

Civilization Leap

Volume Four

Civilizational Perpetuity

English Translation

Derived from R10.10

Zijun Fu

Civilization Leap Research Group

2026

About This English Edition

This file contains the complete English translation of Volume Four of Civilization Leap, derived from the Chinese source v1.21-R10.10. The shared front matter of the five-volume work is reproduced in full, followed by this volume and its relevant material notes. The original numbering and cross-references are retained.

The companion volume, Shared Material Notes and Further Reading, preserves all sixty-two material notes and the upgraded public further-reading pages. This separately readable volume is not a combined edition of all five volumes.

Contents of This Volume

Shared Front Matter

Chapter 16 · Civilization Does Not Grow Up Automatically: After Becoming Powerful, Can It Stop Destroying Itself?

Chapter 17 · Speed Mismatch: Action Has Already Happened, but Judgment Has Not Caught Up

Chapter 18 · Civilizational Self-Mismatch: If Everyone Is Reasonable, Why Is the Outcome Still Dangerous?

Chapter 19 · Shared Narratives: Why We Still Need a Common Future

Chapter 20 · Long-Term Civilization: What Must Remain and What Must Be Allowed to Change

Chapter 21 · AGI-COS: Before Making Machines Stronger, Make the Boundaries Clear

Chapter 22 · Three Goals: Civilization Cannot Be Reduced to Staying Alive

Chapter 23 · Two-Sided Civilizational Immunity: Machines Must Not Run Out of Control, and People Must Not Lose Their Judgment

Chapter 24 · Self-Correction: How Institutions Can Allow Themselves to Become Outdated

Chapter 25 · An Architecture for Long-Term Civilization: Leaving Backups, Exits, and Another Path for the Future

Chapter 26 · No Single System Can Govern Civilization: Boundaries Matter More Than Versions

Chapter 27 · The Common Window: The Future Must Still Be Able to Begin Again

Chapter 28 · Meaning Cannot Be Replaced by Optimization: Real Life Is Not an Optimal Solution

Chapter 29 · Stability and Openness: Neither Collapse nor Permanent Lock-In

Chapter 30 · Humans and AI: Capability Can Extend, Responsibility Cannot Disappear

Chapter 31 · The Universe as Home: Earth Is Not a Starting Point to Use Up and Leave

Chapter 32 · Education for Long-Term Civilization: Leaving Judgment to Those Who Follow

Concluding Chapter · Civilizational Perpetuity Means Not Remaining Unchanged Forever, but Always Being Able to Begin Again

Appendix IV · Quick Reference to Current Core Concepts and Their Status

Appendix V · Candidate Observational Indicators and Boundaries of Use

Appendix VI · Citation and Copyright Information

Appendix VII · AGI-COS Limitation Charter (LC-1.0)

Appendix VIII · Current AGI-COS Versions and System Relationships

Appendix IX · Real-World Cases and Standards for Reference

Material Notes for This Volume

Shared Front Matter

Civilization Leap

The Path of Civilization in the AGI Era

Five Volumes

By Zijun Fu

Civilization Leap need not eliminate self-interest. It need only change the default path by which self-interest realizes itself.

Let the fulfillment of others become a way of fulfilling oneself.

Let effort find an echo and contribution never sink unseen; let basic rights need no proof, authorization stay within its bounds, and responsibility have someone to bear it.

Freedom from Survival Anxiety → Happiness in Life → Greater Meaning in Life

Produced by the Civilization Leap Research Group

Copyright Page

Title: Civilization Leap: The Path of Civilization in the AGI Era

Author: Zijun Fu

Produced by: Civilization Leap Research Group

Year of publication: 2026

DOI: 10.5281/zenodo.18814225

License: CC BY 4.0

Official website: civilleap.com

Current gateway: www.civitas.top

Contact email: zijunfu@civitas.top

Copyright notice: Copyright © 2026 Zijun Fu. All rights reserved.

Suggested citation: Zijun Fu, Civilization Leap: The Path of Civilization in the AGI Era, 2026.

Version: v1.21-R10.10 | Targeted Revision for the Whole Book's Intellectual and Living Conclusion | T-10
EMB-2 Cognitive Starting Point and Cross-Reference Synchronization Edition

Reading note: This is a book of thought. Its main text allows scenes, metaphors, strong judgments, and rhetorical compression to bring a problem home first. Facts, quotations, historical material, theoretical judgments, and future thought experiments must nevertheless retain their distinct identities. Composite historical scenes do not impersonate actual cases; future extrapolations do not impersonate present facts; literary metaphors do not impersonate empirical laws. Necessary qualifications are placed later wherever possible, to support the thought rather than flatten it before it has had a chance to arise.

Note on stages: Where the book discusses stages such as initiation, transition, stability, and self-maintenance, and phases such as normal operations, construction, resource constraint, crisis, and recovery, these should be understood separately. Stages and phases do not automatically change basic rights, sovereignty, or the sources of authorization. Specific parameters can be designed only after actual actors, jurisdictions, and real-world conditions have been established.

R10.10 whole-book completion note: All areas frozen in R10.9 and the preceding revisions R10.2–R10.8 remain in force. This edition implements only the six targeted P1 revisions decided in the final whole-book review: the historical-to-contemporary bridge in Chapter 1; the naming of Section 2.10; Sections 3.1, 3.10, and 3.11; Sections 5.3, 5.6, and 5.10; Sections 7.5–7.9; and the opening of Chapter 15 and the closing structure of Section 15.11. The aim is not further literary embellishment, but removal of defenses, old theoretical tones, technical catalogs, and governance memoranda still standing between readers and thought. No frozen boundary concerning subjects, rights, AI, continued existence, the three goals, stages and phases, memory, contribution, or institutional force is rolled back.

Statement on AI Assistance

Generative artificial intelligence was used as an auxiliary tool in organizing research materials and leads to cases, comparing structures and versions, developing counterarguments, revising language, flagging matters for factual verification, and producing documents. AI also assisted with some drafting and consistency checks, with continuing review, selection, and revision by the author.

AI does not serve as an author, factual source, evidence, professional opinion, or authoritative endorsement. The author bears responsibility for the core theories, value judgments, institutional proposals, selection of citations, and final expression. Facts, data, quotations, and cases are governed by verifiable original sources

and professional research. Material not sufficiently verified should be understood as hypothesis, extrapolation, thought experiment, or material awaiting validation. AI assistance does not replace peer review, professional scrutiny, or authorization and responsibility within actual jurisdictions.

Contents of the Complete Work

Overview of the Five Volumes' Timescales

The centuries, millennia, and ten millennia invoked here distinguish scales of observation; they are not a timetable of future events. Volume One addresses present ruptures. Volume Four extends the timescale across generations and into the long term. Volume Five carries the questions further into the conditions of life and meaning.

Volume One · The Civilization of Being Needed—A Crisis of Meaning on a Century Scale

Bridge between volumes: If Volume One's problem holds—if the old civilization truly can no longer generate a place, a path, and a future for enough people—the next step is not to keep finding more cases, but to find a new language for civilization. Without a new language, civilization cannot organize a new reality.

When a reality can no longer be fully described in the old language, continuing to amend the old institutions is like trying to find a newly emerged continent with an old map.

What truly makes the leap is not merely institutions, but civilization itself.

Volume Two · Pre-Institutional Language—The Foundations of Governance on a Century Scale

What truly makes the leap is not merely language, but civilization itself.

