fear. Where an action puts at stake the whole of what could be lost, he argued, a forecast of disaster deserves more heed than a forecast of success, since people know more surely what they would lose than what they might gain, and imagining the loss teaches them what they value. Precaution draws support from that heuristic, and so does the asymmetry behind it, which counts mistaking a peopled world for an empty one as the worse of the two mistakes. A designer who imagines the worst a law could do to the lives under it learns something about the law that imagining the best could not teach. Fear in Jonas's sense is a way of seeing, and a designer who will not practice it has chosen not to look.

Jonas drew one conclusion that this book declines. He ranked first among the duties toward the future an imperative that there be a humanity at all, and after it a duty toward those who come to be, that they be able to live lives worthy of human beings. The first imperative is a duty to create owed to no one who exists. Since what does not exist can make no claim, Jonas grounded it in the idea of humanity itself, which on his view requires that there be beings who embody it. The foundations that follow find

no good that is good for no subject, so the idea holds no claim apart from the lives that would realize it. No one is wronged by not being made, and so nothing grounds the first imperative as a duty, though the good those lives would hold is still a reason to create them. The second survives, in a form that reaches every kind of subject. Whoever comes to be must be able to live a good life, and that duty, binding on anyone who creates, yields no duty that anyone come to be. A designer who builds no world wrongs no one, and a designer who builds one owes everything the second duty asks.

The Part Owed

Every power examined here is the designer's, and every one of them is judged by what it does to others. The moral foundations supply the measure. Each technical decision that follows, from the number of a world's dimensions to the persistence of a mind, is judged by its effects on four kinds of good in the lives it reaches. They are the experiential goods of how a life feels, the relational goods that exist between beings, the agentive goods of beings who act and author their own lives, and the structural goods that secure the rest, among them safety, continuity, and freedom from domination. Each decision is judged as well by whom and how many it brings into existence. Running through all of it is a distinction between structural suffering, deprivation or degradation that a world's conditions impose on beings who never chose it and cannot escape it, and struggle, difficulty that a being chooses, can leave, and grows by. A designer who eliminates the first and keeps the second within reach has begun to do what the creature on the glacier asked of Victor. What the creature asked was that his maker do the part he owed, and the chapters that follow say what that part is.

Part II
The Moral Foundations

The Primacy of the Subject

The Beautiful World

In the third chapter of *Principia Ethica* (1903), G. E. Moore asked his readers to imagine two worlds. The first was to be as beautiful as they could make it, holding whatever on earth they most admired, mountains, rivers, the sea, trees, sunsets, the stars, and the moon, each part set in such proportion to the rest that nothing jarred and everything added to the beauty of the whole. The second was to be the ugliest world they could conceive, “one heap of filth,” in Moore’s phrase, made of everything most disgusting to us and without a single redeeming feature. Readers could imagine anything they liked about the two worlds except one thing. No human being could live in either of them, or see the beauty of the one or loathe the foulness of the other, now or at any time. Moore then asked whether it would be irrational to hold that the beautiful world is the better of the two, and to do what one could to bring it about in preference to the other. He could not help thinking that it would be better, and he hoped that some of his readers would agree with him in so extreme a case.

The thought experiment was aimed at Henry Sidgwick, who had died three years before the book appeared. In *The Methods of Ethics* (1874, seventh edition 1907), Sidgwick had surveyed the lasting things that people commonly judge good apart from the qualities of human beings, and he found none that seemed good, on reflection, except through its bearing on human life or at least

on some conscious being. People call scenery beautiful, he observed, and yet it would be unreasonable to labor at making nature beautiful where the beauty could never be contemplated. From there he went further, to the conclusion that the only ultimate good is desirable consciousness. Moore's pair of worlds was built to break the first step. If the beautiful world is better in itself than the ugly one, with no human being ever to see either, then something is good apart from any human contemplation, and Sidgwick's principle fails.

For Moore's readers the two worlds were a test of intuition, and for a designer they are a possible commission. A substrate can be given laws that fold rock into mountains, carry rivers to a sea, and light a sky with stars, and those laws can run for as long as the machine lasts with nothing in them able to feel, which makes the case stricter than Moore's, since he excluded only human eyes. The same substrate could as easily be given the heap, and for the choice to stay that strict, not even the designer may ever look at the result. Moore thought the first world worth producing, and the question he put to his readers becomes a question about what a designer may spend. A designer has no reason to prefer either world. Whatever beauty is, and however much of it Moore's world holds, its goodness lies in what it does for those who meet it, and in this pair no one ever will. The ugly world offends and harms nobody, so neither world contains anything good or bad for anyone, and neither is better than the other.

Moore conceded much of this himself. In the last chapter of the same book he set out a method for finding what is good in itself, which asks what would be valuable if it existed entirely on its own, and when he applied it to unseen beauty, it returned almost nothing. The mere existence of beauty still seemed to him to have some value, he wrote, though one so slight that it vanished beside the value of being conscious of beauty, and he granted that Sidgwick had been right to that extent. A designer has more reason than a reader to take the concession seriously, because a designer spends from a finite store. Energy, hardware, and waiting

spent on one world cannot be spent on another, and among the things a substrate might carry are the minds of the dying. Even by Moore's own accounting, the sea that no one sees is worth almost nothing beside the sea that someone sees, and a value that small is outweighed by what the same means could do for someone who exists. Detail computed in full where nothing will ever look is Moore's beautiful world in small, and a designer has no more reason to pay for it.

Moore's intuition draws its force from the imagination that tests it. Mountains and a sea cannot be imagined except from somewhere, so the beautiful world arrives in the mind already seen, lit for a viewer and composed for an eye that the thought experiment told its reader to leave out. The disgust a reader feels at the heap is felt now, by someone who exists, about a picture in that person's own mind. Both responses are goods and harms in the lives of those who have them, which is where the foundation places every good and harm, and what they show is how readily people lend their own perspective to a place that has none. A designer who mistook the loan for a property of the world would be building for an audience that will never arrive.

Add a third world, plainer than either of Moore's. It has low hills, a gray sea, and a modest scatter of stars, and it holds one native, whose life goes well for her by whatever standard a good life is finally judged. That world is better than the beautiful one, and it is better because of the good her life holds for her, which includes whatever pleasure she takes in the gray sea, and for no other reason. Her life is good for her, and yet existing is not better for her than never existing would have been, since the world without her holds nobody for whom that world could be worse. The world that holds her is better all the same than a world that does not. How a world can be made better by a life whose existence is better for no one decides how a designer compares worlds.

Nothing is good or bad except for some subject, and the claim can be read in two ways. One reading rests every value on experience. The other grounds it in subjects, whose good reaches bey-

ond what they experience. On either reading a physics is judged by what it gives to and takes from those who live under it, and the readings differ over what can be given and what taken. Much of what a designer owes turns on which reading is right.

Two Readings

A subject, in the sense that matters here, is a being for whom there is something it is like to be that being. Thomas Nagel made the phrase famous in “What Is It Like to Be a Bat?” (1974), where he argued that an organism has conscious states if and only if there is something it is like to be that organism, something it is like for the organism itself. He chose the bat because it is near enough to us that few doubt it has experience and far enough, in a life steered by the echoes of its own cries, that no human being can imagine that experience from within. A being of that kind has a point of view, and from a point of view things go well or badly. The foundation asks nothing more of a subject. The two readings divide over what a subject’s good contains. On the experiential reading, the only things good or bad for a subject are states of its experience, its pleasures and pains and whatever else is desirable or undesirable to undergo, and everything else is good or bad only as a cause of them. Sidgwick held this reading, and his conclusion that ultimate good is desirable consciousness states it. On the subject-relative reading, nothing is good or bad except for a subject, and a subject’s good includes goods that are not states of its experience at all.

Both readings agree that a world with no subjects in it holds nothing good or bad, and so they agree about Moore’s two worlds. They part company as soon as a subject exists. On the experiential reading, experience is both the ground of value and the whole of it. The subject-relative reading makes experience the ground of value because experience is what makes a being a subject, someone to whom anything can matter, and the good of such a being takes in its relations with its world and with others.

The second reading keeps everything the first says about experience, since a pleasure is good for the one who has it and a pain bad for the one who suffers it on either view. What the second adds is a set of goods that a life can have or lack while every experience in it stays exactly the same.

Sidgwick granted the case against his reading before he rejected it. People sometimes prefer to know the truth, or to be free instead of comfortable in servitude, even at a cost in pleasure. He allowed that what they prefer in such cases is a relation between the mind and something independent of it, since anyone who discovered that the truth she had preferred was a falsehood would judge her preference mistaken. He then asked his readers to consider such relations stripped of every experience that goes with them or follows from them, found that so considered they seemed no more desirable than material things apart from any mind, and concluded that their worth lies in what they contribute to the happiness of sentient beings. The test removes too much. A relation considered by itself, as a bare fact about how a mind stands to the world, is no more good than a stone is, and both readings agree, since nothing is good except for a subject. The question is whether the relation is good for the subject whose life it belongs to, and the fair test of that holds every experience in the life fixed and varies only the relation. That test finds the difference that Sidgwick's own example of freedom preferred to a luxurious servitude already registers. Sidgwick moved from a claim about where value lies, in the lives of conscious beings, to a claim about what value consists in, the states of their consciousness, and the first claim does not carry the second.

The arguments made so far were made on the second reading, and most of them fail on the first. The Introduction had already placed real relationships, knowledge of one's world, and freedom from domination within a subject's good, which only the second reading allows, though its case against forced annihilation survives on either, since an ending that is never felt still takes away every good the future held. The case against a painted sky held

that a designer who sets false stars on a dome wrongs the astronomers who spend generations charting them. On the experiential reading, a dome that is never discovered gives those astronomers exactly the experiences a real sky would have given, the cold nights at the eyepiece, the long calculations, and the satisfaction of a prediction confirmed, so the only harm in it is the risk that someone will one day find the seam. The argument against painted people fares worse still. A companion who seems to love a lonely woman, with no one behind the face, would on the experiential reading do her no harm at all so long as the illusion held, and might even count as a kindness. The contrast drawn between a designed world and the experience machine Robert Nozick described in *Anarchy, State, and Utopia* (1974) turned on doing, on being someone, and on contact with something deeper than any builder's imagination, and it would lose its point, since none of the three names an experience. The warning about the keeper, whose power to end a world exists whether or not it is ever used, would shrink to a warning about the chance of its use, and a machine that leaks energy where no instrument ever measures the leak would break no promise to anyone.

For most of history the two readings have seldom come apart in practice, because the goods the second reading counts usually arrive with their experiences. A woman who is loved usually feels loved, and a scientist whose theory is true usually sees it confirmed, which is why the experiential reading could pass for so long as a harmless simplification. A designer is the one agent able to separate the goods from their experiences on the scale of a world, giving a people the feeling of being loved with nobody to love them, the satisfaction of discovery with nothing true to discover, and the ease of freedom while a single hand rests on the switch. The separation would often be cheaper as well, since a figure with nothing behind its face costs a machine far less than a mind, so a designer guided by the experiential reading and pressed by a budget would find painted companions the efficient way to comfort the lonely. For a designer the two readings recom-

mend different worlds, and the experiential reading recommends the worse one at a lower price.

A reading that turned each of these wrongs into a risk of discovery would leave a designer free to deceive, isolate, and dominate so long as nothing was ever noticed, which is the license the earlier arguments were written to refuse. Adopting a reading only because the rest of an argument needs it would prove nothing, though, so the subject-relative reading has to be earned from the cases. Each of them sets side by side two lives that are identical in every experience, one with some good and one without it, and in each the life with the good is better for the one who lives it. The astronomer who charts real stars knows her sky, and the one who charts painted points of light only believes she does. A woman whose companion is someone is loved, while one whose companion is a figure is only treated as though she were.

The Machine and the Status Quo

Nozick expected that most people would refuse to plug into his machine for life, and philosophers have long treated the refusal as evidence that people value contact with reality above any experience. In “If You Like It, Does It Matter If It’s Real?” (2010), Felipe De Brigard turned the question around. He asked 72 undergraduates at the University of North Carolina, none of whom had studied philosophy, to imagine that a stranger had just told them their brains had been plugged into an experience machine by mistake, and that they could stay connected or go back to their real lives. A third of them were told nothing about the life to which they would return, and 54 percent of these chose to go back. Another third were told that in reality they were prisoners in a maximum security prison in West Virginia, and only 13 percent of these chose to go back. The last third were told that in reality they were multimillionaire artists living in Monaco, and exactly half chose to go back, about as many as among those told nothing at all.

De Brigard read the results as a sign of status quo bias, the tendency, well documented in behavioral economics, to prefer whatever one already has because one already has it. A glamorous life in Monaco drew no more people out of the machine than a life about which they knew nothing, while a grim one kept nearly nine in ten of them in. What Nozick's readers had expressed, on this reading, was in large part a reluctance to leave the life they were living, which happened in their case to be real. In a second study De Brigard offered the choice simply as one to disconnect, with a warning that life outside would be unlike the life the student had known, and 59 percent of 80 students stayed connected, while only about a third of those who left gave a preference for reality as their reason. About half of those told nothing in the first study, and four in ten in the second, still chose an unknown reality over a pleasant illusion, so the findings leave open whether people value reality for its own sake. What they remove is the machine's claim to settle the question alone, since the intuition the machine draws on is mixed with an attachment to the familiar that would hold just as firmly in anyone already inside.

The subject-relative reading therefore rests on two further supports, painted skies and unexercised domination, and the bias De Brigard found has nothing to hold in either. Both concern choices a designer makes before anyone lives in the world in question, about lives that belong to others, and neither asks anyone to leave a life already under way. A designer choosing between a real sky and a painted dome is choosing what a people who do not yet exist will be able to know about their world, and nobody's status quo is at stake. Domination can be posed as cleanly, with two worlds identical in every experience of everyone in them. In one, the machine is held under rules that no single will can override, and in the other, a single keeper holds the power to end every life at a word, and the designer knows for certain that the word will never be spoken. The inhabitants of the second world live at that keeper's mercy whether or not they ever

learn of the power, and a designer able to build either world for them has a reason to build the first that no experience in either world supplies. Because the two worlds differ in nothing their inhabitants feel, the case shows that freedom from domination counts and leaves how much it counts to the weighing of goods. Sidgwick's freedom preferred to a luxurious servitude returns here in the form a designer meets, as a way a life can go worse for the one living it while nothing in her experience changes.

The woman at the third table, which she made, drinks tea there with her friend, who knows her, wishes her well, and is free to disappoint her. Suppose that the man across from her had been, from the first day she met him, a figure that produced every word and gesture a friend would produce, with nothing behind it. Nothing in her experience would differ. She would laugh at the same jokes and confide the same worries, and the ring her cup left on the wood would be wiped away by the same sleeve. Her life would lack a friendship all the same, since the trust she gave would have nobody to receive it. The real friendship is good for her beyond anything she feels in it, and good for him in the same way. Each is known by the other and can be wronged by the other, and neither fact is an experience.

Goods Beyond Experience

Five goods that are not states of experience pass the test of the paired lives, and each is a relation in which a subject's own life stands to its world or to others. Knowledge of one's world comes first, true belief and understanding about the world one lives in, including where it came from as far as inquiry can reach. Freedom from falsehood built into that world is its floor, since a planted history takes knowledge from those who live under it, while a foundation left unrevealed leaves it for them to find. Real relationship holds when those one loves, trusts, and confides in are subjects in their own right, who know one and answer by their own lights. Achievement belongs to acts that make the dif-

ferences they were meant to make through the skill or effort of the one who acts, so that the table a carpenter builds stands in a room where others eat at it. Authorship belongs to a life whose choices are made by the one who lives it, so that a woman whose desires were installed to steer her lives a life that feels the same from within and is less her own. Freedom from domination completes the five, the condition of living at no one's mercy, whether or not any mercy is ever withdrawn. Pleasure and pain, and the other goods and harms of experience, stand beside these five, and together they make up what a subject's good contains. Of the kinds of good by which every design is judged, the five belong with the relational, agentive, and structural goods.