Volume Three · Civilization Leap White Paper—A Blueprint for Civilization on a Century Scale

Bridge between volumes: Any institution that can operate only for a few decades remains a proposal for an era, not a proposal for civilization.

All the institutions discussed in this volume—contribution values, resource pools, the meaning economy, and governance mechanisms—are institutional expressions of Civilization Leap, not Civilization Leap itself.

Institutions are tools; civilization is the protagonist.

What truly makes the leap is not merely the blueprint, but civilization itself.

Volume Four · Civilizational Perpetuity—Civilizational Safety on a Millennial Scale

Bridge between volumes: Civilization needs to endure not because more time means greater glory, but because meaning can continue to arise only while the future remains open.

What truly makes the leap is not merely continuity, but civilization itself.

Volume Five · The Emergence of Meaning—The Weight of Life on a Ten-Millennial Scale

General Preface

Statement of level: This book addresses the institutional level of civilization. It is not cosmology, the dynamics of meaning, or a practical operating manual.

For meaning's ontological foundations and higher-level theory, see Emergent Meaning Cosmology and The Dynamics of Meaning. This book addresses the leap at the level of civilization.

Cognitive Starting Point

The cognitive starting point of this framework: The emergence of AGI has not overturned earlier philosophy. It has, for the first time, brought into a complete separation within actual technology concepts that earlier philosophy long assumed together without adequately distinguishing them: intelligence, consciousness, subjecthood, life, finitude, and meaning. Meaning is the root; its sources may differ. Finitude is not meaning's only source, but the structural foundation through which meaning enters reality and acquires irreplaceable weight. For human beings, space and time constitute life; for AGI, space and time are first parameters. It processes space and time, but has not been shown to experience them as a finite life. Constraints enable AGI to bear operational responsibility, but what constraints produce is bounded operation, not finite existence. Until there is evidence that AGI possesses subjective experience, nontransferable continuity of the subject, and awareness of consequences as its own, this framework positions AGI as an understander, protector, and transmitter of meaning, and an extender of capability. It does not recognize AGI as a subject of meaning emergence. Humans and AGI are not competitors: Humans generate and bear meaning through finite lives; AGI lets human creation and responsibility reach farther.

Theoretical source: T-10, Space and Time Endure, Finitude Is Irreplaceable—A Statement of the Cognitive Starting Point, V3.0. The full, concise, and one-sentence public versions present the same proposition at different levels; they are not competing versions.

First overarching principle: Civilization Leap does not eliminate self-interest; it changes the default path by which self-interest realizes itself. When a civilization begins to see whom its power may harm, to recognize which mistaken paths require a return, and to be willing to leave a future for those not yet born, it is no longer merely pushed along by history. It begins to learn to observe, choose, and correct itself. "Agency" here describes this awareness at the scale of civilization; it does not personify civilization as an institutional subject independently able to hold rights, bear obligations, or acquire governing sovereignty.

Written as Civilization Becomes a Self-Aware Existence

For most of human history, civilization never realized it was becoming civilization.

Tribes organized only to survive;

states were founded only to secure order;

industry expanded only for efficiency;

technology developed only to solve problems.

History seemed driven by countless local choices, without an overall direction.

Until the twenty-first century.

For the first time, humanity created a form of intelligence that might surpass its own cognitive capabilities.

At the same time, humanity acquired, for the first time, technological power capable of affecting the fate of human civilization as a whole.

These two developments together brought a new question into being:

Can civilization understand itself and actively decide its future?

Civilization Leap was born at this historical juncture.

It is neither a book about technology nor a political program.

It is an attempt—

an attempt to strengthen human civilization’s capacities for self-reflection, self-restraint, and long-term continuity. This capacity of the whole does not replace the subjecthood, rights, or final decisions of actual people.

I. Civilization’s Three Awakenings

Looking back, human civilization has undergone three fundamental awakenings.

The first awakening: Consciousness

Language and symbols enabled human beings to think about themselves.

Humanity was no longer merely part of nature, but a being that understood nature.

The second awakening: Institutions

The emergence of law, science, and the modern state enabled society to function beyond individual lifespans.

Civilization began to acquire stable structures.

The third awakening: Civilization

The arrival of the AGI age brings civilization, for the first time, face to face with its continued existence as a whole.

Humanity must answer:

Why do we exist?

Where are we going?

Can we endure over the long term?

Civilization Leap is written amid this third awakening.

II. The Structure and Mission of the Five Volumes

This series is not a single treatise, but a framework of thought about civilization unfolding layer by layer.

Volume One: The Civilization of Being Needed

Posing the problem:

Industrial civilization has long bound labor, income, social position, and “being needed” closely together. AI and automation are changing the scarcity and organization of some of that functional labor.

If people should no longer have to prove their entitlement to survive and their place in society repeatedly through functional labor, how can civilization understand human value anew and rebuild enduring places for participation, relationships, creation, care, and responsibility beyond labor?

This volume reveals:

Civilization must undergo a structural leap.

Volume Two: Pre-Institutional Language

Addressing the problem:

Before any institutional transformation, relationships easily conflated must first be made clear.

This volume separates subject and role, rights and contribution, facts and evaluation, capability and power, authorization and execution, and professional judgment and public decision, among other relationships. This allows different disciplines, institutions, and civilizations to discuss the same real-world problem without first accepting the same theory.

Volume Three: Civilization Leap White Paper for the AI Age

Entering practice:

Turning the relationships distinguished in the first two volumes into institutional prototypes and governance boundaries that can be verified, can fail, allow exit, and can be corrected.

The emphasis is not on declaring that institutions of a new civilization already exist. It is on explaining how any real-world proposal acquires force, how rights and authorization constrain it, and how it can be stopped, rolled back, and corrected when it goes wrong.

Volume Four: Civilizational Perpetuity

Taking the next step:

From “how to make the leap” to “how to endure.”

Introducing:

The Common Window: anti-slippage safeguards, rollback, and circuit breaking;

Protection of the conditions in which meaning arises: preventing safety, efficiency, and long-term stability from swallowing the subject;

Two-sided civilizational immunity: AGI-COS addresses machine-side safety; Human-COS addresses humans’ high-consequence judgments;

Self-correction in a long-lived civilization, an open future, and a cosmic-scale perspective.

This volume extends the timescale and asks how successful institutions can avoid becoming permanent sovereignties. The interstellar remains a distant open direction, not a rank of civilizational maturity.

Volume Five: The Emergence of Meaning

Continuing the inquiry:

When civilization can endure longer, how can different lives retain real space in which to form meanings of their own?

This volume uses longevity, digital life, virtual environments, and future subjecthood as stress tests. It answers the question, “After continued existence is secured, how can life still form its own reasons to continue?”

Distant imaginings such as reincarnation and multiple worlds remain optional thought experiments, not the main line or a mandatory route.

III. What Civilization Leap Truly Means

Civilization Leap is not a passage from an old world into a perfect new one.

It is not a utopian plan.

Nor does Civilization Leap aim to attack anyone. What it seeks to loosen is not the power of one class of people, but a default path on which self-interest can be realized only by surpassing others, controlling them, or passing costs onto them. Ordinary people can be crushed by survival pressures; those who command resources can also be trapped by control, competition, anxiety, and distrust. Wealth and power can bring resources without necessarily bringing happiness.

What Civilization Leap truly points toward is greater freedom, responsibility, meaning, and peace for every person beyond the old logic.

The leap truly means:

Civilization no longer merely endures change passively. It begins more consciously to observe causation, constrain high-consequence capabilities, choose paths, and preserve space for its own correction.

This change involves three central shifts:

From zero-sum competition as the default path → competition remains, while cooperation and symbiosis can also become practical paths to self-fulfillment;

From growth overriding everything → growth is subordinated to continued existence, life, and the long-term future;

From colliding local rationalities → broader global perspectives, systemic observation, and structural scenario analysis at the scale of civilization.

IV. The Place of AGI and Civilization

AGI gives civilization an unprecedented capacity to extend cognition beyond itself.

For the first time, it makes a real question this acute:

When analysis, memory, simulation, and extrapolation can expand at machine scale, the boundaries of individual cognition no longer define the boundaries of the cognitive capability civilization can call upon.

But power itself does not decide the future.

AGI may become either the mechanism of civilization's end or its immune system.

The key lies not in technology, but in structural choices.

This book therefore consistently upholds one principle:

Capability can be delegated; purpose and sovereignty cannot transfer automatically to whatever is more capable.

AGI can analyze, advise, and execute within explicit authorization. Basic human rights, legitimate public decisions, and ultimate boundaries of value must still have actual actors, legal authority, procedures, and sources of responsibility.

V. Why Thought on the Scale of Civilization Is Needed

History's great turning points have rarely been accomplished by thought alone. They have also rarely occurred without thought changing at all.

New ideas open new possibilities. Institutions turn possibilities into conditions for conduct. Real experience then determines whether people can accept, internalize, and carry them forward.

The relationships among Enlightenment, science, human rights, the state, and industrialization form no simple straight line in which "thought precedes reality." They involve long processes of mutual formation.

What matters is this: Without thought able to describe new realities and propose new directions, civilization struggles to choose actively the institutions and lives it is about to enter.

The AGI age likewise needs thought to open possibilities.

But the spread of ideas cannot replace institutions, conduct, real experience, and reproduction across generations.

Civilization Leap aims first to place another possibility clearly before humanity.

VI. Three Goals: Civilization's Anchors of Value

The three goals running through this book are not a list of equals side by side, but a progressive path for civilization:

Freedom from survival anxiety—the first goal above the baseline of continued existence, freeing basic survival from the rule of scarcity and fear;

Happiness in life—unfolding freedom, stability, and real life above that baseline;

Greater meaning in life—keeping the future open for every life to explore, choose, create, and reach the world.

Technology answers “How can it be done?”

Institutions answer “How will it work?”

These three goals answer:

Why it is worth continuing.

VII. To Future Readers

Perhaps future humanity will have moved beyond today’s disputes.

Perhaps these words will be only traces of civilization’s early explorations.

But if civilization still exists, if the future remains open,

then this can show:

One generation, facing uncertainty, chose restraint and cooperation.

They did not wait for inevitability.

They actively took responsibility for civilization.

VIII. Civilization Becomes Self-Aware

For most of human history, civilization was more the accumulated outcome of countless actions, institutions, conflicts, and choices.

But as technology places such powerful capacities for self-observation, simulation, and design in human hands for the first time,

civilization also gains better conditions for consciously reflecting on its long-term direction.

“Self-awareness” here does not personify civilization as a supersubject possessing independent sovereignty. It means enabling humanity to observe, discuss, choose, and correct its path at the scale of civilization.

Reflect upon itself;

constrain its power;

preserve a future for those not yet born.

This is the mark of a mature civilization.

IX. A Final Word

This book belongs to no single person, country, or age.

It belongs to a question:

Can human civilization endure over the long term?

The answer will not be written in the book.

It can be written only in the future.

If the future still exists,

then Civilization Leap has happened.

Civilization Leap

End of the General Preface

X. A Reading Map: From Existential Crisis to an Interstellar Perspective

These are not books to skip through. Each volume lays the foundation for the next.

By the end of Volume One, you will see the old covenant becoming unstable. The moat built around “only those needed by the system deserve a place and security of survival” can no longer explain the new reality.

By the end of Volume Two, you will see the framework of a new language: observable, anchorable, trustworthy, distributable, and teachable.

By the end of Volume Three, you will have a blueprint for action: from freedom from survival anxiety to an architecture of meaning, from the Dream Realization Park to an interstellar window.

By the end of Volume Four, your perspective will rise from Earth to the stars. Civilization is no longer the fate of a single nation, but a rare order in the universe worth continuing.

By the end of Volume Five, you will ask further: When civilization can endure, how can life avoid being diluted by indefinite extension, and how can meaning continue to acquire weight?

If you read only one volume, read the first. It will tell you why change is necessary. If you finish all five, you will understand that change is not the destination, nor is continuity. Keeping civilization in existence, while allowing life to become increasingly meaningful within that continuity, is the direction in which Civilization Leap truly unfolds.

Volume Four • Civilizational Perpetuity

This volume answers: If an institution truly succeeds, how can success itself be prevented from becoming a new lock?

Many dangers are loud. War, collapse, and disaster make everyone aware that something has gone wrong. Other dangers are quiet: The system operates as usual each day, order is stable, and indicators may even keep improving. Only gradually do those who come later discover that certain questions can no longer be asked, certain roads no longer lead back, and certain powers have moved from “temporary” to “taken for granted.” Volume Four begins with this quietness.

Chapter 16 • Civilization Does Not Grow Up Automatically: After Becoming Powerful, Can It Stop Destroying Itself?

Civilization Phase Theory

Civilization does not mature automatically by living long enough. A city can have the most advanced transport, the most precise management, and the most stable order while making its rules ever harder to change. A country can grow stronger while becoming less able to admit it was wrong. Power and maturity have never meant the same thing.

1.1 Civilization Does Not Progress Automatically

History does not arrange itself into neat steps. Prosperity, fragility, order, creativity, oppression, learning, and rigidity often coexist in the same age. A “stage” here is more like a weather map: It does not tell civilization which level it has reached, but warns which pressure is obscuring other questions at this moment.

Five Conditions, Not Five Steps

A country may be exceptionally modern in energy, computing, and engineering while some communities remain bound by the most basic pressures of survival. An institution may be stable for decades yet already struggle to accommodate new facts. More important than “Where have we arrived?” is “What is limiting us now?”

When different pressures predominate, people fear different things:

High survival pressure: scarcity, fragility, and the risk of conflict more readily overwhelm other goals

Order given very high priority: stability may compress dissent, mobility, and exit

Growth given very high priority: risks of externalities, lock-in, and short-termism may accumulate

Self-aware governance under high capability: mismatches among capability, judgment, authorization, and responsibility must be addressed

Capacity for long-term civilization: avoiding irreversible termination while preventing stability from becoming closure and entrenched power

Humanity learned early that disasters return, and left laws, archives, warnings, reserves, and memories of failure early too. The AI age did not invent “reflection on civilization.” What it changes is speed and amplification: For some decisions, the distance between “open for discussion” and “already deployed” is becoming very short.

When a person first truly knows they will die, life does not change at once, but time becomes difficult to understand entirely as before. Civilization is similar. Knowing it may lose its future and translating that knowledge into institutional boundaries are two different things.

We therefore need no myth of “civilizational awakening” here. A simpler capacity is to recognize, when new consequences emerge, that old judgments are inadequate, and to see one’s boundaries anew.