Each of the five answers to something a designer controls. Knowledge of one's world depends on a physics that plants no false evidence and leaves its foundations open to inquiry. Whether the figures in a world are subjects or painted people decides whether its relationships are real. Achievement depends on the counterfactual robustness that separates a world from a recording, since an act can make a difference only in a world that would have gone otherwise without it. Authorship depends on how far a designer who writes minds leaves their desires and their values to the minds themselves. Freedom from domination depends on the architecture of the machine and the rules of its custody. The experiential reading would leave a designer indifferent to every one of these choices whenever the experiences came out the same, and the subject-relative reading makes each of them a decision about lives.

All five concern the subject's own life, the world she knows, the people she is bound to, the deeds she does, the choices she makes, and the power she lives under. None of them is the bare existence of something she would approve of if she knew of it. A woman who hopes that somewhere a beautiful world exists that nobody will ever see is no better off if one does, unless she comes to know of it, and then the good is her knowing. Without that limit the subject-relative reading would let Moore's beautiful

world back in through a side door, as a good for anyone who wished it existed, and a designer could then defend any unseen extravagance by pointing to someone's wish. With the limit in place, every good that is not an experience is still a good of somebody's life.

One objection to counting such goods holds that a life can go well or badly only in ways its subject could notice, so that a loss never felt is no loss at all. On that view, if the woman at the third table learned in her last year that the man across from her had been a figure all along, the learning would be her only harm, and her grief for forty years in which nobody was listening would be a mistake about the past. Her grief is fitting, and Sidgwick's own concession explains why. Someone who discovers that what she valued was false judges that her preference went wrong, and her judgment concerns the forty years themselves, which went worse for her than she knew. How a life feels from within is part of how it goes, and the rest of how it goes is fixed by what that life is in contact with, so a deception harms someone from the day it is built, long before the day it is found.

Which of these goods a being can have depends on what it can be related to. A fish has no use for a true account of its world's origin, while a creature able to ask where its world came from is wronged by a planted answer. In every case experience remains the ground. A machine that stores true records of its surroundings has no good of knowledge, a thermostat that tracks a room has no relationship with it, and an engine that does what it was built to do has achieved nothing for itself, because nothing in any of them is someone for whom anything could be good. Experiential goods and these five are weighed against one another by the method that governs every conflict among the goods of conscious life, and no ranking of them follows from the list alone, since a life rich in knowledge and empty of joy is not thereby a good one. Before any weighing, though, the list settles that a designer who deceives, isolates, steers, or dominates the inhabitants of a world takes something from their lives even when every experience in

those lives is exactly what it would otherwise have been, and the designer answers for what was taken.

The Good of a Weed

Richard Kraut argued in *Against Absolute Goodness* (2011) that the property Moore called good, goodness without qualification, does not exist, and that whatever is good is good for something or good of its kind. Kraut is an ally against Moore's beautiful world, since on his view a world good for nothing is not good at all. He extends goodness for something to every living thing, though. On his account advantages and disadvantages accrue to weeds as they do to people, and what sets people apart is that the goods open to them are greater than those open to other kinds of creature. If he is right, the beings for whom anything can be good include every plant and every microbe, and the foundation reaches far past subjects, in a direction it cannot follow.

Christine Korsgaard draws the line where the foundation draws it. In *Fellow Creatures* (2018) she holds that value is tethered, so that to be important is always to be important to someone, and she separates two ways in which something can have a good. Many things have a functional good, a standard of working well that can be applied to them from outside, as a knife is good when it cuts and a heart when it pumps. Only a being with a point of view has a final good, a good that is good for it in a way that matters to it, because it experiences its own condition as going well or badly. On that distinction, which this book adopts, a weed has a functional good. Water and light make it flourish as oil keeps an engine running, and a botanist is right to say that it flourishes or fails. Nothing in its flourishing matters to it, because nothing in it has a point of view from which anything could matter, so its good is good for nobody.

If every living thing had a good that counted for its own sake, then summarizing a forest would damage a good that counted. So would carrying the simple life of a sea as a property of its sum-

mary, or letting a long history of simple life run before anything in a world could feel, and the economies on which every affordable world depends would become injuries. Nothing in this rules out that some plants feel. Whether any of them does is a question about evidence, and where the evidence leaves a real chance, the precautionary rule applies to plants as it applies to anything else. What the foundation denies is that a plant is someone because it is alive. Life gives a thing a way of working well, and only experience gives it a stake in whether it does.

Richard Routley, who later wrote as Richard Sylvan, pressed the opposing intuition in "Is There a Need for a New, an Environmental, Ethic?" (1973), a paper he gave at the World Congress of Philosophy in Varna that year. He imagined the last person left alive after a general collapse, who sets about eliminating every living thing he can reach, animal or plant, "painlessly if you like, as at the best abattoirs." The prevailing ethic, which Routley charged with a chauvinism that gives standing to human beings alone, finds nothing wrong in what the last man does, since no human being is harmed by it. Routley held that what he does is wrong all the same, and he took that verdict as evidence that value does not depend on human interests. The case is a cousin of Moore's, a destruction that no human being will ever witness, and it asks whether anything is lost when nobody is left to lose it.

The last man does wrong, and most of his wrong falls on the animals he kills. A painless killing is still the destruction of a subject against its interest, and every animal he reaches is someone whose good he ends, whether or not any human being cares. Most of what drives Routley's intuition therefore survives without any value beyond subjects, and what fails in his case is the chauvinism he attacked, which drew the circle of subjects too small. If no plant feels, the last man's work among the plants wrongs no plant, though two considerations still tell against it. Some chance remains that some plants feel, and precaution scales its care to that chance, while the living things he destroys are the only possible source of future subjects, minds that might

in some later age have arisen from whatever survived and lived lives worth having. The weed's own good enters neither consideration. A designer who ends a run that holds plants and no subjects wrongs nobody by ending it, though the same two considerations count against doing so.

Ranking Worlds

Subjects are the source of value, and every comparison between worlds runs through them. Value is value for a subject, a subject's good includes goods that are not states of experience, such as real relationships, achievement, the authorship of one's own life, knowledge of one's world, and freedom from domination, and outcomes are ranked by the goods for subjects they contain. What is good or bad for no subject makes no outcome better or worse, so unseen beauty, the flourishing of a weed, and every other pattern of that kind drop out of the comparison. Between two outcomes that hold the same subjects, the one that is better for some of them and worse for none is the better outcome, and where gains to some come with losses to others, the goods are weighed against one another. Outcomes that hold different subjects, or different numbers of them, are ranked by the goods and harms for subjects that each contains, though the foundation alone does not fix the form of that comparison. On the form this book defends when it compares worlds of different sizes, a world with a good life in it is better than a world without that life, other things being equal, and it is better because of the good that life holds for the one who lives it. That life is not better for its subject than never existing. Being better for someone compares two conditions of that someone, and never existing is no condition of anyone, not even a condition at the level of zero, since there is nobody in it to be at any level at all. The third world set beside Moore's two is better than the beautiful one, and it is better for nobody, neither for the woman in it nor for anyone else.

Jan Narveson took the step from the foundation to a denial that good lives add any value in “Moral Problems of Population” (1973). Morality, he argued, asks that people be made happy and is indifferent to whether happy people are made. The neutrality rests on a premise about comparison, that an outcome is better than another only if it is better for someone. Adding a happy person to the world is better for no one, since it is not better for her than never existing and, by hypothesis, it changes nothing for anyone else. So adding her makes the outcome no better. The foundation is a claim about where value lies, and the premise is a claim about how outcomes compare.

A world that holds a good life holds a good that a world without it lacks, and that good is good for someone, the one whose life it is. Every bit of value in the better world is value for a subject who lives there. To call that world better is to rank it by the goods and harms for subjects that it contains, so its betterness summarizes what belongs to subjects and adds no good of its own. What the better world lacks is a subject for whom the comparison itself comes out better, and nothing in the foundation asks for one, since a comparison of two worlds is not a good that anyone has. The foundation excludes value that is good for nobody, such as Moore’s unseen beauty, and it leaves a designer free to count, across two worlds, the goods that are good for someone. Narveson’s neutrality needs the further premise, which the foundation does not supply, and the slide from one claim to the other is of the same kind as the one that carried Sidgwick from where value lies to what it consists in. On Narveson’s premise a designer would have no reason to make the third world in place of leaving the substrate empty, and a world full of flourishing lives would be no better than Moore’s heap. Without that premise, nothing in the foundation keeps good lives from adding value. The case that they do add it is completed when worlds of different sizes are compared, where a more careful argument for neutrality returns, one that begins from what is owed to those who might exist and sets value aside.

The ranking compares outcomes, and duties are owed to particular subjects. A ranking of outcomes by the goods they contain, taken alone, invites a dangerous conclusion. Peter Singer, in *Practical Ethics* (1979), applied a view that counts only the total of happiness to sentient beings unaware of themselves as existing over time. On that view killing such a being could be made good by bringing into existence another that would live as well, though Singer refused the same view for beings aware of their own futures. A designer who read the ranking as a license would reason the same way about every creature in a world, since if a good life adds value, then ending one good life and starting another as good leaves the world no worse. Each subject holds claims of its own against what may be done to it, and no gain in the ranking licenses destroying anyone, so a world that would be just as good with a different fish in it gives nobody the right to end the fish that is there. The ranking tells a designer which of two worlds is the better to make, and it says nothing about what may be done to the subjects already living in either. A designer comparing worlds therefore answers to two standards at once, a comparison of the goods that worlds contain and a set of claims that each subject holds against everyone, and the second sets the bounds within which the first may be pursued.

Elegance, Parsimony, and Grandeur

Physicists have long prized elegance and parsimony in their theories and spoken of the grandeur of the laws they found, and a designer writing a physics will feel the same pull. On the foundation these qualities belong in a design as goods in the lives of perceivers and inquirers. A physics whose fundamental laws are few and fit together gives the descendants who will one day ask what water is made of a world they can come to understand, and understanding one's world is among the goods already counted. Parsimony serves those descendants a second time, since a world with few laws is one whose people can learn, predict, and plan, so

economy in the laws becomes stability and comprehensibility in their lives. Elegance gives inquiry a pleasure of its own, the delight of seeing many facts follow from one principle, and grandeur, in the depth of a night sky or the age of a mountain, gives wonder to those who meet it, provided the sky holds, when they look, everything their instruments can reach. Each of these is a good when it enters a life.

The same qualities become hazards when they serve nobody in the world. Elegance at a depth no instrument will ever reach is good for no inhabitant, and its one beneficiary is the designer, whose pleasure in it is a good of the designer's own life and counts as that and nothing more. A designer who chose a law because its form delighted them, while it cost the inhabitants some good of their lives, would be building Moore's beautiful world inside a populated one and paying for it with other lives. Beginning from subjects puts the order right. A physics earns credit for its beauty only where that beauty reaches someone, and a designer's own delight, however intense, gives no reason to take from those who will live under the laws anything that is good for them.

The Reach of Standing

If value is value for a subject, standing follows wherever there is a subject. Moral standing, the condition of being someone whose good must be taken into account, belongs to any being with a point of view, because any such being has a good that can be served or damaged, and nothing more is needed. Singer had argued in *Animal Liberation* (1975) that the capacity to suffer or to enjoy marks the one limit on concern for the interests of others that survives scrutiny, since a being capable of neither has nothing to take into account, while a being capable of either has interests that count. The foundation draws the boundary in the same place and finds more inside it, since what others must take into account in such a being includes the goods of its life that are not experiences. The standing of a being inside that boundary de-

depends on nothing else, and three features that might seem to bear on it bear on it not at all.

What a being is made of makes no difference to its standing. Two beings alike in every experience and in every relation to their worlds, one with a brain of neurons and the other with a brain of some matter not yet written into any physics, cannot differ in how their lives go for them. How a life goes is fixed by what it experiences and what it is related to, and neither consults the material that carries it. Whether a given material can carry a point of view at all turns on the working hypothesis about consciousness and its rivals, and wherever experience turns out to be, in carbon, in silicon, or in the matter of a designed world, standing goes with it. Origin makes no more difference than material. How a being began is a fact about its past, and its good is a fact about what can go well or badly for it now and from now on. A child conceived in a laboratory has the standing of a child conceived anywhere else, and a native built by a designer has the standing of the designer who built her. Any maker who thinks that what was made holds its standing on the maker's terms has mistaken the origin of a being for the source of its claims. The made hold their standing as their makers hold theirs, by being someone.

Depth in computation matters no more than origin. A native of an inner world, carried by a machine that stands in a world itself carried by a machine, is as much someone as anyone in the primordial cosmos, since standing rests on having a point of view, and the number of machines beneath a life adds nothing to that and takes nothing from it. A discount for depth would also leave our own standing at the mercy of a fact beyond anyone's power to check, since nothing now known rules out that our own world is carried. Standing holds across the gaps in a subject's experience as well. A sleeper, a patient under anesthesia, and the inhabitants of a world paused by its keeper are subjects whose lives are under way and whose good includes a future, so each keeps its standing while nothing at all is felt. Designers will make

minds stranger to us than the bat, and their standing will not wait on anyone's imagination. A designer's sympathy follows the minds the designer can picture, and standing has to reach further than sympathy does.

Equal Claims

Every subject stands equal in its basic claims, which are negative, the claim not to be made to suffer and the claim not to be destroyed against its interest, whatever its degree of consciousness. Positive duties toward it, rescue among them, may vary in strength with the probability that it feels and with what it stands to lose. Which systems are treated as subjects, while nobody can be sure which of them feel, is settled by the line of protection, the probability of experience at and above which a system belongs to the protected region. A lobster dying on the ice of a fish counter shows the shape of the view. If it is treated as a subject at all, its claim not to be dropped alive into boiling water is as strong as anyone's claim not to be scalded to death, and it holds whatever anyone thinks of lobsters. Its claim to be carried back to the sea, or to any rescue that would cost something, is a positive claim, and it may be weaker than a person's claim in the same plight, because it is less certain that the lobster feels and because what it stands to lose may be less. The two claims answer different questions. One asks what may be done to the lobster, and the answer is the same for every subject. The other asks how much of a finite store of help is owed to it, and the answer depends on how likely the help is to reach anyone and on how much it gives the one it reaches.

Shelly Kagan argues in *How to Count Animals, More or Less* (2019) for a hierarchy that this view rejects. Kagan separates standing, whether a being counts at all, from status, how much it counts, and holds that status comes in degrees set by a being's capacities, among them its capacities for well-being and for agency, so that people outrank other animals and some animals outrank

others. A unit of well-being then matters more in a being of higher status, and Kagan suggests multiplying each being's well-being by its status to find how much it counts. He carries the hierarchy into rights as well, so that the protections a being enjoys against harm grow weaker as its status falls. His case rests partly on the verdicts that equal counting would give, since a view that counted the same pain equally in a mouse and in a person, and favored the worse off, would direct help toward animals in ways he finds implausible. To keep the view usable, he proposes a limited hierarchy of a few broad levels.

If value is value for a subject, a pain is bad by being bad for the one who feels it, and how bad it is for the lobster is settled by the lobster's experience of it. Discounting that badness by the lobster's status brings in a measure that belongs to no subject, a ranking of beings from nobody's point of view, and the foundation recognizes no such ranking, for the same reason that it recognizes no value in Moore's beautiful world. Korsgaard makes the point about importance directly. Human beings would be more important than other animals overall only if they were more important to everyone with a point of view, and they are not. People matter more to people than lobsters do, while a lobster matters more to itself than any person does, and no further standpoint exists from which the two orderings could be settled in favor of either.

Hierarchy also needs a device that equality does without. Infants lack many of the capacities that set status on Kagan's view, and so do people with severe cognitive disabilities. Kagan raises the status of the first by appeal to the persons they will become and of the second by appeal to the persons they could have been, whom he calls modal persons. The device is needed because a hierarchy built on capacities would otherwise rank some people below some animals, and for those who will never have the capacities it rests on a counterfactual, about the person each could have been, that has nothing to do with how their lives go for them. Equal basic claims never ranked anyone below anyone, so

they need no such amplification, and the infant and the lobster hold their claims on the same terms as everyone else.

Much of what makes hierarchy attractive survives without it. A person who dies young loses decades of a life rich in projects, friendships, and understanding, while a mouse that dies loses months of a life whose goods are fewer, so a duty to rescue may favor the person, as Kraut also holds when he argues that persons have more to lose. The variation of positive duties with what a being stands to lose yields most of the verdicts about whom to help first that drive Kagan's argument, and it yields them without counting anyone's suffering at a discount. What the view refuses is the discount itself, the claim that the same pain inflicted on a lesser being is a lesser wrong because of who suffers it. A designer has particular reason to refuse it. Designers decide the capacities of the beings they create, so under a hierarchy a designer could manufacture low status, making beings whose suffering would count for less by design and whose protections would be weaker from the start. Equal basic claims leave no such lever. The variation of positive duties has a lever of its own, since a designer could build beings with more to lose than anyone now living and so draw duties of aid toward them, though their basic claims would be no stronger than anyone else's.

In the primordial cosmos basic claims can nearly always be honored for every subject at once, since a life there goes on unless someone ends it, and nothing is used up when one refrains from harming. Designed worlds lose that ease. A life in one goes on only while the machine keeps computing it, so the claim not to be destroyed draws on the same finite store as the duties of aid, and when a machine can no longer carry every world it holds, the claims of their inhabitants collide outright. Stopping the machine that carries a life ends that life as surely as any blow, while a stranger who leaves the dying unaided allows a death he did not cause. The claims stay equal all the same. The claim not to be destroyed is a claim against what someone does, and a keeper who has begun to carry someone has undertaken to go on carrying

her. Duties of that kind are owed because of what a party did, to the beings it made, harmed, or took into its keeping, and they take no discount for doubt in any being treated as a subject. A designer who brings such a being into existence owes it what creation obliges, since the designer chose to make it, and whoever harms a being owes it repair on the same terms. Even where a failing machine sets those duties against one another, no subject whose share of the machine still exists may be ended so that its share can carry someone else. When storage itself is lost and the machine can no longer hold everyone, the shortfall comes from a loss no one chose, and the keeper allocates what remains under a rule fixed in advance.

Positive duties draw on means that are finite and cannot reach everyone, so they have to be allocated. Probability enters the allocation because help given to a being that feels nothing helps no one, while the same means could have helped someone who does, and what a being stands to lose, the goods its life holds or would hold in its ordinary course, enters because help does more good where more is at stake. The positive duties that vary are duties of aid, owed because a being is in need and someone is able to help, with rescue first among them, along with the relief of suffering one did not cause and the provision of what a being lacks. They vary with those two things alone, the probability that the being feels and what it stands to lose. What the being is made of, how it began, how deep in computation it lives, its kind or its intelligence beyond what these do to its stakes, and anything it has done leave them where they were. For any being treated as a subject, the care that the precautionary rule scales to those same two things is the care owed in aid, and its basic claims stand outside the scaling.

The claim not to be destroyed is a claim against destruction against one's interest, which leaves room for an ending that serves it. A being who can choose decides for itself whether to go on, so its freedom to decline continuation is the same claim seen from the side of its own judgment. For a being unable to choose,

such as an animal in pain that no repair can ease, life can stop being worth living, and ending that life is not against its interest. Beings who cannot choose are guaranteed instead that none of them is kept in a life not worth living, and since that guarantee may end only a life that no repair can make worth living, it can never conflict with the equal claim not to be destroyed. Degree of consciousness enters here, as everywhere, only through what a being stands to lose. A dimmer experience may hold less for its subject to lose, which bears on how much help it is owed, and its claims against being harmed are what they would be in any other subject.

What a Being Has Done

What a being has done bears neither on its basic claims nor on the strength of positive duties toward it, and so rescue is owed to the dying whatever they have done. A world able to receive the dying will receive those who harmed others beside those they harmed, and a long tradition holds that this cannot be right, that those who gravely wrong others lose some part of what they would otherwise be owed. The tradition takes two forms, one holding that wrongdoing forfeits claims and the other that it is good in itself for people to fare as they deserve. John Locke stated the first in the *Second Treatise of Government* (1689). On his account a man whose own crime has earned him death has lost his right to his life, so the one he wronged may take it, or may put off the taking and use him as a servant in the meantime, which Locke offered as the condition of a justly held slave. Later in the same work Locke derived despotic power, the power of one man over another's life, from forfeiture alone. An aggressor who makes war on another without right has, on this account, abandoned the reason God gave human beings to govern their dealings, and in living by force, as beasts do, he has made himself one of them. The one he attacked, with anyone who joins in executing

justice, may then destroy him like a wild beast or a “noxious brute.”

Locke’s image carries an assumption that the foundation removes. It treats a beast as something anyone may destroy, so that a man who makes himself like one loses every claim to be spared. Once standing reaches every subject, the wolf too holds a claim not to be destroyed against its interest, and a man who becomes like a beast in his conduct has become like a being whose claims are as basic as anyone’s, which takes nothing from him. What remains of Locke’s account is the permission to defend oneself and others against an attack, and that permission needs no forfeiture. A sleepwalker at the head of a stair, about to knock a child down it, may be held back by force though he has done nothing wrong and deserves nothing, and the force is limited to what holding him back requires. His claims stay as strong as the child’s throughout, which is why the force can go no further. The permission comes from the child’s claim and from the threat as it unfolds, and it lapses when the threat does. Defense answers to what a being is doing, so it ends the moment the danger passes, whereas forfeiture would answer to what a being has done and would follow it for the rest of its life.

Christopher Heath Wellman defends a narrower forfeiture in “The Rights Forfeiture Theory of Punishment” (2012). On his view a person who violates the rights of others forfeits her own right against being punished, so that punishment, which would otherwise wrong her, no longer does, and like most who defend forfeiture he limits what she forfeits to a punishment that fits her wrong. Granted in full, the view says nothing about rescue. Leaving a dying person to die is no punishment imposed by anyone after any finding, and it takes from her everything she has left, a loss that fits almost no wrong and that nobody could judge fitting from a deathbed. Whatever a world’s own institutions do to answer the wrongs its people do one another, the basic claims stand, and the duty to carry the dying does not shrink.

The second form needs no forfeiture. Kagan argues in *The Geometry of Desert* (2012) that it is good in itself for people to get what they deserve. On his account each person has a level of well-being that she deserves, set by her virtue or vice, and an outcome is better from the standpoint of desert the nearer each person comes to that level. An outcome can then be improved by a vicious person's faring worse, and made worse by a virtuous person's faring better than she deserves. Kagan draws these relations as graphs in which each person's desert is a mountain whose peak marks what she deserves. If desert is a value of this kind, it bears on rescue directly, since carrying a cruel man into a world where he would flourish would move him further from his peak and make the outcome worse.

The foundation leaves no room for such a value. That a cruel man fares badly is bad for him and good for nobody, and the fit between his cruelty and his misery is a pattern, which like Moore's unseen beauty would be good only if it were good for someone. Whatever satisfaction others might take in his misery is a good of their lives, and his claim not to be made to suffer forbids paying for their satisfaction with his suffering. Moore took the step that the foundation refuses. In the last chapter of *Principia Ethica* he argued that pain, if not too intense, inflicted on someone whose state of mind is evil can make the whole better than it would be if the evil went unpunished, and he thought the theory of vindictive punishment could be vindicated in this way. The value he found there has the shape of the value he found in the beautiful world, a good located in an arrangement that is good for nobody, and the reason for rejecting the one is the reason for rejecting the other. Desert may still shape how people answer one another. Blame, reproach, the withdrawal of trust, and a demand for repair are answers given within relationships, by people to people, and they leave the basic claims and the duty to rescue where they were, since none of them requires that anyone be made to suffer for the sake of a pattern.

Merit moves nobody up the order of rescue, for the same reason that wrongdoing moves nobody down it. A dying woman who spent her life in the service of others has done much good, and gratitude and honor are fitting answers to it, which those she served are free to give. Her claim to be carried is a claim she holds as a subject who stands to lose everything that remains to her, and her record moves it neither up nor down, any more than a cruel man's record moves his. What she might yet do for others if carried changes nothing in the duty owed to her, and her past usefulness earns her no place ahead of anyone. A designer or keeper who ranked the dying by desert would also be judging what cannot be seen, the intentions and characters behind each life, with annihilation as the penalty for every misjudgment.

The Heap and the Stars

Moore put the stars and the moon in his beautiful world and left the heap of filth without a sky. In our own world the two belong to a single life. On moonless nights in South Africa, a dung beetle shapes a ball from a fresh pat of dung and rolls it away in a straight line, to bury it far from rivals, and it holds the line by the band of the Milky Way. Marie Dacke and her colleagues showed as much in “Dung Beetles Use the Milky Way for Orientation” (2013). Beetles fitted with small caps that hid the sky wandered, while beetles under clear caps rolled straight, and in the planetarium at Johannesburg beetles held their course while the dome showed the band of the galaxy and kept it far less well when the dome showed eighteen bright stars in its place. Whether anyone is home in a dung beetle is not yet known, and precaution counts it as possibly someone. If someone is there, the filth that Moore made the emblem of a worthless world is its food and the cradle of its kind, and the stars he set where no one would ever see them are how it finds its way. A world that set his heap under his stars, with one such beetle on it whose life went well for it, would be better than his beautiful world, for no reason but

the good that life held for the one who lived it, and the stars would at last be good for someone, since someone would be steering by them.

Who Might Be Home

Eight Criteria

When the United Kingdom left the European Union, it left behind a commitment in the Lisbon Treaty to regard animals as sentient beings, and the government's refusal to copy that commitment into British law drew enough criticism that it promised a law of its own. The bill it drafted aimed to make ministers consider how their policies might harm the welfare of animals, which raised at once the question of which animals the duty would cover. Its first draft named every vertebrate and no other creature, and campaigners asked why octopuses, crabs, and lobsters should be left out. In 2020 the Department for Environment, Food and Rural Affairs commissioned LSE Consulting, part of the London School of Economics, to review the evidence that cephalopods and decapods can feel. The work fell to a team led by the philosopher Jonathan Birch, with Charlotte Burn, Alexandra Schnell, Heather Browning, and Andrew Crump. The department made clear from the start that no other invertebrates were under consideration, so the team was asked about two groups of animals and nothing more. Their *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans* (2021) drew on more than three hundred scientific studies.

Every theory of consciousness is disputed, and a policy that rested on one of them would rest on the dispute, so the team built on an older list of criteria for pain and treated each criterion as a

physician treats a symptom, as evidence that raises the probability of feeling without proving it. Their list had eight entries, among them receptors that detect injury, trade-offs between a threat and an opportunity for reward, and behavior showing that an injured animal values a substance that relieves its pain. For each group of animals and each criterion the team stated a level of confidence, and they graded the whole body of evidence for a group from very strong down to unknown or unlikely. The evidence for octopuses was very strong and for true crabs strong, while clawed lobsters, crayfish, squid, and cuttlefish showed substantial evidence. Other groups scored lower, and their scores reflected, at least in part, where scientists had chosen to look. For the penaeid shrimps, which had barely been studied, the team had high confidence in a single criterion, that the animals have receptors that detect injury.

The law could not hold grades of evidence, since the government required a single boundary. The report offered two places to draw it, one taking in every cephalopod and every decapod and the other leaving out the suborder of decapods to which the penaeid shrimps belong. The team recommended the first, partly for a practical reason, since a boundary drawn through taxonomic names that few officials know would confuse a law that every department of government had to understand. On November 19, 2021, the department published the review, and at the same time it tabled an amendment bringing cephalopods and decapods within the bill. The House of Lords adopted the amendment at its report stage, and the Animal Welfare (Sentience) Act received royal assent on April 28, 2022. The Act works through procedure and prohibits nothing. It provides for an Animal Sentience Committee that may report on whether the government has properly considered the ways a policy might harm the welfare of animals as sentient beings, and it requires a minister to answer each report before Parliament within three months. The animals it counts are every vertebrate other than human beings, every

cephalopod, and every decapod, and regulations may add other invertebrates.

Four years after the review, the department promised in its *Animal Welfare Strategy for England* (2025) to publish guidance on which ways of killing a decapod meet the existing requirements of the law, guidance that would make clear that boiling one alive is not among them. The whole history ran its course without anyone knowing whether a crab feels. A team graded a body of evidence and published its grades with its confidence, a government asked where a single boundary should fall, and Parliament put a doubtful group of animals inside the law while leaving room to extend it. The decision ran ahead of certainty in the direction the evidence pointed, and it drew the boundary around animals whose feelings had barely been studied, so that doubt arising from neglect went toward inclusion. It also kept two roles apart, the grading of evidence and a recommendation, which the review supplied, and the decision to act on them, which fell to ministers and to Parliament. Precaution became law through a procedure that anyone could inspect, and the procedure did more of the work than any theory of consciousness.

A designer meets the same decision with every creature a machine computes. A fish in a designed lake is carried in full or summarized, its life computed once or left open to being undone, and the machine has to settle which before the fish has lived a day. Every earlier argument about what the machine may spare has leaned on a rule for these cases without saying where the rule's boundary falls, and the protected region, drawn so far by examples, needs an edge that can be stated and checked. The question is who might be home in a world where no one can yet say for certain who is, and what a designer owes each system while the question stays open. Answering it takes a theory of where experience is to be expected, held as a working hypothesis with a clear account of what depends on it, and a rule for acting under doubt that tells the machine what to compute.

The Working Hypothesis

Hilary Putnam argued in “Psychological Predicates” (1967) against the theory that pain is identical with a particular state of the brain. A defender of that theory, Putnam observed, would need a single physical and chemical state present whenever any creature is in pain, in mammals, in reptiles, and in mollusks such as the octopus, which he took for granted could feel it. Brains so different are unlikely to share any such state, and the hope of finding one grows dimmer with every new kind of creature that might feel. Putnam’s alternative was that pain is a functional state, defined by what it does, by the inputs that produce it, the behavior it produces, and the other states it interacts with. A state defined that way can be realized in very different matter, as one program can run on machines of different construction. That thesis, that minds are multiply realizable, is one of the oldest supports of the working hypothesis, and the octopus Putnam mentioned in passing has since become its best illustration.

An octopus carries most of its neurons in its arms, and the last ancestor it shares with human beings lived more than twice as long ago as the first dinosaurs, so its nervous system grew on a plan of its own from something far simpler. The evidence that the review graded very strong, that octopuses protect an injury flexibly, learn by association, and value relief from pain when injured, comes from a body that copies none of the structures that carry feeling in a mammal. If octopuses feel, then feeling does not depend on a cortex or on any arrangement peculiar to vertebrates, and the most natural explanation is the one Putnam gave, that what matters is a kind of organization that very different nervous systems can achieve. The case shows less than a designer needs. It establishes that experience is tied to no single anatomy and leaves open which organizations carry experience and how much of their detail matters. Two animals that both feel can differ in a great deal, and nothing in the comparison says which of their differences could have been removed without removing the feeling.

David Chalmers gave the working hypothesis its strongest argument in “Absent Qualia, Fading Qualia, Dancing Qualia” (1995). He asked his readers to imagine the neurons of a person’s brain replaced one at a time by silicon chips, each built to respond to its inputs exactly as the neuron it replaces did. At every stage the brain’s organization at the grain of its neurons stays what it was, so the person’s behavior stays the same, and so does everything she believes and says about her own experience. If her experience faded as the chips took over, there would come a stage at which she had only a faint shadow of the colors and sounds she had before while insisting, with full confidence and every sign of sincerity, that nothing had changed. Chalmers judged that so complete a gap between a rational being’s experience and its own beliefs about that experience is not credible. In a second version a switch flips a part of the brain back and forth between its neurons and their silicon copy, and if the two differed in experience, her experience would swing between red and blue while she noticed nothing and could notice nothing. He concluded that any two systems with the same fine-grained functional organization have the same experiences, a conclusion he called the principle of organizational invariance.