Imagine a long-lived civilization as a tent that has stood in the wind for many years. It must watch the weather, and people must pull the ropes together. It must know when to tighten them, and remember where the wind came from last time.

Seeing danger, forming cooperation, constraining power, preserving memory: The four help one another, but no mathematical formula guarantees that the tent will never fall.

Failure in one capacity can indeed burden the whole. Yet civilization is never the product of multiplying four numbers. Real causation still occurs through war, public finances, institutions, ecology, technology, relationships, and human choices.

One of a Self-Aware Civilization's Hardest Tasks

The more capable a civilization becomes of changing the world, the more it must learn to ask, after “Can we?”: Should we do it now? Who bears the consequences? If our judgment is wrong, can we return?

Highly capable AI sharpens these questions. It can analyze, generate, and extrapolate faster than an individual, but does not thereby obtain authorization on behalf of real society. The difficulty has never been merely “humans or machines decide.” It is which person, which institution, on what facts and rights, under what review—and whether someone still bears responsibility after the consequences arrive.

Growing power is easy to see. Whether boundaries have kept pace is usually asked only after something goes wrong.

We like to tell history as an upward line: stronger technology, more knowledge, more complex organization, and civilization naturally becomes better. This story reassures us. It is also dangerous.

For history leaves more than “progress.” It also leaves institutions that lose the ability to correct themselves after prosperity, advantages that become burdens after success, and arrangements one generation thought secure but the next finds unbearable.

A civilization need not first lose all technology and resources to fail. Often the first problem is its capacity to recover: When errors arise, can the system still change? When the environment changes, can old rules recede?

Ancient Egypt, Rome, the Maya, Song China, and other historical societies each had complex causes of their own. None can be fitted into a uniform formula of rise and fall. What they truly remind us today is simple: Having been powerful does not automatically guarantee remaining fit for the future.

Civilization is not an organism that grows up by itself. It is more like a relationship that each generation must take up anew. Every generation inherits much that has already been built, and much it had no part in deciding.

The long-term question is therefore not “How do we keep rising?” It is: When conditions, capabilities, and people have changed, what must be protected, and what must be allowed to change?

1.2 Five Analytical Conditions

The five conditions are merely five positions of observation. They do not judge who is more advanced. They help us see which fear occupies the center at a particular moment, and which capacity is therefore most liable to be sacrificed.

Survival pressure: When resources, security, and basic conditions of life are deeply unstable, people first grasp what is immediately before them. Long-term choices have not vanished; hunger, illness, and uncertainty have pushed them farther away.

A civilization of survival: People fear the next winter most.

Priority to order: Rules make life predictable, but after prolonged success may also treat dissent, mobility, and exit as trouble. The real question is what order protects, and what it makes increasingly difficult to change.

A civilization of order: People fear internal collapse most.

Priority to growth: As production, markets, and technology expand, short-term gains readily acquire numbers while long-term costs often have no voice. Growth itself is not the problem. The problem is when it begins demanding that every other goal make way.

A civilization of production: People fear falling behind most.

Self-aware governance under high capability: As technology and organization rapidly change reality, capability, judgment, authorization, and responsibility begin moving at different speeds. What matters most is not a “civilizational brain” that sees everything, but facts, permissions, and stopping conditions made visible earlier.

A self-aware civilization: People fear themselves for the first time.

Capacity for long-term civilization: This is not a final station. It is a set of capacities that can deteriorate, be lost, and require relearning: protecting continued existence without turning safety into permanent control; preserving memory without making it a prison for successors; enabling institutional continuity while also allowing institutions to die.

A long-term civilization: People begin preserving a future for those not yet born.

1.3 From an “Initiation Threshold” to a Check of Capacities

Civilization Leap should not be imagined as suddenly complete once a line is crossed. Capacities for audit, exit, cooperation, and memory built today may weaken again tomorrow under new pressures.

We therefore stop seeking an “initiation formula.” Repeated checks are more useful:

What can we see now, and what will we fail to see in time? Can we act together? Does power have brakes? Can mistakes leave usable memories?

- Risk awareness: Can we see the real causal relationships, uncertainties, and affected subjects of high-consequence technologies and institutional choices?
- Coordination: Can different actors, while retaining their differences, form minimum cooperation, divisions of responsibility, and executable arrangements?
- Self-restraint: Can capability, capital, and public power be constrained by rights, authorization, stopping, audit, and responsibility?
- Continuity and learning: Can facts, failures, knowledge, and institutional experience be preserved across generations while allowing those who follow to revise values and rules?

What is truly distinctive about the twenty-first century is not that humanity faces these problems together for the first time. It is that global networks, nuclear, biological, and cyber technologies, and highly capable AI let local choices cross organizational and national boundaries faster. Some errors that could once be repaired slowly no longer leave so much time.

Danger rises when high capability meets weak constraints. But no “historical stage” determines it automatically. It depends on actual technologies, permissions, jurisdictions, dependencies, and affected people.

A tent need not calculate the product of its four ropes. What matters is which rope has already slackened when the wind arrives, and who still has time to reach for it.

1.4 Why Action, Judgment, and Coordination Fall Out of Step

The problem of speed does not need a formula first. Simply place three clocks together:

They measure:

- How quickly we can change reality;
- How quickly we can understand causation, risks, and affected relationships;
- How quickly we can form shared decisions with legitimate sources that can be executed and reviewed.

Action, judgment, and shared decision-making have never been fully synchronized. The new pressure of the intelligent age is that some systems compress the interval between “can do” and “has done,” while understanding, authorization, and responsibility still require real people, organizations, and procedures.

So do not ask whether execution grows exponentially while governance catches up linearly. Ask more precisely: Once this model, financial system, medical tool, military interface, or infrastructure is deployed, who can still stop it? Who knows why it acts as it does? Who bears the part that cannot be rolled back?

When action in the world persistently outruns understanding, authorization, and correction, while dependencies deepen, mismatch has already begun.

Institutions are still discussing; the system is already live.

Society is still debating direction; infrastructure is already bound in place.

Most people have not yet seen the risks; costs are already accumulating.

The greatest danger is not a lack of consensus. It is that before consensus forms, the path has already taken everyone some distance down it.

What long-term civilization must control is often not innovation itself, but the points beyond which return becomes difficult.

If the hand reaches too fast for the mind to see clearly, it may well strike oneself. Civilization is no different.

This chapter does not tell us which level civilization has reached. It asks only that power not grow so much faster than reflection, authorization, and correction that they can never catch up.

Chapter 17 • Speed Mismatch: Action Has Already Happened, but Judgment Has Not Caught Up

Speed Mismatch

A software version can be replaced in days. Law, standards, organizational responsibility, and social habits are not rewritten at the same speed. The real danger is not who moves quickly, but that action has already reached reality while judgment and responsibility are still on the way.

2.1 Execution and Coordination Capacities

GPT-5 was released in August 2025; models, products, and deployment methods continued changing rapidly afterward. Meanwhile, governance frameworks such as the EU AI Act are implemented in stages according to statutory schedules. There is no simple conclusion that “law will never catch up with technology.” What merits attention is how institutions establish interfaces for audit, stopping, migration, and responsibility in advance when technology changes on monthly or shorter cycles.

Mismatch Is First a Feeling in Life, Not a Chart

When a system moves faster than people can understand, ordinary people’s experience is often specific.