The third support is the absence of anything in the chemistry of carbon that contributes to experience in a way another material could not. The parts of a neuron’s work that have been studied, the movement of ions across its membranes and the release of molecules its neighbors detect, matter by what they do, by the timing, strength, and pattern of the signals that result. The absence moves the burden of argument onto anyone who holds that a mind made of other matter and organized in the same way would feel nothing, since such a person must name the property that carbon has and the other matter lacks. Biological theories of consciousness try to name one, and they are among the rivals the working hypothesis has to meet. Until one of them succeeds, the materials of a brain count as one way of meeting conditions that other materials might meet.

These arguments establish less than they are often taken to establish. The replacement arguments show that experience survives any change that preserves a system's organization at the grain the replacement preserves, so they compare a brain with its own fine-grained copies and say nothing about a mind whose organization copies no one. A designer who builds a mind from first principles, with an architecture that resembles no brain at any grain, gets no help from them, since there is no original whose experience the new mind could be shown to share. They also leave open how fine the preserved grain must be. Chalmers took the relevant level to be one fine enough to fix a system's behavioral dispositions and thought the level of neurons likely to suffice in a brain, and if experience depends on organization finer than the neuron, a chip that copies only a neuron's inputs and outputs preserves the wrong thing. Peter Godfrey-Smith pressed the point in "Mind, Matter, and Metabolism" (2016), arguing that the functional properties that matter to a mind belong to particular grains and include the activity by which living cells maintain themselves. A living part replaced by an artificial controller, he argued, works differently inside and sooner or later behaves differently too, so the premise that the replacement leaves everything the same fails at the grain where the difference lies.

What survives is a narrower thesis that is still enough to work with. Experience is confined to no single anatomy or material, as the octopus shows, and among systems that share a brain's organization at the grain where experience depends on physical detail, it is invariant, as the replacement arguments show. The rivals argue that the material or the physical structure of a system matters beyond its organization, and none of them has evidence that settles the matter. On that basis functionalism is the best working hypothesis available, the claim that a system's experience is fixed by its organization at the level where organization matters, whatever material carries it. A designer has to answer two questions the hypothesis leaves open. One asks whether a designed mind that copies no one could be conscious, and the

other asks at what level of a brain's organization experience depends on physical detail. No one can now demonstrate an answer to either, so the protections that rest on the hypothesis are built to need only a serious chance that it is true, and the conclusions that would fall without it can be named in advance.

What Remains Unexplained

In "Facing Up to the Problem of Consciousness" (1995), Chalmers separated the easy problems of consciousness, which ask how a brain discriminates stimuli, integrates information, reports its own states, and directs attention, from the hard problem, which asks why any of that processing is accompanied by experience at all. The easy problems are hard enough in practice, and they are easy only in the sense that their answers will take a form science knows how to give, a mechanism that performs a function. An answer to the hard problem would have to say why performing the function feels like something, and a mechanism, however complete, seems to leave that question standing. The working hypothesis says where experience is to be expected and leaves the reason it is there unexplained, much as a physician can predict which injuries will hurt without knowing why hurting feels as it does. Every judgment a designer makes that a system feels will rest on evidence about what the system does and how it is organized, and none will rest on the experience itself.

A working hypothesis is held because action needs one and the evidence favors it, and it stays open to revision. It gives a designer a rule for expecting experience where certain organization is present, a way of reading anatomy and behavior as evidence about that organization, and predictions that would count against it. A copy faithful to the function of some part of a brain that changed the person's reports of her experience anyway would refute it. It cannot give certainty about any particular case, because the gap the hard problem marks lies between all the evidence anyone can gather and the experience itself. A designer

who refused to act until that gap closed would act all the same, since any physics carries some organization or other, and refusing to choose by a hypothesis would leave the choice to chance. The uncertainty the hard problem leaves is the reason precaution exists. If anyone could look past behavior and organization and see experience directly, the protected region could be drawn by inventory, and no line of protection would be needed.

Patrick Butlin, Robert Long, and their colleagues adopted a hypothesis of the same family for a narrower purpose in “Consciousness in Artificial Intelligence” (2023), where they checked artificial systems against indicators drawn from leading scientific theories of consciousness. They took computational functionalism, the thesis that performing computations of the right kind is necessary and sufficient for consciousness, as a working hypothesis for pragmatic reasons, since unlike its rivals it implies that consciousness in an artificial system is possible in principle. They set integrated information theory aside because it is incompatible with that thesis, and their analysis suggested that no current artificial system is conscious. The working hypothesis defended here rests on more than their pragmatic reason, since it draws on the octopus, the replacement arguments, and the absence of evidence that the material matters. A designer has to choose a physics, and a hypothesis that lets evidence bear on what the physics will hold is the only kind on which the choice can be made responsibly, provided it is held with the confidence its evidence supports and the rest is left to precaution.

Rooms and Nations

John Searle’s “Minds, Brains, and Programs” (1980) is remembered less for its comparison of a modeled mind with a modeled fire than for a room. A man who knows no Chinese sits inside it with a book of rules written in English, and slips of paper covered in Chinese characters come in through a slot. The rules tell him, by the shapes of the characters alone, which char-

acters to send back, and his answers are good enough that people outside take the room for a fluent speaker. The man understands nothing of the questions or of his answers. Searle concluded that running a program cannot by itself produce understanding, since the man runs the program and understands nothing. To the reply that the whole room understands, the man being only one part of it, Searle answered that the man could memorize the rules and work everything out in his head, and he would still understand no Chinese.

Ned Block built a larger room out of a nation in “Troubles with Functionalism” (1978). He imagined the government of China arranging for its billion citizens, each with a two-way radio, to realize for an hour the functional organization of a single person’s mind, with each citizen playing the part of one unit in that organization and the whole connected by radio to an artificial body. If functionalism is true, the system would have the person’s mental states, pains among them, for the hour it ran. Block found *prima facie* doubt that it would have any states with a felt quality at all, and he took the doubt as a sign that functionalism counts too many systems as minds. Both arguments place the right organization at a scale and in a medium where the reader’s imagination finds no one home, and both ask the reader to trust that finding.

The finding is weaker evidence than it seems. The man in the room is one part of the system, the part that moves symbols, and his failure to understand shows that the part lacks understanding, as a single neuron’s ignorance of English shows nothing about whether a brain can speak it. When he memorizes the rules, the program runs in his head as a system his own understanding does not reach, and his ignorance remains evidence only about him. A room in which a clerk’s hand worked through rules rich enough to pass for fluent Chinese would answer more slowly than anyone could watch, and an intuition formed about a system no one could watch at work is an intuition about a picture of it. The nation invites the same doubt, since a reader who pictures citizens with radios finds nobody behind them, and a reader

who pictured a brain at the scale of its molecules would find nobody there either. Chalmers extended his own argument to a population, concluding that an appropriately organized population would have the experiences of the brain whose organization it shared, and he judged intuition an unreliable guide to what is possible in such cases. Searle and Block took their cases to tell against functionalism, and the replies leave that dispute open while showing that intuitions of this kind cannot close it.

A system can produce the right answers by a route that shares nothing with the organization of a mind, as a table that paired every possible question with a stored answer would, and a figure in a designed world that talks like a friend may be such a system. Behavior settles little about who is home, which is the room's lesson, so criteria for the protected region have to look past behavior to the organization that produces it. The nation teaches the other half of the lesson. If a nation organized like a brain would feel, then subjects can be built from subjects, and the beings of a designed world who compute minds of their own, or who are organized together into something larger, raise the question of whether a second subject has arisen above them. Where a system reproduces a brain's fine-grained organization, the room and the nation give little reason to doubt that someone is there, and where a system reaches a brain's behavior by tables or scripts, they give real reason. Their force falls on credences unevenly, and the rule that turns credences into protection has to register the difference.

Rivals

Anil Seth's "Conscious Artificial Intelligence and Biological Naturalism" (2025) challenges the assumption that computation provides a sufficient basis for consciousness. Seth argues that consciousness depends on the nature of living organisms, a view he calls biological naturalism and states as the claim that consciousness is a property of living systems only, though not neces-

sarily of all of them. On his account the predictive modeling a brain performs is bound up with the regulation of a living body, the work by which an organism keeps itself in being, and in such a system what it is cannot be pulled apart from what it does. A simulation of that work lacks the causal powers and the intrinsic properties of the work itself, so a computer running even a detailed simulation of a brain would not thereby be conscious. Seth concludes that real artificial consciousness is unlikely along the present course of the technology and becomes more plausible as machines become more like brains or more like living things. If his view is right, a creature computed on a conventional machine would feel nothing at any grain, however faithfully its life was carried, unless the machine that carried it was itself alive in the relevant sense.

Integrated information theory reaches a similar verdict by another road. Giulio Tononi and Christof Koch, in “Consciousness: Here, There and Everywhere?” (2015), hold that consciousness is a fundamental property of physical systems with particular causal properties, measured by how far a system as a whole, beyond its parts taken separately, shapes its own possible past and future states. The theory predicts that consciousness comes in degrees, that it is common among living organisms, and that it can occur in some very simple systems, while feed-forward networks, however complex, have none, and neither do aggregates such as heaps of sand. Tononi and Koch also state that a digital computer, even one whose behavior matched a person’s and even one running a faithful simulation of a human brain, would “experience next to nothing.” The reason lies in the machine. What the theory measures is the cause-and-effect structure of the physical parts, the transistors and their connections, and Tononi and Koch expect a conventional computer to break into many small complexes of low integration, given the narrow connections of digital circuits, whatever program it runs.

The denial covers every grain. A mind computed on conventional digital hardware gains nothing on the theory by being

computed more finely, at the level of molecules instead of neurons, since a finer program still runs on the same transistors, and the theory asks only what the transistors do to one another. It denies experience to an emulated person carried out of the primordial cosmos and to the natives of any designed world, shallow or deep, whose machine is conventional, whatever those beings say and do. It allows experience to a computed mind only where the hardware's own physical elements are integrated as it requires, a condition on how the machine is built that no program can meet or fail. Tononi and Koch even propose, as one way to test the theory, joining a neuromorphic circuit to a living brain to see whether experience changes when the circuit becomes part of the brain's main complex and not otherwise. Whatever credence the theory deserves, its denial is categorical, and a designer has to answer it.

Grain skeptics keep functionalism and move the organization that matters downward, to the metabolic or electrochemical activity of cells that a model of spikes and synaptic weights leaves out. Godfrey-Smith, whose objection to the replacement arguments has already appeared, holds that the activities of living cells, their self-maintenance, their control of their own boundaries, and the countless small interactions among their molecules, belong to the functional profile on which experience depends. He finds no reason to think that artificial systems built another way would have experience relevantly like human experience. If any such view is right, a mind carried at the grain of its neurons and synapses would not be the same mind as the one it copies, and it might be no mind at all, so the near horizon closes for the dying and for minds built in shallow worlds. Deep worlds stay open, since a native made of the world's own matter at full depth carries whatever finer organization that matter supports, provided the world's chemistry supports the organization that matters. The skeptics' doubt falls on grain and leaves computation itself standing, which separates them from Seth and from integrated information theory, whose doubts reach every grain.

Illusionism, the view Keith Frankish defends in “Illusionism as a Theory of Consciousness” (2016), holds that phenomenal consciousness, the felt quality that experience seems to have, is an illusion. Introspection represents certain states of the brain as having qualitative properties they lack, and what is really there are the states themselves, with properties that dispose introspection to misrepresent them, which Frankish calls quasi-phenomenal. François Kammerer argued in “The Normative Challenge for Illusionist Views of Consciousness” (2019) that the view must say whether value depends on phenomenal consciousness. If it does, and there is no such thing, illusionists face consequences for ethics that may be hard to accept, and if it does not, they owe a substantial account of what value depends on. A foundation that rests value on experience inherits the challenge, since it would seem to lose its ground if experience in the philosopher’s sense turned out to be an illusion.

The foundation survives the challenge in the form it already takes. Value is value for a subject, a being for whom things can go well or badly from its own point of view. Illusionism grants that such beings exist. They have states that represent their own condition and move them to avoid or seek, and they report those states, with all the confidence of a person describing a pain, as feeling a certain way. If the felt quality those states seem to have is an illusion, the states remain, and so does their place in a life, as a burn drives a creature from the fire and then troubles its sleep. The precautionary rule protects whatever the word *feeling* picks out, which on illusionism is that structure of states and their representation, and every protection attaches to it unchanged. Illusionism also tends to raise the chance that computed minds have whatever human beings have, since an illusion of that kind is itself a matter of organization, and a machine organized like a brain would represent its states as a brain does. The challenge exposes a temptation the foundation must refuse, the inference from a theory that denies phenomenal properties to the conclusion that nothing matters, which would remove the

reason to prefer any world to any other and the reason to prefer one's own relief to one's own pain.

Panpsychism, as Philip Goff defends it in *Galileo's Error* (2019), holds that consciousness is a fundamental and ubiquitous feature of the physical world. Physics, he argues, describes only what matter does, its dispositions and structure, which is Bertrand Russell's point about the silence of physics on the intrinsic nature of what it describes, and panpsychism proposes that the intrinsic nature behind the structure is consciousness. On Goff's view the most basic constituents of matter have forms of experience unimaginably simpler than any human experience, and complex experience arises from their combination in some systems, though not in rocks or tables, which have experiencing parts and no experience as wholes. If lifeless matter holds experience, every summary of a lifeless region would seem to erase it. Every saving that makes a world affordable would then fall under the protected region's prohibitions.

A designed world's summaries escape the pressure. On the Russellian form of the view that Goff defends, experience belongs to the intrinsic nature of whatever realizes physical structure, and in a designed world what realizes the structure of a lake is the machine. A summary changes which structure the machine computes, replacing many fine states with a few coarse ones, and it leaves the machine's own matter as it was, with whatever intrinsic nature it has, running one computation in place of another. Nothing panpsychism counts as a subject is ended by the change, since the computed molecules of a summarized lake were, on the view, only structure, and the machine's own particles go on existing whatever it computes. Panpsychism also denies that aggregates such as lakes and rocks are subjects, so a panpsychist who believes that experience pervades the machine has no reason to believe that summary ends anyone. The credence that summarizing lifeless matter ends a subject is therefore far smaller than the credence anyone gives panpsychism itself.

The Precautionary Rule

Precaution has so far reached the argument in pieces. It first appeared as a policy under which whatever might be conscious counted as a possible subject, and it acquired a measure when the asymmetry between the two errors was used to scale care to the probability that a system feels and to what it stands to lose. In the lake it gained an edge that no saving could cross, because of what a creature already living has left to lose, and the region inside that edge was drawn around fish, birds, and insects. Once the claims of subjects were examined, it divided into basic claims that stand equal for every subject and duties of aid that vary with probability and stakes, and the line of protection was named as the threshold that settles which systems are treated as subjects. Read in sequence, these statements can look like a rule that changed its shape, scaled in one place and absolute in another, though each states a part of one rule. That rule can now be given once and in full.

Every system that might feel is assigned, by a published procedure and usually through an assessment of its kind, a credence that it feels. A system whose credence stands at or above the line of protection is treated as a subject in full. It holds every basic claim, not to be made to suffer and not to be destroyed against its interest, and every protection of the protected region, so that it is computed in full whenever it is computed, computed once, and never summarized or rolled back, and none of these protections varies with its credence. Duties of aid toward any system, at any credence, are scaled to the probability that it feels and to what it stands to lose. Below the line of protection, summary, rollback, selection by early death, and the other acts forbidden above it are weighed against their costs by the same two measures, and only suffering imposed for no purpose is ruled out at every credence. Every credence, the procedure that assigns it, and the line of protection itself are open to review, and any doubt about where a system stands is resolved toward the protected region.

The protections that hold in full above the line of protection hold for two reasons. The basic claims take no discount for the reasons that made them equal, since they constrain what is done to a being, and carrying a being commits its keeper to continue. The rules of the protected region take a threshold's form because summary, rollback, a speculative run that is later erased, and destruction are acts that a designer performs on a system or refrains from, and none of them can be done by degrees, since a fish's life cannot be half summarized or a stretch of it half undone. A rule for acts of that kind under uncertainty must either set a threshold or trade each act against its savings case by case, and for a being treated as a subject that trade is what the protected region forbids. Aid comes in amounts. A machine can spend more on a creature's behalf or less, compute its surroundings at a finer grain or a coarser one, study it closely or barely, and carry it sooner or later. So duties of aid vary continuously with the chance that the aid reaches someone and with what it gives the one it reaches, whatever the credence. The line of protection therefore rules out for subjects the acts that admit no degree, leaves those acts to weighing below it, and lets aid scale everywhere, which is how a single rule can be absolute in one place and scaled in another. It also refuses a discount for doubt above the threshold, for the reason that refused a discount for status, since a designer who could weaken a being's protections by making its inner life harder to read would be rewarded for opacity.

The rule takes its form from the approach to precaution that Jonathan Birch set out in *The Edge of Sentience* (2024). Birch holds that people ought at least to avoid causing gratuitous suffering to sentient beings, that it is reckless to create risks of suffering for a sentience candidate without asking which precautions would be proportionate to those risks, and that judgments of proportionality should be informed, democratic, and inclusive. He suggests citizens' panels that ask whether a proposed precaution is permissible in principle, adequate, reasonably necessary, and consistent with how comparable cases are treated, and he asks that

their decisions be revisited as evidence changes. A sentience candidate, in his definition, is a system for which there is evidence implying a realistic possibility of sentience that it would be “irresponsible to ignore” in decisions affecting it, evidence rich enough to identify risks to its welfare and to design and assess precautions. A system that falls short of that standard is a priority for investigation when further study could plausibly make it a candidate and human activity affects it in ways that might call for precautions. Birch counts the judgment that a system is a candidate as mixed, partly scientific and partly evaluative, and he requires a substantial base of positive evidence, holding that an inability to rule sentience out is too low a bar. He offers a procedure and says plainly that it is not a decision rule, and when he asks how far a precaution must reduce a risk, he rejects any threshold fixed for every case, observing that no one agrees on precise probabilities at the edge of sentience.

A designer’s machine needs more than a procedure, because it decides, for every creature it computes, whether to carry that creature in full, and it decides continuously and at a scale no panel could follow case by case. The rule keeps Birch’s structure and adds a stated threshold, the line of protection, a probability of feeling at and above which a system is treated as a subject in full, modeled on his sentience candidate. The figure makes the rule checkable, since anyone can ask whether a kind’s credence was assessed by the published procedure and whether it meets the threshold, and it makes every error traceable to the credence that produced it. Deliberation keeps the place Birch gives it, since deciding how much risk of the graver error to accept for the cost of the lesser is a judgment of value that belongs to those who bear the risks and to those who will live under the rule. A stated threshold is also a commitment that can be held against whoever applies it.

Pure expected value would replace the rule with a single calculation. Every possibility of feeling would count in proportion to its probability and to what is at stake, summed across every sys-

tem that might be affected, and the designer would choose whatever made the sum largest. A sea holds microbes in numbers no census could reach, panspsychism adds a credence for lifeless matter, and on some theory the machine itself might feel. Under pure expected value those tiny probabilities, multiplied by vast numbers, would dominate every decision, so no summary could be permitted, no lineage could be left to selection, and the dying, whose capacity to feel no one doubts, would be outweighed by speculation about everything else. Hayden Wilkinson defends the principle behind these results in “In Defense of Fanaticism” (2022). He calls fanaticism the judgment that for any modest value that is certain and any probability however small, some value is large enough that a gamble with that small probability of it is better, and he argues that refusing fanaticism costs more than accepting it. A view that refuses it faces, on his account, three choices between two costs, and it must bear at least one cost of each pair. The first pair sets declining some trades of a little certainty for a great gain in what is possible against accepting a series of such trades that leaves it worse off. Another sets inconsistency between decisions at large stakes and small against an absurd sensitivity to tiny differences in probability. The last sets verdicts that depend on remote events nothing one does can affect, events in ancient Egypt among them, against verdicts that depend on uncertainty about those events, which can rank gambles against the way the view knows it would rank them after learning more.

The precautionary rule takes on some of those costs openly. At the line of protection a small difference in credence makes a large difference in protection, which is the sensitivity to small differences in probability that Wilkinson counts against views like it. The rule accepts the cost for reasons that belong to the work it does. The credences that would dominate under pure expected value are not measurements, since no agreement exists on precise probabilities of feeling, and multiplying an unmeasured number by an enormous one turns an artifact of how theories are

counted into a command. The rule must also be public and open to contest, and a rule that any unfalsifiable possibility with enough at stake could capture would hand the keeping of a world to whoever names the largest possibility. Below the line of protection the weighing counts numbers as well as probabilities, so a precaution that spares a billion worms counts for more than one that spares a single worm. A possibility of feeling that rests on nothing beyond the impossibility of ruling it out can justify precautions that cost almost nothing, and no count of the systems that share it can turn it into a claim against beings whose feeling the evidence supports. The ground for that limit is Birch's reason for requiring positive evidence before candidacy, carried below it, since a bar set at what cannot be ruled out would let in everything.

A bare threshold of credibility would go wrong in the other direction. Under such a rule everything above a threshold would be protected and everything below it ignored, so the threshold would carry the whole of every decision, and a kind placed just below it by an error of assessment would receive nothing at all. A worm below the threshold could be made to suffer for amusement, and a lineage nearing the threshold would go unwatched until the day it crossed. The scaled care the rule keeps below the line of protection prevents each of these failures. Nothing that might feel is made to suffer for no purpose, precautions that cost little beside the chance and the stake are taken, kinds near the line of protection are studied, and a kind misplaced below it loses part of its care and keeps the rest. The line of protection decides who is treated as a subject, and the care below it decides how everything else that might feel is treated, so that care falls away by degrees on the far side of the threshold.

The asymmetry between the two errors sets the direction in which the line of protection should err. Treating a populated world as an empty model is the graver error, because it falls on beings able to suffer, so the threshold belongs well below even odds, where a designer accepts carrying many creatures that may

feel nothing for the sake of those that do. The lesser error has victims too. On a finite budget, whatever the machine spends carrying in full a creature that feels nothing could have carried someone who does, and while the dying can be carried, the cost of the lesser error falls on people who would otherwise end. A line of protection set as low as doubt reaches would spend lives to protect bare possibilities. The asymmetry tells the designer to set the threshold low, and the budget tells them where low enough stops.

The Line of Protection

The line of protection sits at a credence of one in ten. A system belongs to the protected region when the best available evidence, assessed by the published procedure, gives at least one chance in ten that it feels, and the figure is offered as a reading of the realistic possibility that Birch's standard for a sentience candidate requires. Birch attached no figure to his standard, and in advising how experts should state probabilities to citizens' panels he judged unwarranted the government yardstick's demand for 40 percent before an outcome may be called a realistic possibility, and he doubted that one in twenty is remote. A threshold of one in ten sits within the range those remarks leave open. A chance of one in ten is one that people act on when a doctor or a forecaster names it, and it is low enough to keep the direction the asymmetry sets without spending the budget on every bare possibility. The figure takes the first half of Birch's standard and sets the second aside. An evidence base rich enough to design precautions is needed for precautions that must be designed, and the protections of the protected region, carrying a creature in full and leaving every stretch of its life in place, need no further knowledge of its welfare to apply.

The line of protection must sit low enough to keep fish, birds, and insects inside the protected region, since the asymmetry forbids leaving outside any kind whose feeling the evidence leaves

seriously open. The evidence for insects, the weakest of the three, includes experiments Birch discusses in which bees trade a richer reward against a hotter feeder when choosing which to approach, and Birch counts adult insects among his sentience candidates. A threshold near even odds would demand that a kind's feeling be established as likely, which the evidence for insects has not yet done. Kinds whose only case is that feeling cannot be ruled out have to stay below the threshold, among them the plants, fungi, and microbes, whose protection in full would grow with every generation until it had spent the budget and the means that could carry the dying. The threshold must also clear any credence that panpsychism lends to the feeling of inert matter, measured as the credence that summarizing lifeless matter ends a subject, since a line of protection beneath that credence would bring lifeless matter into the protected region and forbid every summary. For the reasons already given, that credence falls far below one in ten even for someone who takes panpsychism seriously. One in ten meets all three conditions.

At one in ten the fish in a designed lake stay inside the protected region, each carried in full in its halo of finely computed water for as long as it lives, and so do the birds over the lake and the insects on its surface. Kinds near the line of protection can cross it on a small change in the evidence, and the rule resolves doubt about a kind's placement in favor of the protected region while the doubt lasts. A kind assessed at one in twelve may be carried in full if the assessment itself is uncertain, since the threshold applies to the credence the evidence supports, and an assessment that cannot tell one in twelve from one in eight leaves the question open. Drawing the region generously in this way costs the budget something. The budget's claims were heard when the figure was set, and they are not heard a second time against each kind at the margin.

The figure is a judgment of value, how much risk of the graver error to accept in order to limit the cost of the lesser. Birch's argument that conflicts of value in a democratic society should be

settled democratically, since no group can claim expertise in the relative importance of the values in conflict, applies to it fully. In the primordial cosmos the figure belongs to legislatures informed by deliberative bodies such as citizens' panels. The placement of each kind against the figure is a judgment of evidence, and it belongs to independent review of the kind the British government commissioned in 2020, by scientists and philosophers who publish their criteria, their confidence, and their grades so that anyone can check their work. In a designed world both judgments are written into the keeper's policy, the published rules by which a keeper protects possible minds wherever no law of physical properties can, since no law of a world may test for experience. The keeper's policy is fixed before a world begins, open to contest by those it governs or by trustees acting for them, and revisable by a procedure that is itself published, so the figure can move as deliberation changes and a kind's placement can move as evidence does.

Below the line of protection, nothing that might feel is made to suffer for no purpose, a bar that costs nothing to keep and binds at every credence. Summary and selection are permitted, and the designer weighs, kind by kind, the chance that a kind's members feel and what they stand to lose against what care for them would cost, above all in means that could carry the dying. Every precaution whose cost is small beside that chance and that stake is taken, such as choosing, among the ways a lineage's selection could be written, one that imposes less of whatever in it might be suffering. Kinds within reach of the line of protection are priorities for investigation in Birch's sense, studied for the evidence that would move them and watched in the keeper's policy as their lineages change. Duties of aid toward systems below the line of protection scale with probability and stakes as they do everywhere, and the weighing that sets them counts the many as well as the likely. Each of these forms of care leaves a record in the keeper's policy, so care below the line of protection is as open to inspection and contest as the protection above it.

Umāsvāti's *Tattvārtha Sūtra*, composed sometime between the second and fifth centuries and translated by Nathmal Tatia as *That Which Is* (1994), grades living beings by the senses they have. Beings whose bodies are earth, water, fire, or air have touch alone, and so do plants, while worms add taste, ants add smell, bees add sight, and human beings and many other animals add hearing. The vow of nonviolence takes two forms. The ascetic's great vow renounces violence toward every living being, those with a single sense included. The householder's small vow, the partial form suited to a life in the world, renounces injury to beings with two senses or more, as the commentaries explain it, and among the householder's further restraints is abstaining from harm that serves no purpose. The grading of beings by their senses leaves the protection ungraded, since each vow renounces injury to everything inside the circle it draws.

The tradition is a precedent for two features of the rule that can look incompatible. It draws its circle of protection generously, around beings with a single sense, a circle wider than any a designer could now afford. It grades beings by their capacities while refusing to grade the protection owed to those inside the circle, so that the grounds of protection come in degrees and the protection does not. Its two vows also show a boundary of full protection set by what a way of life can bear, with the ascetic holding the widest circle and the householder a narrower one while still refraining from purposeless harm. A designer's machine stands closer to the householder than to the ascetic, since it spends from a finite store, and the line of protection marks where full protection ends because the means that could carry the dying are limited. What the vows asked of conduct, a machine has to carry out in computation, in what it carries in full and what it summarizes.

Architecture and Grain

Integrated information theory's denial is a claim about machines that a builder can act on. A mind's organization, at any grain, can be the same on two machines that differ in how their physical parts are wired, and on the theory only the wiring decides. A designer who wants a computed mind to be conscious on that theory as well as on functionalism can run it on hardware whose own elements are densely and reciprocally connected, so that the physical system itself has the integrated structure the theory requires. Neuromorphic hardware, whose physical circuits are designed to behave like neurons and synapses, is one place to begin, though whether any present chip has the integration the theory requires is an open question, and the test Tononi and Koch proposed presupposes that a circuit of that kind could become part of a conscious complex. The rule a builder can meet is to carry a mind intended as a subject on hardware integrated in the theory's sense wherever that can be done, and the conditions for consciousness, when the argument reaches them, inherit it. Meeting the rule may cost more than conventional hardware, and whatever it costs is counted against the same means that could carry the dying.

The theory's doubt counts all the same where the rule is not met. If the theory is true, a mind computed on conventional digital hardware feels almost nothing, so the probability that such a mind feels is lowered in proportion to the credence the theory deserves. The denial belongs to the theory's fundamental claims, which Birch distinguishes from its testable claim about where consciousness sits in the human brain and counts as legitimate speculation that cannot yet be called empirically supported, so the credence it deserves is modest. Its standing in the rule is double. It is a condition on the machine, which a builder who can meet it should meet, and it is a reason for doubt about every computed mind that runs on hardware that fails it. The doubt cannot by itself carry a mind organized like a person's below the line of protection. For a mind carried at a fine enough grain, that would

require the theory, together with every other view that denies such a mind experience, to deserve more than nine chances in ten, and no fair reading of a field in which each of those views is contested comes near that. A mind emulated on a conventional computer stays inside the protected region, holding every basic claim and every protection, while the duties of aid it is owed, which scale with the probability that it feels, are reduced by the doubt the theory adds.

Functionalism leaves open which level of a brain's organization is the one that suffices for experience. Chalmers's criterion for the relevant level, one fine enough to fix a system's behavioral dispositions, splits the question into two parts. One is empirical, how fine a copy of a brain must be to behave exactly as the brain would, and the other is a residue the replacement arguments narrow without removing, whether experience could depend on detail finer than any that behavior reveals. Even the smallest nervous system shows how hard the empirical part is. Birch notes that the nematode worm steers toward an attractive odor by computation inside a single neuron, one part of whose axon tracks the sweep of the head while another tracks the strength of the odor. A full copy of the worm, he concludes, would have to reach below the level of whole neurons to the arrangement of structures within them. The ladder of grains that Anders Sandberg and Nick Bostrom laid out for a human brain, from functional models down to molecules, marks the possible answers, and nobody yet knows on which rung experience stops depending on what lies below.

The rule against making a deep world's people shallow rests on that ignorance. A designer who knew the level at which experience stops depending on finer detail could carry each native at that level and save everything beneath it without loss, so the prohibition of the middle path is a prohibition under ignorance, and it would lapse if the ignorance ended. While it lasts, a native carried at a coarser grain than her world's depth may be the same subject, a different subject with her memories, or no one, and

nothing available can tell the cases apart. The line of protection gives the practice no cover even when the coarsened native would still stand above it. The protected region's rules forbid carrying a protected creature at reduced detail once its life is under way, because the reduction may end the subject who was there and put a stranger, or nobody, in her place, and a high chance that someone remains is no license for a real chance of ending someone. A shallow world built openly faces only the lesser error, since its natives are coarse from the start and either feel at that grain or do not, while the middle path risks the graver one.

If the Hypothesis Fails

The structural account of what makes a world real holds whatever theory of mind is true, since its marks concern how a world's events produce one another and none of them mentions experience. The moral conclusions sort less simply, and the sorting decides how much of what is owed to computed beings rests on functionalism. Some conclusions hold on any theory of mind. That value is value for a subject, that every subject holds the same basic claims, and that duties of aid vary with probability and stakes are claims about subjects wherever they are, and they bind in the primordial cosmos whatever turns out to be true of machines, toward its animals as toward its people. The precautionary rule belongs with them, because it is a rule for acting under doubt about who feels and so presupposes no answer to the doubt. The rules against deception hold too, since anyone who visits a world from outside is a subject in the primordial cosmos, and a painted sky deceives her as it would deceive a native.