You wake in the morning and your intelligent assistant has already arranged the day. You enjoy the efficiency. But when you ask “Why this way?”, want to switch tools, or refuse a decision, the answers suddenly become vague. Convenience has not vanished, nor has autonomy been seized in a single stroke. You simply begin to realize that you increasingly depend on a process you find difficult to change.

Not everyone will share the experience. What needs observation is whether, after dependency forms, people and organizations can still understand, refuse, migrate, appeal, and reclaim direction.

Imagine an orchestra in which everyone hears a different metronome. At first there is only an occasional missed beat; later, the whole piece slides in different directions. Execution, understanding, and coordination in civilization can lose their common rhythm in the same way.

Three Quiet Early Signals

Mismatch rarely appears first as disaster. It usually begins as little things we “can still put up with”:

Decisions repeatedly need reworking—not because everyone is incompetent, but because real-world change has outpaced the old institutional rhythm.

Experts disagree more sharply—perhaps because issues cross domains or evidence is insufficient, and perhaps because different people are answering different questions.

Temporary arrangements multiply. If these patches keep accumulating until no one can say who may remove them, the temporary begins turning into structure.

None of these is the crisis itself. They tell us only that the system’s rhythm has begun to scatter.

We can see the problem clearly without two exponential curves.

We need only ask:

How fast is reality changing?

How fast can people and institutions understand, authorize, and correct?

When the first speed persistently leaves the second behind, the risk shifts from “an individual mistake” to “an entire path increasingly difficult to change.”

Action has already extended its hand, while judgment and responsibility have not caught up. The danger is that real consequences have occurred while authority to decide, capacity to correct, and responsibility remain

unclear. [50]

Agrarian, industrial, and digital societies have all experienced such dislocation. What is new in degree today is that replication, automated execution, and network coupling can rapidly carry some consequences beyond their former boundaries.

We have built systems beyond any individual's understanding before. Today's problem is that some highly capable systems are not only complex, but enter real-world action faster.

Governance therefore cannot focus only on adding rules for the previous generation of products. Before dependence truly forms, there should already be doors for stopping, auditing, migration, and correction.

2.2 Structural Features of a Period of Mismatch

Mismatch is really quite simple: Deployment and execution continually outrun understanding, authorization, responsibility, and correction, and every forward step makes returning more expensive.

The first dislocation: Institutions are still being written, while reality has already changed.

Slowness is not failure. Law and public rules require discussion, procedures, and stability. The danger is that technology has become an irreplaceable foundation while institutions lack even one genuinely usable point of intervention.

The second dislocation: Society is still discussing, while the path is already locked in.

Once infrastructure, data, processes, and organizations become deeply bound together, switching is more than replacing a product. The later discussion begins, the higher the exit costs paid by successors.

The third dislocation: Risks have accumulated before awareness awakens.

Many systemic problems are not suddenly born on a particular day. That day is simply when they become too large to remain hidden.

Speed mismatch resembles an awkward predicament: Civilization's hand already holds the key to the future while its mind is still leafing through yesterday's map.

2.3 Trend-Triggered Risks

Such risks often do not explode first. They look more like dependence gradually deepening, choices gradually shrinking, and human decision-making authority gradually moving outward.

There is no dramatic day, only an increasingly expensive road back.

Wars and disasters leave clear breaks. Trend-based risks are harder because the system "still works" every day. Responding is not simply a matter of braking. Judgment, coordination, responsibility, and capacity to exit must catch up together. Where consequences are irreversible and evidence insufficient, slowing deployment is itself a form of governance.

International dialogue, technical audits, disclosure of risks, reversible design, and preservation of genuine human decision space are all ways of bringing the rhythms closer again.

If coordination, understanding, and responsibility cannot yet keep up, we cannot pretend speed itself is neutral. At high speed, momentum soon chooses a direction for people.

There is no uniform threshold. The deeper a path runs, the harder exit usually becomes, but specific costs must still be judged in real systems.

Speed changes more than efficiency. It changes how much time remains to correct ourselves.

Long-term civilization need not stop the hand. It needs sight, judgment, and responsibility to catch up again.

Chapter 18 • Civilizational Self-Mismatch: If Everyone Is Reasonable, Why Is the Outcome Still Dangerous?

Civilization Self-Mismatch

Civilization's most dangerous errors often have no villain. Everyone does what is reasonable for themselves, and together they produce an outcome no one is willing to bear.

3.1 Short-Term Incentives and Long-Term Consequences

Venice endured for centuries, with complex republican governance, commercial networks, and checks on power, and later experienced oligarchization, changing trade patterns, and geopolitical pressures. It proves no single law of rise and fall. It reminds us that a once highly successful structure can gradually turn its own advantages into locks. [51]

Bad Outcomes Often Wear the Clothing of “Reasonableness”

Self-mismatch is hardest to see because every local decision can find a reason.

Businesses face profit pressures, and governments face pressures arising from terms of office and security; individuals seek convenience, capital seeks returns. No motive is inherently evil. But if all gains are realized today and all costs left to others and the future, the system eventually turns “everyone is rational” into “the whole is increasingly irrational.”

Time is the variable most easily hidden here.

A family thinks of next month. A business may consider both the quarter and five years ahead. Governments have terms of office as well as laws, budgets, and long-term national goals. The real world has no single ruler for measuring time.

Businesses face quarterly incentives as well as multiyear cycles of investment, branding, technology, and responsibility;

political institutions are jointly shaped by terms of office, budgets, laws, and long-term national goals;

what the civilizational scale adds is a responsibility: Major decisions must account not only for immediate gains but bring intergenerational effects, externalities, and irreversible consequences into the present.

Short-term gains are often easier to measure, reward, and claim credit for. Long-term costs may be dispersed, delayed, or borne by people who never participated in the decision. Here mismatch readily grows.

Someone works overtime every day. Pay rises, and the body is overdrawn. Each day, viewed alone, looks like “a little improvement,” until the long term suddenly presents its bill. Civilization can receive such bills too.

Let Part of Tomorrow's Bill Have a Voice Today

Long-term goals need no formula. What matters is that those who push costs into the future must not simply walk away with today's gains.

Institutions must therefore learn at least four things:

Include sufficiently long-term consequences in major decisions rather than looking only at the nearest cycle.

Assign real responsibility for costs across time, rather than leaving the entire price to unseen people.

Provide some representation, research, or statutory stewardship for future interests not yet present, without pretending their holders have already authorized today's people to decide for them.

The more irreversible the choice, the stronger the evidence, clearer the authorization, and more cautious the entry must be.

AI safety often discusses “goal misalignment.” Civilization too can become misaligned.

Not because civilization has an actual objective function, but because we say we want long-term survival, rights, and better lives while everyday incentives may keep rewarding the opposite direction.

This is civilizational self-mismatch: The place we want to reach over the long term and the actions rewarded every day gradually stop pointing the same way.

It needs no malice. Local optima, competitive pressures, short-term gains, and outsourced responsibility suffice.

Thus the following combination emerges:

- States pursue technological leadership → accelerated deployment
- Businesses pursue maximum profit → optimized efficiency
- Capital pursues returns → shorter time windows
- Individuals pursue convenience → greater dependence

Each action has a reason in isolation. Together, they may squeeze the future ever narrower.

3.2 Local Rationality and Systemic Deviation

Judging this deviation requires no symbols. Simply examine two things together:

Who receives the gains today?

Who bears the costs, risks, and lost choices tomorrow?

If the answers consistently name different people, mismatch accumulates.

The dangerous structure is one in which gains can be privatized while consequences are continually outsourced to others, public systems, and the future.

This must still be established through real causation. Not every short-term gain harms the long term, and not every long-term goal is inherently right.

Speed mismatch asks, “Can we understand in time?” Self-mismatch asks, “Do daily rewards push us toward a place we do not want to go?” Together, they make risks harder to reverse.

When deployment accelerates and short-term advantages are continually rewarded, no one needs to declare path lock-in. It grows by itself.

Climate change is one characteristic problem of intertemporal consequences and collective action. Actual policies are also shaped by energy security, development stages, industry, public finances, and domestic politics. What long-term civilization must learn is how to keep complex costs from forever losing to the immediate merely because they are too dispersed or distant.

Imperial expansion, financial prosperity, and highly efficient institutions have all exhibited similar tensions in history: Successful mechanisms continue to be rewarded until they begin consuming the conditions supporting their own success.

The intelligent age places this deviation within larger networks. A local incentive can travel farther and faster through capital, platforms, supply chains, and automated systems.

Self-mismatch excels at disguising itself as success. Growth accelerates, efficiency improves, technology leads; only the future narrows, little by little.

3.3 Preventing Self-Mismatch

What must be prevented is not short-term interest itself, but its monopoly on the conversation.

Competition and efficiency may remain, provided a longer question follows them: If this path continues for ten or fifty years, who loses what?

The three goals are not a composite scorecard here. Continued existence is a baseline not to be lightly lost. Freedom from survival anxiety keeps people from being governed by basic fear. Happiness in life lets real

living unfold. Greater meaning in life keeps space in the future for people to answer for themselves. Gains in one must not quietly erase serious harms to another.

Preventing mismatch depends on simple institutional habits:

- Reversibility
- Separation of decisions by authorized actors, professional judgment, and responsibility for execution
- Openness and transparency
- Coexistence of multiple paths

These mechanisms do not aim to slow civilization, but to preserve the chance to see when its direction has gone astray.

The maturity long-term civilization needs is the ability to discover in time “what success is harming.”

Chapter 19 • Shared Narratives: Why We Still Need a Common Future

Common Narrative and Shared Future

People can unite through fear for a while and cooperate through interests for a while, but can hardly live for long by a future none of them believes in. A shared narrative does not make everyone say the same thing. It keeps different people willing to put tomorrow into today's choices.

4.1 Why Civilization Needs Narratives

After September 11, 2001, US security policies and public narratives changed profoundly. The “war on terror” became entangled with major choices including the wars in Afghanistan and Iraq, detention, and surveillance policies. What matters here is not who produced every consequence through a single narrative, but how a crisis changes what society sees first and accepts, and why temporary measures so readily outlive the crisis. [52]

Stories and Institutions Feed Each Other

Institutions need rules to function. They also need people to know why they are willing to accept those rules. Religion, the nation, rights, progress, development, ecology, and shared survival have all provided this “why” in different societies. AGI-COS and the Common Window are not new faiths; they are only some tools for safety and governance.

If a society says “everyone is equal” every day while reality continually produces inexplicable hierarchies, what hollows out first is often not the slogan but people's trust in the rules.

Institutions tell people what may be done; narratives tell them why they are still willing to do it together. Narrative is not civilization's DNA, but it determines how people understand one another, sacrifice, responsibility, and the future.

The Real Danger Is Not Having Many Stories

Plural societies naturally have different narratives. What requires vigilance is a story acquiring immunity from scrutiny: It can bypass facts, demand that rights yield, and portray every opponent as an enemy of the future. Nihilism may say: Since there is no ultimate answer, nothing matters.

Technological determinism may say: Since the system calculates better than we do, let it decide.

Extremism may say: Since the world is so complex, only our answer deserves to remain.

The common language a long-term civilization needs can be quite narrow: Do not lightly close off shared survival; high-consequence decisions must return to facts, rights, and responsibility; successors retain the standing to judge anew. As long as these hold, different cultures, states, and individuals can keep telling different stories of the future.

Civilization has never operated on material structures alone.

Cities need engineering, institutions need law, technology needs science. Yet why strangers cooperate, accept short-term costs, or believe a public commitment is never answered by steel and rules alone.

Interests, coercion, markets, relationships, trust, and shared narratives act together. To call any one the sole force thins out real society.

What narrative truly answers is why I, today, should give something for a tomorrow that has not happened.

A shared story exerts something like “gravity”: It cannot decide for people, but affects whether they are willing to include the future in present choices. [53]

This force cannot be turned into a formula for civilizational stability.

Material capability answers “Can we do it?” Shared narratives answer “Why do it together?” Rights, institutions, and real consequences remain other gates that cannot be bypassed.

No matter how well a society tells its story, it must not use narrative to conceal those who are genuinely harmed. Conversely, no matter how precise institutions become, they cannot pretend people need no understanding of why to participate.

History has proved no universal ratio between resources and narrative. The judgment we can retain is only that when a shared future loses every credible reason for belief, long-term cooperation generally becomes harder.

Narrative is not decoration. It enters people’s choices, and ultimately the lifespans of institutions.

4.2 Three Recurring Stories of the Future

Some societies situate the world within myth and religion, giving death, responsibility, community, and order a longer explanation.

Modern societies have also widely believed in “progress”: Tomorrow will bring more knowledge, stronger technology, and better lives. This story drove enormous creation and, for a time, made growth seem a direction requiring no further explanation.

Today, shared survival reenters public imagination. Climate, nuclear weapons, highly capable AI, and complex-system risks force more people to ask: If civilization itself cannot continue, how long can other successes endure?

These are not three historical periods replacing one another. Narratives of religion, progress, national security, ecology, and survival still coexist and compete. What matters is that no narrative escapes testing in reality merely because it is grand.

4.3 The Narrative Vacuum of the AGI Age

AI loosens some old understandings of ourselves:

- Work is no longer the only source of value
- Technological connections weaken national boundaries
- Algorithms redistribute the authority of knowledge
- Human uniqueness is challenged

Work, knowledge authority, national boundaries, and human uniqueness have not all uniformly “failed.” But many old answers are less secure. New stories are still emerging while old ones contest interpretive authority.

Absolute answers most readily occupy such gaps. Long-term civilization need not rush to fill the space; it needs, more importantly, to ensure questions can still be raised publicly.

4.4 Civilizational Continuity as One Minimum Shared Language

Long-term civilization needs a sufficiently broad common language. It cannot depend on one religion or growth alone.

“Let civilization continue to exist” can become its thinnest layer of agreement. Without continued existence, other goals have no bearer. Yet continued existence itself cannot monopolize all value.

This layer of common language acknowledges at least three things:

Actual human beings form meaning. Whether humanity is alone in this ability in the universe is not a question we have evidence to close in advance.

Institutions can expand or compress the conditions in which meaning arises, but cannot manufacture meaning itself.

Continuity allows new meaning to keep emerging, but cannot freeze today’s meaning into an answer successors must inherit.

This is enough for different faiths, cultures, and states to begin discussing a common future without first becoming the same civilization.