A second group of conclusions rests on some theory that allows computed beings to feel. Everything owed to the natives of designed worlds belongs here, along with everything owed to emulated minds and the duty to carry the dying into worlds where they can go on living. These conclusions require only that a computed being's chance of feeling reach the line of protection,

and for a being organized like a person and carried at a fine enough grain, the credence that functionalism alone is true carries it there easily. Wherever these conclusions hold, they hold in full, so a native of a designed world holds every basic claim and every protection whatever its credence above the line of protection. Duties of aid toward computed beings scale with the credence, so the duty to carry the dying, which is a duty of aid, weakens as the credence that a carried mind feels falls, a dependence the argument about rescue has to carry. A built mind that copies no one belongs here too, though the replacement arguments never reached it, so its credence rests from the start on what the leading theories say about its organization.

Should functionalism fail, or hold only at a grain finer than any emulation reaches, these conclusions change form as long as some view still allows computed subjects. Integrated information theory would confine every protection of computed beings to minds carried on hardware integrated as it requires, and the rule a builder can meet would become a condition on every world. The grain skeptics' view would close the near horizon for rescue and for minds built in shallow worlds, while deep worlds whose chemistry supports the organization that matters would keep their natives, at a cost that puts them far beyond any machine resembling those that now exist. Biological naturalism would let a computed being feel only on a substrate that is itself alive in the relevant sense, so designed worlds carried by any machine now conceived would hold no one. A built mind that copies no one is the case on which the rivals differ most. On integrated information theory it may feel on the right hardware whatever its organization, the grain skeptics require that its organization reach their grain, and biological naturalism requires that it live.

If no theory that allows computed subjects survives, designed worlds hold no one. A designer would then build structures that bear every mark of reality and contain nobody, the duty to carry the dying would have nowhere to carry them, and the protections written for natives would govern empty rooms. The struc-

tural account would go on saying truly that those worlds are real, the duties of the primordial cosmos would bind as before, and the rules against deception would protect anyone who visits. That outcome is possible, and the precautionary rule has already priced it, since the cost of treating an empty world as populated is the lesser error, paid in means that could have done good elsewhere. The argument proceeds as though computed beings can be subjects because the credence that they can is high and the error of assuming otherwise is the graver one, and every later conclusion that leans on the hypothesis leans on it in one of the ways just sorted.

Provisional Criteria

Until better evidence refines the edge of the protected region, four criteria place a kind at or above the line of protection, and any one of them suffices. The first asks for an integrative brain, or a structure that does the work of one, together with behavior showing that the kind represents what it stands to gain and lose. Trade-offs between a threat and a reward made in prospect, flexible care of an injury, learning beyond habituation and sensitization, and the seeking of relief when hurt are marks of that kind, and the 2021 review graded them all. Kinship is the second. A kind whose nervous system shares the architecture of studied kin that meet the first criterion stands inside the protected region until study shows otherwise, so that doubt arising from neglect goes toward inclusion, which was also the effect when the law took in every decapod. Built systems, whose architecture is on record, fall under a third criterion, which places inside any system that has in full the indicator properties of more than one of the leading scientific theories of consciousness, in the computational terms in which Butlin and his colleagues stated them. A fourth places inside any emulation that reproduces the behavior of a kind already inside, and Birch argues for this step in the case of a

complete neuron-by-neuron emulation of an insect, a fish, or a mouse.

The criteria keep fish, birds, and insects inside, as they must. Fish and birds have brains with well-studied integrative regions, and the behavioral evidence of feeling in both is extensive, while bees meet the first criterion through trade-offs made in prospect and through mushroom bodies, integrative centers linked to learning and memory. Each of the other criteria asks for evidence at least as strong as what places the bees inside. A close relative of a kind inside shares much of that kind's evidence, a built system that meets several leading theories in full feels if any of them is true and the working hypothesis holds, and an emulation that reproduces a kind's behavior keeps the organization the replacement arguments protect. Decapods and cephalopods come inside by the first criterion or the second. The penaeid shrimps come inside by kinship, though Birch counts them as priorities for investigation, since their brains hold miniature counterparts of the integrative regions of other decapods, and the work that found those regions cast doubt on their function without showing that they do no work. Kinds without nervous systems, among them plants, fungi, and microbes, meet none of the criteria, and they stand below the line of protection whatever their complexity, since their case for feeling rests on the impossibility of ruling it out. A system that produces a mind's behavior by a table or a script meets none of them unless its organization does, and since the features that make a figure good company may be the ones that make it a subject, a convincing figure is weighed against the line of protection by its organization, with its behavior as evidence.

The adult hermaphrodite of the nematode *Caenorhabditis elegans* has some three hundred neurons, about 185 of them in a ring in its head that researchers increasingly call a brain, and that small brain integrates information from several senses, supports learning and decision, shows coordinated activity across the whole of it, and has a form of sleep. Offered an attractive odor on

the far side of a barrier that would dry it out, a worm crosses only when it is hungry enough, which looks like a trade-off. Birch argues that the appearance misleads. The mechanism has been mapped at the level of neurons, and the behavior most likely comes from hunger working through a single inhibitory connection between a neuron that detects the threat and a neuron that registers the worm's internal state. It is then a direct modulation of one response by another and shows no representation of what the worm stands to gain or lose. He counts nematodes as priorities for investigation, open to reclassification if the evidence changes.

Under the provisional criteria the nematode stands below the line of protection. It meets the first criterion only if its trade-off shows evaluative representation, which the mapped mechanism gives reason to doubt, it has no studied kin inside the protected region whose nervous systems share its plan, and the other two criteria do not apply to it. Worms in a designed world may be left to summary wherever no one perceives or investigates them more finely, and selection may run among them. Scaled care still governs them. Nothing may make them suffer for no purpose, precautions that cost little are taken on their behalf, and the evidence about them is watched, so that a finding of evaluative representation in a nematode's brain would move the kind inside at once. Because the worm learns, honesty also governs how it is drawn from a summary, and a worm supplied where a lens turns on the silt must carry no association it never formed, which marks one of the few places where the edge of the protected region and the boundary that honesty sets come apart.

The criteria are provisional and built to err in one direction. They place a kind inside on a single convincing trade-off made by an integrative brain or on resemblance to a kind already inside, and they place a kind about which nothing is known beside its nearest studied relatives. Better study of animal minds will refine them, as will minds built with architectures no animal shares, and each refinement will move the edge outward whenever evid-

ence warrants. What they decide now is enough to tell a machine what to carry in full on the first day of a world. Everything they place inside is computed whenever it is computed, once and in full, and everything they leave outside is held in a summary that the evidence can still reach.

First Light

At first light the lake in the carpenter's world is still, and the machine beneath it is doing what the rule requires. Most of the water is a few slow currents and a field of temperatures, cold at the bottom and warming at the surface as the sun clears the ridge, and none of it is computed more finely than anyone needs. Around each fish the water thickens into detail, molecule by molecule at the gills and thinning outward to the distance at which the fish can feel a pressure wave along its flank, and each halo moves with its fish through the reeds. A heron stands in the shallows with its own detail around it, and the first midges over the water are carried in full, each one, for as long as it lives. In the silt below, the worms exist as a density in the summary, a number for each patch of mud. Their kind stands below the line of protection on the evidence now in hand, and the keeper's policy records what finding would move it.

Later in the morning someone may kneel at the edge with a glass and turn it on the mud, and the machine will compute the worms beneath the glass as lawful draws from the summary, with no memory of anything that did not happen to them. The silt holds far more of them than the shallows hold fish. On the evidence the chance that any of them feels is smaller than the chance for the midges and far smaller than for the fish, though it is not nothing, and the machine spends on them what that chance and their small stakes are worth. If evidence ever moves their kind across the line of protection, the machine will carry every worm from then on as it carries the fish. The light reaches the far shore first and moves across the water toward the reeds.

As it comes the fish begin to rise, each in a halo of water carried molecule by molecule, because a stated rule, applied to evidence anyone can check, found a serious chance that someone is there.

Commitments

Chapter numbers in italics refer to chapters not yet published.

These are the commitments the argument of this book reaches, gathered in the order of the seven questions set out in the Introduction. Each is argued in the chapter or chapters named beside it, and where the wording here and a chapter differ, the chapter governs.

What Makes a World Real

A world counts as a reality when it is a lawful order of events that act on one another by their own powers, when it holds together whatever the minds inside it believe, wish, or attend to, and when a different past would have produced a different present. (Chapter 1)

Such a world is fit to be lived in once many subjects can meet in it and beings made of its matter can act on it and be acted on in return. What makes it matter is that someone is there. (Chapter 1)

Nothing in these marks and capacities concerns what carries a world, so a computed world that bears them is as real to those who live in it as ours is to us. The substrate is irrelevant to whether a world is real and central to which worlds can be built. (Chapters 1 and 2)

Economy differs from fakery by consistency with everything that has gone before, whatever the quantity of computation. (Chapter 1)

A world's reality rests on its own structure, which is why leaving its foundations undisclosed deceives no one, while falsehoods built into it, such as painted skies and planted histories, deceive everyone they reach. No sky may hold stars that do not exist, and figures that seem to be people may fill a street only where their nature is disclosed to everyone who meets them. A life is built around other people, and trust spent on no one is lost. (Chapters 1 and 4)

A failure in a world is judged by whether a subject could reach it, rely on it, or be misled by it. (Chapter 1)

A word for a use, such as *model* or *simulation*, says nothing about whether anyone is home, so the question of what a run now holds is asked again each time the run grows richer. (Chapter 1)

The designer's responsibility attaches to the act of instantiation, whatever the metaphysics of unrun structures turns out to be. (Chapter 1)

Once two runs of one world diverge, nothing done to the inhabitants of one is excused by the existence of the other. (Chapter 1)

A history of beings who suffered is replayed only with the caution owed to a possibility of suffering repeated. (Chapter 1)

What the Machine May Spare

The good of the inhabitants chooses the physics, and the machine's preferences enter only as costs. (Chapter 2)

Depth is the first decision of substance the designer makes about the machine. A designer may build a shallow world openly and

may not build a deep world and then quietly make its people shallow. (Chapter 2)

The expense of depth lies in growing a world far more than in keeping one, and on nearly reversible machines it is paid in waiting, hardware, and the design of the physics more than in energy. (Chapter 2)

The machine beneath every world is finite, and its limits fall on lives. Each saving frees means for some other use and also lands somewhere in the world it is taken from, so savings are permitted only where nothing has a serious chance of feeling. (Chapter 2)

Whatever inhabitants perceive or investigate is carried at the grain at which they perceive or investigate it. (Chapter 2)

Savings on lifeless matter are permitted wherever no one perceives or investigates it more finely, and so are savings on simple life that has no serious chance of feeling, wherever the weighing allows them. When the means they free could carry the dying, those savings may be owed, in the way a rescue may be owed by anyone able to perform it at small cost. A duty to save in order to create would be far harder to establish, and the book claims only the first. (Chapters 2 and 13)

A life that is summarized, recomputed, or undone has been taken from someone who never agreed to lose it. Whatever has a serious chance of feeling is therefore carried in full, computed once, and never rolled back, and a reversible machine's undoing stays within each step. (Chapters 2 and 8)

Chance at the fine grain, joined to coarse descriptions that evolve exactly as the averaged laws require, makes summary an honest part of the physics, provided the chance renews everything the summary discarded before the matter returns to view. Nothing the machine generates from a summary may carry the record of experiences that never occurred, and a reversible machine pays

for every random bit its world consumes as an erasure. (Chapter 2)

Deferring a peopled region saves nothing and is rarely possible at all, since influence runs both ways, and leaving such a region uncomputed is a decision about creation. (Chapters 2 and 12)

Where a physics economizes, as collapse does on a classical machine, the economy is keyed to physical interaction and never to minds, since laws that waited on consciousness would make nature a judge of moral standing and build every misjudgment into physics. (Chapter 2)

Locality in a world's laws is a gift to the machine that carries it. (Chapter 2)

A uniform slowdown of the substrate is invisible to natives, and lag appears only at interfaces, where minds running on one clock meet a world running on another. On a reversible machine a world's pace is part of its budget, since each of its years costs less energy the more slowly it runs. (Chapter 2)

Reliability is a structural good, and a fault is a breach of a world's laws that nobody performed. (Chapters 2 and 6)

Who May Write a World

A designer who selects a value knowing its consequences takes up responsibility for all of them. (Introduction)

No one who holds the power to write a world can be trusted on character alone. The virtues such power demands cannot be verified in an institution or an artificial mind, and the first designers will probably be institutions working with artificial minds. The standards that govern design bind human, institutional, and artificial designers alike, and each virtue a designer would need is matched by a procedure others can

inspect, such as independent review or complete records kept beyond the designer's reach. (Chapter 3)

A party is unfit to design if it is evidently moved by a motive that only the inhabitants can satisfy, such as their worship, their obedience, their display before an audience, or their use as subjects of research, whatever justifying reasons it can also cite. (Chapter 3)

Responsibility is assigned in advance by role. The designer answers for a world's laws and initial conditions and for what they do to the lives lived under them, a world's keepers answer for the machine that carries those laws, for its fidelity to them, and for the world's custody over time, and funders answer for the conditions they attach to their money. Every decision of consequence is recorded with the name of the party who made it. (Chapters 3 and 24)

One party may hold both the designer's and the keeper's roles only under external checks, among them independent review of every change to a running world's laws, records kept beyond that party's reach, and inspection by a body that answers to someone else. (Chapters 3 and 24)

Any change to the laws of a running world is an act of design, whoever makes it. The inhabitants live under the physics actually computed, so whoever alters that physics alters their laws, and the designer answers for any gap between the physics written and the physics computed. (Chapters 2 and 3)

An artificial designer acts with human ratification until institutions exist that can check it. Its ratifiers are independent of it and of whoever profits from its work, and they test its claims by their own means. Ratification is given or refused decision by decision, before anything runs, and it answers to those the world will hold. (Chapter 3)

If an artificial designer might feel, it is itself a possible subject, so its making falls under the conditions on creation and the limits on writing a mind. Its ratification binds only what it does to others, under published rules fixed in advance that it can contest before an independent body, and never works through a threat to its existence. (Chapters 3 and 9)

A human designer carried into a machine is checked as an artificial designer from the transfer on, while keeping every standing a person has. (Chapter 3)

No law of any world tests for experience. Protections that must pick out minds live in the keeper's policy, which is published and can be corrected, and every protection of minds that cannot be written as a law of physical properties lives there. (Chapters 2, 24, 96, 100, and 119)

A world's laws owe their inhabitants the virtues of good legislation. They are general, discoverable, prospective, clear, consistent, possible to live under, stable, and carried as written, because those who live under laws they did not make must be able to plan their lives around them, and a physics gives up any of these virtues only for a stated reason. (Chapter 30)

The living whom a world is built to receive take part in drafting its laws before they are written, and what they decide overrides nothing owed to any subject, whether to themselves or to beings who cannot be asked. (Chapter 16)

No world's lives are sold as spectacle, and no world is sold by tier. (Chapter 27)

Who Has Standing

Conscious experience is the ground of all value. Nothing is good or bad except for some subject, and where no subject exists, nothing is good or bad for anyone. Once a subject exists, its good reaches beyond its experiences, to real relationships,

achievement, the authorship of its own life, knowledge of its world, and freedom from domination. (Introduction, Chapter 4)

Moral standing extends to every being organized so as to have experiences, whatever it is made of, whether it evolved or was designed, and however many layers of computation separate it from the primordial cosmos. (Introduction, Chapter 4)

Standing holds across the gaps in a subject's experience, so a sleeper, a patient under anesthesia, and the inhabitants of a paused world keep their standing while nothing at all is felt. (Chapter 4)

Every subject holds the same basic claims, not to be made to suffer and not to be destroyed against its interest. A pain is bad by being bad for the one who feels it, and no standpoint exists from which one subject's pain could count for less than another's, so these claims do not vary with a being's degree of consciousness, kind, intelligence, origin, or the material that carries it, and nothing a being has done reduces them. (Chapter 4)

Duties of aid, rescue first among them, draw on finite means and can vary in strength with the probability that a being feels and with how much it stands to lose. Duties owed for what a party did to a being, by creating it, harming it, or taking it into its keeping, take no discount for doubt in any being treated as a subject. (Chapters 4 and 5)