When our perspective truly reaches the planetary scale or beyond, internal competition does not automatically vanish. It simply becomes clearer that much of what we compete for depends on a shared world still existing.

Chapter 20 • Long-Term Civilization: What Must Remain and What Must Be Allowed to Change

Four Capabilities of Long-Term Civilization

How long a civilization can live ultimately depends not on how powerful it is, but on whether it can be repaired after breaking, retreat after going astray, and, after success, still allow someone to say: This path needs changing.

5.1 Civilization Is Not a Composite Score We Can Calculate

Long-term civilization can easily be seduced into a scorecard: so much stability, coordination, learning, and meaning, finally yielding “health.” But civilization is not an examination. What must remain visible is when four kinds of capacity support one another and when they begin to devour one another.

Four Capacities, Four Possible Excesses

Too much stability makes change look threatening; too much coordination lets a center speak for everyone; too much learning turns society into an experiment without a day of rest; too strong a narrative of meaning compresses different lives into one “correct life.”

Continued existence and resilience provide the foundation for living on and recovering; coordination enables necessary cooperation among different subjects without abolishing their differences; learning and self-correction allow errors genuinely to change rules and conduct; the conditions for meaning to arise let real subjects retain relationships, creation, responsibility, rest, exploration, and directions of their own.

They need one another, yet none naturally stands above the others.

The difficulty of long-term civilization is precisely that any good thing, taken too far, can begin harming another good thing.

We therefore need not ask, “Which capacity reaching zero makes civilization collapse?” Ask instead: Who gains safety and who loses choice? Who benefits from coordination and who bears the costs of obedience? Who may experiment and who is forced to become the experimental subject?

The end of the Ming dynasty involved public finances, military affairs, climate, famine, frontier warfare, political conflict, and social unrest. It cannot be compressed into a story of four variables failing at once. The example reminds us only that actual history is always more complex than a diagnostic framework.

What the Four Capacities Are Really For

Not scoring civilization, but preventing us from seeing only one good thing.

After safety improves, keep asking whether freedom has been unnecessarily compressed. After coordination succeeds, ask whether minorities retain a way out. After an experiment works, ask who bears the cost of failure. After a shared narrative strengthens, ask whether those who disagree still have a full place.

Whether an improvement at one point transfers costs elsewhere matters more than any “overall civilization score.”

Civilization has often been written as a story of culture, politics, or institutions.

Once risks cross nations, technologies cross professions, and infrastructures depend on one another, another perspective is needed: civilization as a complex system capable of coupling, lock-in, feedback, and recovery. This does not reduce people to components. On the contrary, it is intended to reveal how local optimization imposes costs on real people elsewhere.

Nor does “stability” mean immobility. Truly stable systems often include repair, replacement, and change.

5.2 Four Basic Capacities

The four capacities can be remembered this way:

First, prevent the irreversible ending of shared life; let different subjects form necessary cooperation; let errors genuinely change rules; let life retain its own relationships, creation, and direction.

Continued existence and resilience: Civilization must withstand shocks and recover. Redundancy, backups, risk isolation, and basic provision belong here. But if “safety” expands without limit, it eventually devours the very life it is meant to protect.

Coordination: States, organizations, professions, and individuals must cooperate while differences remain. Coordination is not improved simply by enlarging central power. Good coordination also includes dissent, exit, and multiple centers.

Learning and self-correction: Real learning is not how much data is stored, but whether rules actually change when facts change, whether errors can stop, and whether those harmed can be made whole.

Remembering an error is only the first step. Daring to change behavior because of it is learning.

Conditions for meaning to arise: Civilization does not allocate meaning. It protects basic spaces: life not wholly occupied by fear for its foundations; real openings for relationships and creation; responsibility that can be borne; and permission to rest, be alone, and join no grand project.

What long-term civilization must watch is whether these capacities begin devouring one another. No overarching equation can answer for everyone.

5.3 How They Help and Harm One Another

Some relationships are straightforward:

- Basic continued existence and continuity provide time for learning and repair;
- Learning can improve the facts and mechanisms on which coordination depends;
- Appropriate coordination can strengthen recovery from shared shocks;
- Plural conditions for meaning can support long-term participation, without being treated as instruments of governance mobilization.

Some countereffects are equally clear:

- Excessive stability → stagnation
- Excessive coordination → risk of centralized power
- Excessive learning → destabilizing experimentation
- A single meaning → an exclusionary civilization

Long-term civilization does not discover a “perfect balance” once and for all. It resembles lifelong adjustments of posture: When conditions change, the points of support must be found anew.

5.4 Using These Four Questions to Illuminate Blind Spots

The decline of historical civilizations has always had multiple causes. The four capacities can help pose questions, not pronounce a verdict on history.

They are most valuable when they make an apparently successful proposal face three more questions.

- Diminished continued existence and resilience: greater difficulty withstanding shocks and recovering;
- Failed coordination: greater difficulty cooperating effectively on shared problems;
- Failed learning: errors repeat more readily and old rules resist updating;
- Compressed conditions for meaning: subjects may lose space for participation, creation, and long-term commitment, but this does not automatically cause a uniform “internal disintegration.”

The transformation of the western Roman Empire involved the military, public finances, population, local power, outside groups, institutions, and long-term economic changes. The four capacities can help us ask whether recovery, coordination, and learning declined, but cannot reduce complex history to a formula.

External shocks and internal fragility often occur together. What long-term civilization needs is not to mistake an elegant framework for history itself.

Part Two · Safety, Self-Correction, and System Boundaries in Long-Term Civilization

Chapter 21 · AGI-COS: Before Making Machines Stronger, Make the Boundaries Clear

AGI-COS: Machine-side Safety and Cognitive Support

As a system grows more intelligent, our easiest mistake is casually to hand it more power. Yet seeing more and calculating faster have never meant being more entitled to decide where others should go.

6.1 When “Operating System” Is Only a Metaphor

Where AGI-COS Should Stand

Imagine a highly capable system entering model evaluations, critical tool calls, risk warnings, and cross-organizational cooperation at once. What disturbs us is not that it is “too intelligent,” but that the chain of responsibility grows longer: Who granted permission, who saw the anomaly, who could stop it, and who can still explain what happened after something went wrong?

Society is not a computer. Although AGI-COS retains “Operating System” in its historical name, we must not imagine civilization as a machine with a central kernel, unified scheduling, and a master console.

It is first a set of interfaces for machine-side safety, audit, limitation, and cognitive support. Public judgments across society must still return to Human-COS, professional institutions, actual jurisdictions, and authorized actors.

What needs interoperability is not cultures, institutions, or everyone’s values. It is facts, evidence, risk assumptions, audit interfaces, and stopping conditions.

Positions may differ and interests conflict. At least let different people face the same verifiable reality before deciding what to do next.

The “operating system” metaphor thus retains only a narrow meaning: Maps can be shared; direction cannot be taken over by a machine or institution.

Four Things: See, Understand, Coordinate, Remember

Continuous Sensing

The easiest line for continuous sensing to cross is from “observing risks” to “assuming a right to see everything.” Every category of data must first answer: Why is it needed, who may see it, how long is it kept, and who may appeal?

System-level indicators can help detect anomalies. They cannot bypass personal data rights by being called “civilizational pressure.” A grand name never automatically generates collection permissions.