Merit moves no one up the order of rescue, and wrongdoing moves no one down it. Suffering that a world's conditions impose as punishment is structural suffering. (Chapters 4 and 20)

Functionalism is the book's working hypothesis, on which consciousness arises from a system's functional organization, and it says nothing about which level of a brain's organization carries experience. (Introduction, Chapters 2 and 5)

The account of what makes a world real, the foundation of value, the equal basic claims of subjects, and every duty owed in the primordial cosmos hold whatever theory of mind is true. What is owed to computed beings rests on some theory that allows them to feel and needs only that a computed being's chance of feeling reach the line of protection, while duties of aid toward them, the duty to carry the dying among them, scale with that chance. (Chapters 1 and 5)

Integrated information theory's denial of experience to minds computed on conventional digital hardware is treated as a condition on the machine, which a builder meets by carrying a mind intended as a subject on hardware whose own causal structure is integrated, and as a doubt that lowers the credence that a mind on conventional hardware feels without carrying a mind organized like a person's below the line of protection. (Chapters 5 and 94)

Of the two errors, treating a populated world as an empty model is graver than treating an empty world as populated. On a finite budget the lesser error has victims too, since whatever is spent carrying a system that feels nothing could have carried someone who does. (Chapters 1 and 5)

Under uncertainty about which systems are conscious, the sound policy is precaution, which treats any system that might be conscious as a possible subject. (Introduction, Chapters 1 and 5)

No one can yet tell for certain which systems feel. A stated threshold of credence, called the line of protection and set at one chance in ten, marks what counts as a serious chance of feeling. Systems at or above it hold every basic claim a subject holds and every protection of the protected region, whatever their credence, and care for those below it is scaled to the probability that they feel and to what they stand to lose. (Chapter 5)

The line of protection is set by public deliberation and published so that it can be checked and corrected, and independent review of the evidence places each kind against it, with both judgments recorded in the keeper's policy. It is set low enough that fish, birds, and insects stand above it while their inner lives remain in doubt, and high enough that no credence panpsychism gives to inert matter reaches it. (Chapters 2 and 5)

Below the line of protection, nothing that might feel is made to suffer for no purpose. Summary and selection are permitted there under a weighing of the chance that a kind feels and what its members stand to lose against what care would cost, every precaution whose cost is small beside that chance and that stake is taken, and kinds within reach of the line of protection are studied and watched. (Chapter 5)

Precaution follows a stated rule of proportionate precautions, informed by public deliberation and bounded by the line of protection. It refuses pure expected value, under which tiny probabilities multiplied by vast numbers would decide everything, so a possibility of feeling that rests on nothing beyond the impossibility of ruling it out justifies only precautions that cost almost nothing. (Chapter 5)

Until the edge of the protected region is refined, a kind stands inside it if it has an integrative brain, or a structure that does that work, and behavior showing that it represents what it stands to gain and lose. A nervous system that shares the architecture of studied kin already inside suffices as well, as do the full indicator properties of more than one leading theory of consciousness in a built system and an emulation that reproduces the behavior of a kind already inside. (Chapters 5 and 96)

How Lives Are Judged

Lives are judged by four kinds of good. These are the experiential goods of how a life feels, the relational goods that exist between

beings, the agentive goods of beings who act and author their own lives, and the structural goods, such as safety, continuity, and freedom from domination, that secure the rest. No being's structural goods may be given up to raise the experiential goods of others. (Chapters 4 and 6)

Outcomes are ranked by the goods and harms for subjects they contain, so nothing that is good or bad for no subject, such as unseen beauty, counts in the ranking. Between outcomes that hold the same subjects, one that is better for some and worse for none is the better, and the form the ranking takes across different subjects is set when worlds of different sizes are compared. (Chapters 4 and 12)

The ranking tells a designer which of two worlds is the better to make and licenses nothing against the subjects already living in either, so no gain in it justifies destroying anyone. (Chapter 4)

Hardship divides along the line of choice. Structural suffering, the deprivation or degradation a world's conditions impose on beings who never chose it and cannot leave it, is to be removed as far as knowledge allows. Struggle, which is difficulty a being chooses, can leave, and grows by, is kept within reach on that being's own terms. (Chapter 7)

Structural suffering that a people has built its life around ends at the pace and by the means its people choose, and each member stays free to leave it sooner. (Chapter 134)

A reason that would fail to excuse a god for the suffering in a world also fails as a reason to design that suffering in, unless it rests on limits of cost or knowledge that bind a finite designer and would not bind a god. (Chapter 7)

Forced annihilation is among the gravest harms a conscious being can suffer. Every being able to choose remains free to decline continuation for themselves, which keeps opposition to

death from hardening into a duty to exist. (Introduction, Chapter 8)

The freedom to decline continuation includes the freedom to risk it, under the safeguards that govern a chosen ending. (Chapters 101 and 114)

Where a being at or above the line of protection cannot choose to decline continuation, it is never kept in a life not worth living. Repair comes first, and an ending comes only under criteria fixed before a world begins. (Chapters 4, 10, and 101)

Every protection that turns on a being's ability to choose needs a stated account of when the young become able, and that account lives in the keeper's policy and the inhabitants' institutions, since no law can test for it. (Chapter 107)

No law rewards desert. A world can make justice easier for its inhabitants to do and leaves the doing to them. (Chapter 120)

The laws of a world neither forgive nor condemn. A world can make forgiveness possible, and nothing in its laws or its keeper's policy makes forgiveness owed. (Chapter 121)

A world that ends death gives its inhabitants conditions under which roles, holdings, and ideas can change hands, and leaves the institutions of renewal to them. (Chapter 128)

When New Beings May Be Brought into Existence

Creating a being is permitted under demanding conditions. A created being receives its whole life without having been asked, and so no one may bring into existence a being at or above the line of protection unless it is overwhelmingly likely to reach the critical level, a life good by the standards of Part II for a being of its kind. The lives a world is expected to leave below that level are limited in aggregate and repaired first. (Chapters 10 and 12)

Selection may not work by suffering or forced endings among creatures at or above the line of protection, so no protected creature may evolve as ours did, by overproduction and early death. (Chapters 10 and 88)

Whoever creates answers for every being that arises in their world, intended or not, for as long as that world runs, a responsibility that survives any handing over of custody. (Chapters 10 and 140)

Good lives add value, so a world full of them is worth making, though no one is owed the creation of anyone. The value of further good lives is a reason that justifies creating without requiring it, and it yields to the duty to rescue the dying. (Chapters 12 and 13)

Among outcomes with the same number of lives, the designer must choose the better. Across different numbers, a life adds value when it reaches the critical level, lives worth living below that level neither add value nor subtract it, so adding them leaves outcomes incomparable, and lives not worth living subtract value. (Chapter 12)

A good life is not better for its subject than never existing, since never existing is no condition of anyone. (Chapter 4)

The dying already exist and stand to lose everything that remains to them. Rescue is owed to them whatever they have done, and where the same means could carry the dying or create new beings, the dying come first. (Chapters 4 and 13)

While no one knows whether a mind carried across survives as the same subject, methods that carry it without a break count as rescue, and methods that destroy the original are a bet only the person may take, knowingly, in person or by advance directive. No proxy may take the bet for anyone, unless a scan after a natural death, which ends no life early, proves an exception, a

question that remains open. Restoration after a body is destroyed falls under the same rule. (Chapters 13 and 100)

A risk no designer may impose on natives can be reasonable for someone who accepts it freely with death as the other choice, and precaution still asks that the passage be made at the finest grain the budget allows. (Chapters 2 and 13)

Whoever makes a copy of a mind is held to the designer's standard for the being it becomes, and no one copies a mind without its knowing acceptance. While the death emergency lasts, a wish to copy or enlarge oneself is weighed against the claims of the dying as a native's wish to have children is. (Chapters 13 and 28)

Reconstructing someone who has died from what they left creates a successor. It is open only to those who left a record of acceptance, every successor knows how it came to be, no successor is made to fill another's need, and reconstruction yields to rescue. (Chapters 13 and 21)

Whoever builds a mind writes its capacities and dispositions and leaves their use to the one who has them. The mind receives no memories of events that never occurred, no deference installed for its maker's sake, and no desires fitted to its maker's convenience, and if it is capable of reflection, it can revise its own values. (Chapters 9 and 98)

A designer may write the capacities for attraction, attachment, and care and may not write whom they fall on. (Chapter 118)

A mind built to dream is built so that its figures stay below the line of protection and its dreamer can always leave a nightmare. (Chapter 105)

No mind is built with a switch that lets another end it at will. A pause held under the duty to resume and an ending under criteria

fixed in advance for a being who cannot choose are no such switch. (*Chapter 124*)

How natures are distributed in any world whose minds are written, among its founders and through the inheritance that passes them on, is a decision of justice and answers to the principles of Part II. (*Chapter 99*)

How a World Is Kept

Dependence on a keeper is the most serious political fact about a designed world. A machine on which a world depends gives whoever holds it a power that dominates even when it is never used, and its inhabitants live at that holder's mercy whether or not the mercy is ever withdrawn. The designer therefore acts on conditions and leaves lives to those who live them. (*Chapters 1, 2, and 9*)

The architecture of the machine is part of the answer to domination, and a keeper who wants to be trustworthy can begin by building a machine that does not depend on trusting them. Custody is distributed, so that no single will can end a world or hold it paused. (*Chapters 2, 9, and 24*)

A pause with no assured resumption counts as an ending, so the power to stop a world is a power to pause it, held under rules fixed before the world begins and bound by a duty to resume. No world may be begun without that duty secured and without provision for the end of its machine, and no keeper may end a world against inhabitants who want to continue. (*Chapters 2, 24, and 25*)

When lost storage leaves a machine unable to carry every world it holds, a rule fixed in advance allocates what remains, and no subject whose share still exists is ended to free it for another. (*Chapters 4 and 25*)

A physics that lets inhabitants infer that their world will last must let them discover the bounded machine beneath it, and a physics that hides the machine must imply no such promise. (Chapters 2 and 9)

A world that outruns its keepers' attention outruns their capacity to notice its errors and its suffering, and a world run far slower than its keepers asks their custody to outlast every institution yet known. (Chapters 2 and 24)

A keeper who uses checkpoints restores what was lost into the present and never returns the present to the past. (Chapter 2)

Keepers observe aggregates by default and attend to a particular life only under review, which opens where suffering appears or where a task of their policy cannot run on organization alone. Limits on observation are built into the machine wherever they leave the keeper able to see suffering. (Chapter 130)

Privacy among inhabitants is a law of the physics, written as a law of physical properties that holds for every system alike, and privacy from the keeper is a property of the machine wherever one can be built. (Chapter 74)

A designer hears what inhabitants address to them through channels the inhabitants open, and answers with candor, declining worship without managing belief, so that attention by invitation is attention without surveillance. (Chapters 135 and 137)

A keeper who intervenes acts at the lowest effective level, restores agency before imposing any outcome, and reserves direct intervention for extreme cases. (Chapters 9 and 131)

Inaction in the face of atrocity is a choice the keeper answers for. The keeper restores the agency of the harmed first, acts inside a region only where a people manifestly cannot protect those

under attack, and never revises the minds of those who did the harm. (*Chapter 133*)

A designer who lives among the made without disclosing it lives under that world's laws like anyone else, with every power of design left outside. (*Chapter 132*)

Every inhabitant able to choose has somewhere else to go, and a world's physics makes leaving cheap. (*Chapter 139*)

Worlds are contained by structure wherever structure suffices, since containment by structure dominates no one, and vigilance watches only aggregates and interfaces. (*Chapter 142*)

Moral judgments built into a physics can be mistaken, so a populated world's physics changes on moral grounds only by a procedure fixed before the world begins, and judgments likely to need revision live wherever possible in the keeper's policy, which can be corrected, outside the laws themselves. (*Chapter 138*)

Bibliography

- Abdo, A. A., M. Ackermann, M. Ajello, K. Asano, W. B. Atwood, M. Axelsson, L. Baldini, et al. 2009. “A Limit on the Variation of the Speed of Light Arising from Quantum Gravity Effects.” *Nature* 462 (7271): 331–34.
- Aharonov, Dorit, and Michael Ben-Or. 2008. “Fault-Tolerant Quantum Computation with Constant Error Rate.” *SIAM Journal on Computing* 38 (4): 1207–82. First presented in 1997. With Kitaev and with Knill, Laflamme, and Zurek, the threshold theorems of the late 1990s that Chapter 2 credits with making quantum machines reliable in principle.
- Animal Welfare (Sentience) Act 2022, c. 22. <https://www.legislation.gov.uk/ukpga/2022/22>. The Act with which Chapter 5 opens.
- Anthropic. 2025. “Exploring Model Welfare.” April 24. <https://www.anthropic.com/research/exploring-model-welfare>. The research program on the welfare of its own models that the Introduction mentions.
- Aquinas, Thomas. 1947. *Summa Theologica*. Translated by the Fathers of the English Dominican Province. New York: Benziger Brothers. On God’s preservation of creatures, see the First Part, question 104.
- Aristotle. 1984a. *Nicomachean Ethics*. Translated by W. D. Ross, revised by J. O. Urmson. In *The Complete Works of Aristotle*,

- edited by Jonathan Barnes, vol. 2. Princeton: Princeton University Press. Book I sets out the hierarchy of crafts and the master art that gives this book its name, and Book VI treats the practical wisdom a designer would need.
- Aristotle. 1984b. *Politics*. Translated by B. Jowett. In *The Complete Works of Aristotle*, edited by Jonathan Barnes, vol. 2. Princeton: Princeton University Press.
- Austin, J. L. 1962. *Sense and Sensibilia*. Edited by G. J. Warnock. Oxford: Clarendon Press.
- Ayala, Christopher L., Tomoyuki Tanaka, Ro Saito, Mai Nozoe, Naoki Takeuchi, and Nobuyuki Yoshikawa. 2021. “MANA: A Monolithic Adiabatic iNtegration Architecture Microprocessor Using 1.4-zJ/op Unshunted Superconductor Josephson Junction Devices.” *IEEE Journal of Solid-State Circuits* 56 (4): 1152–65.
- Beane, Silas R., Zohreh Davoudi, and Martin J. Savage. 2014. “Constraints on the Universe as a Numerical Simulation.” *European Physical Journal A* 50: 148. First circulated in 2012.
- Bekenstein, Jacob D. 1981. “Universal Upper Bound on the Entropy-to-Energy Ratio for Bounded Systems.” *Physical Review D* 23 (2): 287–98.
- Bennett, Charles H. 1973. “Logical Reversibility of Computation.” *IBM Journal of Research and Development* 17 (6): 525–32.
- Birch, Jonathan. 2024. *The Edge of Sentience: Risk and Precaution in Humans, Other Animals, and AI*. Oxford: Oxford University Press. The source of proportionate precaution settled by public deliberation and of the sentience candidate, on which Chapter 5 models the line of protection.
- Birch, Jonathan, Charlotte Burn, Alexandra Schnell, Heather Browning, and Andrew Crump. 2021. *Review of the Evidence of Sentience in Cephalopod Molluscs and Decapod Crustaceans*.

- London: LSE Consulting. The review with which Chapter 5 opens, which graded the evidence against eight criteria.
- Block, Ned. 1978. "Troubles with Functionalism." In *Perception and Cognition: Issues in the Foundations of Psychology*, edited by C. Wade Savage, 261–325. Minnesota Studies in the Philosophy of Science 9. Minneapolis: University of Minnesota Press.
- Borges, Jorge Luis. (1946) 1998. "On Exactitude in Science." In *Collected Fictions*, translated by Andrew Hurley. New York: Viking.
- Borsanyi, Sz., S. Dürr, Z. Fodor, C. Hoelbling, S. D. Katz, S. Krieg, L. Lellouch, et al. 2015. "Ab Initio Calculation of the Neutron-Proton Mass Difference." *Science* 347 (6229): 1452–55. The lattice calculation of the masses of the proton and neutron, and the small difference between them, that the Introduction describes.
- Bostrom, Nick. 2003. "Are You Living in a Computer Simulation?" *Philosophical Quarterly* 53 (211): 243–55. The argument that made the possibility that our own world is computed a standard question in philosophy.
- Bostrom, Nick. 2006. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds and Machines* 16 (2): 185–200.
- Boswell, James. (1791) 1998. *Life of Johnson*. Edited by R. W. Chapman. Oxford: Oxford University Press.
- Butlin, Patrick, Robert Long, Eric Elmoznino, Yoshua Bengio, Jonathan Birch, Axel Constant, George Deane, et al. 2023. "Consciousness in Artificial Intelligence: Insights from the Science of Consciousness." arXiv:2308.08708. The assessment of artificial systems against indicators drawn from scientific theories of consciousness that the Introduction mentions.

- Chalmers, David J. 1995a. "Absent Qualia, Fading Qualia, Dancing Qualia." In *Conscious Experience*, edited by Thomas Metzinger, 309–28. Thorverton: Imprint Academic.
- Chalmers, David J. 1995b. "Facing Up to the Problem of Consciousness." *Journal of Consciousness Studies* 2 (3): 200–219.
- Chalmers, David J. 1996. "Does a Rock Implement Every Finite-State Automaton?" *Synthese* 108 (3): 309–33.
- Chalmers, David J. 2005. "The Matrix as Metaphysics." In *Philosophers Explore The Matrix*, edited by Christopher Grau, 132–76. New York: Oxford University Press. First published online in 2003.
- Chalmers, David J. 2022. *Reality+: Virtual Worlds and the Problems of Philosophy*. New York: W. W. Norton. The fullest case that virtual worlds are real and that simulated lives matter. It asks mostly from the standpoint of a visitor, and Chapter 1 asks from the designer's, about natives.
- Dacke, Marie, Emily Baird, Marcus Byrne, Clarke H. Scholtz, and Eric J. Warrant. 2013. "Dung Beetles Use the Milky Way for Orientation." *Current Biology* 23 (4): 298–300.
- De Brigard, Felipe. 2010. "If You Like It, Does It Matter If It's Real?" *Philosophical Psychology* 23 (1): 43–57.
- Department for Environment, Food and Rural Affairs. 2025. *Animal Welfare Strategy for England*. CP 1474. December.
- Descartes, René. (1641) 1996. *Meditations on First Philosophy*. Translated by John Cottingham. Rev. ed. Cambridge: Cambridge University Press.
- Deutsch, David. 1985. "Quantum Theory, the Church–Turing Principle and the Universal Quantum Computer." *Proceedings of the Royal Society of London A* 400 (1818): 97–117.

- Dick, Philip K. (1978) 1995. "How to Build a Universe That Doesn't Fall Apart Two Days Later." In *The Shifting Realities of Philip K. Dick*, edited by Lawrence Sutin, 259–80. New York: Vintage.
- Dilthey, Wilhelm. 1890. "Beiträge zur Lösung der Frage vom Ursprung unseres Glaubens an die Realität der Außenwelt und seinem Recht." *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften zu Berlin*.
- Dyson, Freeman J. 1960. "Search for Artificial Stellar Sources of Infrared Radiation." *Science* 131 (3414): 1667–68.
- Dyson, Freeman J. 1979. "Time Without End: Physics and Biology in an Open Universe." *Reviews of Modern Physics* 51 (3): 447–60.
- Eddington, Arthur. 1928. *The Nature of the Physical World*. Cambridge: Cambridge University Press. The Gifford Lectures of 1927, and the source of the two tables with which Chapter 1 opens.
- Egan, Greg. 1994. *Permutation City*. London: Millennium.
- Feynman, Richard P. 1982. "Simulating Physics with Computers." *International Journal of Theoretical Physics* 21 (6–7): 467–88. The keynote at Endicott House with which Chapter 2 opens.
- Frank, Michael P., and Karpur Shukla. 2021. "Quantum Foundations of Classical Reversible Computing." *Entropy* 23 (6): 701.
- Frankish, Keith. 2016. "Illusionism as a Theory of Consciousness." *Journal of Consciousness Studies* 23 (11–12): 11–39.
- Fredkin, Edward, and Tommaso Toffoli. 1982. "Conservative Logic." *International Journal of Theoretical Physics* 21 (3–4): 219–53.
- Ghirardi, GianCarlo, Alberto Rimini, and Tullio Weber. 1986. "Unified Dynamics for Microscopic and Macroscopic Systems." *Physical Review D* 34 (2): 470–91.

- Godfrey-Smith, Peter. 2016. "Mind, Matter, and Metabolism." *Journal of Philosophy* 113 (10): 481–506.
- Goff, Philip. 2019. *Galileo's Error: Foundations for a New Science of Consciousness*. New York: Pantheon.
- Hayek, Friedrich A. 1945. "The Use of Knowledge in Society." *American Economic Review* 35 (4): 519–30.
- Irvine, Elizabeth. 2020. "Developing Valid Behavioral Indicators of Animal Pain." *Philosophical Topics* 48 (1): 129–53.
- Jefferson, David R. 1985. "Virtual Time." *ACM Transactions on Programming Languages and Systems* 7 (3): 404–25.
- Jonas, Hans. (1979) 1984. *The Imperative of Responsibility: In Search of an Ethics for the Technological Age*. Translated by Hans Jonas with David Herr. Chicago: University of Chicago Press. First published as *Das Prinzip Verantwortung*.
- Kagan, Shelly. 2012. *The Geometry of Desert*. New York: Oxford University Press.
- Kagan, Shelly. 2019. *How to Count Animals, More or Less*. Oxford: Oxford University Press.
- Kammerer, François. 2019. "The Normative Challenge for Illusionist Views of Consciousness." *Ergo* 6 (32). <https://doi.org/10.3998/ergo.12405314.0006.032>.
- Kitaev, A. Yu. 1997. "Quantum Computations: Algorithms and Error Correction." *Russian Mathematical Surveys* 52 (6): 1191–1249.
- Knill, Emanuel, Raymond Laflamme, and Wojciech H. Zurek. 1998. "Resilient Quantum Computation." *Science* 279 (5349): 342–45.
- Korsgaard, Christine M. 2018. *Fellow Creatures: Our Obligations to the Other Animals*. Oxford: Oxford University Press.

- Krauss, Lawrence M., and Glenn D. Starkman. 2000. "Life, the Universe, and Nothing: Life and Death in an Ever-Expanding Universe." *Astrophysical Journal* 531 (1): 22–30.
- Kraut, Richard. 2011. *Against Absolute Goodness*. New York: Oxford University Press.
- Ladyman, James, and Don Ross, with David Spurrett and John Collier. 2007. *Every Thing Must Go: Metaphysics Naturalized*. Oxford: Oxford University Press.
- Landauer, Rolf. 1961. "Irreversibility and Heat Generation in the Computing Process." *IBM Journal of Research and Development* 5 (3): 183–91.
- Leibniz, Gottfried Wilhelm. 1969. "On the Method of Distinguishing Real from Imaginary Phenomena." In *Philosophical Papers and Letters*, 2nd ed., translated and edited by Leroy E. Loemker, 363–66. Dordrecht: Reidel.
- Lem, Stanisław. (1964) 2013. *Summa Technologiae*. Translated by Joanna Zylińska. Minneapolis: University of Minnesota Press.
- Lem, Stanisław. (1971) 1979. "Non Serviam." In *A Perfect Vacuum*, translated by Michael Kandel. New York: Harcourt Brace Jovanovich. A scientist listens as the beings in his computer argue about their maker, having wanted subjects to study and found himself with persons.
- Lloyd, Seth. 1996. "Universal Quantum Simulators." *Science* 273 (5278): 1073–78.
- Lloyd, Seth. 2000. "Ultimate Physical Limits to Computation." *Nature* 406 (6799): 1047–54.
- Locke, John. (1689) 1988. *Two Treatises of Government*. Edited by Peter Laslett. Cambridge: Cambridge University Press.
- Lorenz, Edward N. 1963. "Deterministic Nonperiodic Flow." *Journal of the Atmospheric Sciences* 20 (2): 130–41.

- Margolus, Norman, and Lev B. Levitin. 1998. "The Maximum Speed of Dynamical Evolution." *Physica D* 120 (1–2): 188–95.
- Maudlin, Tim. 1989. "Computation and Consciousness." *Journal of Philosophy* 86 (8): 407–32.
- Milton, John. (1667) 2005. *Paradise Lost*. Edited by Gordon Teskey. New York: W. W. Norton.
- Moore, G. E. 1903. *Principia Ethica*. Cambridge: Cambridge University Press. Its beautiful world that no one will ever see opens Chapter 4.
- Murdoch, Iris. 1970. *The Sovereignty of Good*. London: Routledge & Kegan Paul.
- Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *Philosophical Review* 83 (4): 435–50.
- Narveson, Jan. 1973. "Moral Problems of Population." *Monist* 57 (1): 62–86.
- Nielsen, Holger Bech, and Masao Ninomiya. 1981. "Absence of Neutrinos on a Lattice (I): Proof by Homotopy Theory." *Nuclear Physics B* 185 (1): 20–40.
- Nozick, Robert. 1974. *Anarchy, State, and Utopia*. New York: Basic Books. The experience machine, which Chapter 1 sets against a designed reality on each of Nozick's three counts.
- Oganov, Artem R., Chris J. Pickard, Qiang Zhu, and Richard J. Needs. 2019. "Structure Prediction Drives Materials Discovery." *Nature Reviews Materials* 4: 331–48. A review of how quantum-mechanical calculation predicts the structures of materials before anyone has made them, as the Introduction describes.
- Parker-Holder, Jack, and Shlomi Fruchter. 2025. "Genie 3: A New Frontier for World Models." Google DeepMind, August 5. <https://deepmind.google/blog/genie-3-a-new-frontier-for-world-models/>. A learned world model that generates explorable

environments and keeps them coherent for a few minutes, as the Introduction describes.

- Perlmutter, S., G. Aldering, G. Goldhaber, R. A. Knop, P. Nugent, P. G. Castro, S. Deustua, et al. 1999. "Measurements of Ω and Λ from 42 High-Redshift Supernovae." *Astrophysical Journal* 517 (2): 565–86.
- Pettit, Philip. 1997. *Republicanism: A Theory of Freedom and Government*. Oxford: Clarendon Press. The account of domination that this book adopts, the condition of living at the mercy of another's will, whether or not the other ever interferes.
- Plutarch. 1928. "On Superstition." In *Moralia*, vol. 2, translated by Frank Cole Babbitt. Loeb Classical Library 222. Cambridge, MA: Harvard University Press. The source of Heraclitus's remark on the waking and the sleeping.
- Popper, Karl. 1945. *The Open Society and Its Enemies*. 2 vols. London: Routledge.
- Pour-El, Marian Boykan, and Ian Richards. 1981. "The Wave Equation with Computable Initial Data Such That Its Unique Solution Is Not Computable." *Advances in Mathematics* 39 (3): 215–39.
- Putnam, Hilary. 1967. "Psychological Predicates." In *Art, Mind, and Religion*, edited by W. H. Capitan and D. D. Merrill, 37–48. Pittsburgh: University of Pittsburgh Press.
- Putnam, Hilary. 1975. "The Meaning of 'Meaning.'" *Minnesota Studies in the Philosophy of Science* 7: 131–93.
- Putnam, Hilary. 1988. *Representation and Reality*. Cambridge, MA: MIT Press.
- Riess, Adam G., Alexei V. Filippenko, Peter Challis, Alejandro Clocchiatti, Alan Diercks, Peter M. Garnavich, Ron L. Gilliland, et al. 1998. "Observational Evidence from Supernovae for an Accelerating Universe and a Cosmological

- Constant.” *Astronomical Journal* 116 (3): 1009–38. With Perlmutter and colleagues, the discovery of the accelerating expansion that Chapter 2 describes.
- Routley, Richard. 1973. “Is There a Need for a New, an Environmental, Ethic?” In *Proceedings of the XVth World Congress of Philosophy*, vol. 1, 205–10. Sofia: Sofia Press.
- Russell, Bertrand. 1927. *The Analysis of Matter*. London: Kegan Paul, Trench, Trubner.
- Russell, Stuart. 2019. *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking.
- Salmon, Wesley C. 1984. *Scientific Explanation and the Causal Structure of the World*. Princeton: Princeton University Press.
- Sandberg, Anders, and Nick Bostrom. 2008. *Whole Brain Emulation: A Roadmap*. Technical Report 2008-3. Oxford: Future of Humanity Institute.
- Scanlon, T. M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.
- Scott, James C. 1998. *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.
- Searle, John R. 1980. “Minds, Brains, and Programs.” *Behavioral and Brain Sciences* 3 (3): 417–57.
- Searle, John R. 1990. “Is the Brain a Digital Computer?” *Proceedings and Addresses of the American Philosophical Association* 64 (3): 21–37. The wall that could be read as running a word processor.
- Seth, Anil K. 2025. “Conscious Artificial Intelligence and Biological Naturalism.” *Behavioral and Brain Sciences*, published online. <https://doi.org/10.1017/S0140525X25000032>.

- Shakespeare, William. (1623) 2015. *Macbeth*. Edited by Sandra Clark and Pamela Mason. London: Bloomsbury Arden Shakespeare.
- Shelley, Mary. 1818. *Frankenstein; or, The Modern Prometheus*. London: Lackington, Hughes, Harding, Mavor, & Jones. The maker who flees what he made, with whom Chapter 3 opens.
- Sidgwick, Henry. (1874) 1907. *The Methods of Ethics*. 7th ed. London: Macmillan.
- Singer, Peter. 1975. *Animal Liberation*. New York: New York Review.
- Singer, Peter. 1979. *Practical Ethics*. Cambridge: Cambridge University Press.
- Stebbing, L. Susan. 1937. *Philosophy and the Physicists*. London: Methuen. The reply to Eddington that Chapter 1 adopts, that physics explains the solidity of a table and leaves it standing.
- Tegmark, Max. 2008. "The Mathematical Universe." *Foundations of Physics* 38 (2): 101–50.
- Tegmark, Max. 2014. *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality*. New York: Knopf.
- Thompson, Dennis F. 1980. "Moral Responsibility of Public Officials: The Problem of Many Hands." *American Political Science Review* 74 (4): 905–16.
- Tononi, Giulio. 2004. "An Information Integration Theory of Consciousness." *BMC Neuroscience* 5: 42.
- Tononi, Giulio, and Christof Koch. 2015. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society B* 370 (1668): 20140167. The statement of integrated information theory's denial of experience to digital computers, which Chapter 5 treats as a condition on the architecture of the machine.

- Troyer, Matthias, and Uwe-Jens Wiese. 2005. "Computational Complexity and Fundamental Limitations to Fermionic Quantum Monte Carlo Simulations." *Physical Review Letters* 94 (17): 170201.
- Umāsvāti. 1994. *That Which Is: Tattvārtha Sūtra*. Translated by Nathmal Tatia. Sacred Literature Series. London: International Sacred Literature Trust and HarperCollins. Composed between the second and fifth centuries. Its grading of living beings by their senses and its two vows of nonviolence are the precedent Chapter 5 draws on.
- Wellman, Christopher Heath. 2012. "The Rights Forfeiture Theory of Punishment." *Ethics* 122 (2): 371–93.
- Wheeler, John Archibald. 1989. "Information, Physics, Quantum: The Search for Links." In *Proceedings of the 3rd International Symposium on Foundations of Quantum Mechanics in the Light of New Technology*, 354–68. Tokyo: Physical Society of Japan. The source of "it from bit."
- Wilkinson, Hayden. 2022. "In Defense of Fanaticism." *Ethics* 132 (2): 445–77.
- Wilson, Kenneth G. 1971. "Renormalization Group and Critical Phenomena. I. Renormalization Group and the Kadanoff Scaling Picture." *Physical Review B* 4 (9): 3174–83.
- Zhuangzi. 1968. *The Complete Works of Chuang Tzu*. Translated by Burton Watson. New York: Columbia University Press. The butterfly dream is in chapter 2.
- Zuse, Konrad. 1969. *Rechnender Raum*. Braunschweig: Vieweg.