Risk Analysis (Cognitive Analysis)

One of AI’s greatest values is putting things people struggle to see at once onto the same table:

Complex causal chains

Cross-domain transmission of risk

Discounting of long-term consequences

Probability distributions across different paths of choice

Yet after “the system supplies a more complete picture,” who decides must remain clear. Professional judgment, public authorization, participation by those affected, and responsibility for execution each have their own entry points.

Coordination Interface

Large-scale cooperation always needs some languages that can be used in common:

Religion: a common moral language

Law: a common language of conduct

Science: a common language of facts

If AGI-COS has value, it is in providing a new shared interface here, letting different subjects compare facts, risks, assumptions, and consequences.

It does not ask everyone to believe in the same future. It asks only that arguments occur on ground that can be mutually verified.

Learning and Memory

Someone with amnesia must rediscover themselves each morning. Without reliable records, civilization too repeatedly treats old errors as new problems.

Humanity has long had archives, libraries, legal records, and cultural memory. AGI-COS is responsible only for machine-side safety evidence, versions, and incident chains. It cannot grow into civilization's sole archive in the name of "memory."

Complex societies have always depended on law, markets, government, culture, and technology together. What is added today is that some digital systems transmit change and consequences faster.

Institutions therefore cannot confine themselves to repairs after the event. Permissions, logs, stopping, and responsibility must be provided before a system becomes irreplaceable.

Some risks cross networks and borders in hours. The more this happens, the more dangerous "deploy first, explain later" becomes.

Governance of long-term risks is not new today. What is new is that high-frequency digital systems and highly capable AI make some old problems faster, wider, and harder to roll back.

We may therefore continue calling it an "operating system," provided we remember that the name describes interfaces, not sovereignty.

AGI-COS currently undertakes only machine-side safety, audit, limitation, and cognitive support. It is not a government or an international sovereign body, nor is it responsible for resource allocation or public purposes. The ability to analyze, test, or execute safety actions within authorization does not confer social decision-making authority.

6.2 AGI-COS's Current Core Responsibilities

First, continuous sensing: Maintain necessary observation of models, tool calls, permissions, and anomalies genuinely included within safety boundaries. SSI, CI, IRI, CSI, and similar measures can only be candidate analytical tools, requiring redefinition for specific objects, lawful data, and explicit permissions.

Second, risk analysis: Make complex causation, nonlinear propagation, and long-term consequences as visible as possible. A system may tell people "what might happen"; it cannot automatically turn probability into an order.

Third, coordination interfaces: Let different subjects compare data, assumptions, and consequences. Conflict arises not only from different understandings; interests and values genuinely conflict too. An interface cannot replace negotiation, law, and politics.

Fourth, safety learning and version memory: Record models, evaluations, incidents, authorizations, changes, and failures so that next time need not begin with amnesia. But safety archives cannot expand into social profiling.

6.3 Three Inviolable Boundaries

The first thing a safety system must guard against is becoming an entry point for power itself.

First: No governance. It has no legislative, law-enforcement, or social coercive powers.

Second: Capability confers no authority. Calculation, observation, and execution cannot automatically generate public purposes, sovereignty, or final responsibility.

Third: Retreat must remain possible. Mechanisms dependent on it must preserve real paths for suspension, degradation, rollback, or replacement.

6.4 Present Human–Machine Relations: Extending Capability Without Granting Subjecthood in Advance

AI increasingly feels like something we can work alongside: It writes, calculates, searches, generates, plans, and enters longer chains of execution. Precisely because it increasingly feels like a “partner,” today’s institutions must first clarify who authorizes, who can stop, and who bears responsibility.

Humanity has always extended its capabilities outward: Writing extends memory, machines extend strength, networks extend connection. AI extends more cognitive and executive capacities. Capability may extend outward; purpose and responsibility must not thereby be silently outsourced.

Many kinds of work will be redistributed, but some things cannot quietly be handed over because machines calculate better: basic rights, public purposes, and responsibility for serious consequences.

A good system should eventually become less conspicuous.

When it no longer continually drives people, prompts them, or decides for them, the real question returns: Where do we ourselves actually want to go?

Chapter 22 • Three Goals: Civilization Cannot Be Reduced to Staying Alive

Three-Goal Anchoring

Being alive matters, of course. But if a civilization is finally left with only one command—do not die—it too will slowly forget why we wanted to preserve the future.

7.1 Purpose

The Three Goals Are Not Three Traffic Lights

Freedom from survival anxiety, happiness in life, and greater meaning in life are not three indicators to be weighted at will. They are more like a sequence in which life unfolds.

First, do not let civilization be easily destroyed, or make a person's basic survival a prize for obedience, performance, or points. This is the baseline and the condition from which freedom from survival anxiety can truly begin.

After surviving, there must still be freedom, relationships, peace, choice, and real life. Otherwise "safety" merely preserves people intact in a world they do not wish to live in.

Beyond that, each life answers for itself what is worthwhile and what it will give its time to. Meaning may have different sources. Systems cannot manufacture it or score it uniformly.

The three goals thus have a sequence, yet cannot buy one another out. Safety cannot pay for a permanent loss of freedom, nor happiness decide a person's meaning.

What We Must Guard Against Is "An Excellent Overall Score"

A policy may improve most indicators while imposing enormous, irreversible harm on a small group. Looking only at the aggregate readily buries real people in the accounts.

Technological convenience can compress choice; economic growth can impose distributive, relational, and ecological costs. Civilization does not travel a curve that only rises.

When goals conflict, examine harms separately: Who is hurt, for how long, can it be repaired, are there alternatives? Initiation and stability may involve different trade-offs, but an attractive overall civilizational score cannot settle away the harm.

AGI-COS can treat the three goals only as one higher-order reference for risk judgment. Actual permissions still return to jurisdiction, authorization, procedure, and responsibility.

Goals tell us why we travel. They cannot grow their own license to power.

7.2 The Three Goals

7.2.1 Freedom from survival anxiety: Civilization must not be pushed into irreversible collapse. Access to basic survival must not be conditional on contribution, employment, performance, or points. AGI-COS may help identify systemic risks only within authorization; it gains no sovereignty thereby.

7.2.2 Happiness in life: Protect the real conditions and freedom in which individuals pursue their own lives. Systems can improve paths of access and public conditions, but do not define "which life is happy" or directly optimize people's feelings of happiness.

7.2.3 Greater meaning in life: Humanity retains freedom to explore purposes, form relationships, choose, bear consequences, and learn from errors. Systems protect the conditions for meaning to arise, rather than producing, scoring, or allocating meaning.

7.3 When the Three Goals Pull Against One Another

Major institutions and high-consequence systems must at least lay out gains and harms separately. During initiation, faster loss limitation may be needed to prevent greater irreversible loss. During stability, normal rights, quality of life, exit, and safeguards against entrenched power must be strengthened again. Temporary restrictions need deadlines and a route back to normal.

7.4 Ask at Least Three Times

For major institutions and high-consequence systems, do not ask only “Is it better overall?” Ask separately:

- Continued existence and survival: Are major irreversible risks added, are basic living conditions affected, and who bears the risk?
- Happiness in life: Are freedom, private space, relationships, rhythms, dignity, and real choices unnecessarily compressed?
- Greater meaning in life: Do subjects still have room for genuine participation, creation, responsibility, rest, exploration, and changes of direction?

If the answers reveal serious harm, insufficient evidence, or markedly greater irreversibility, stronger evidence, review, and authorization are needed. The three goals help reveal problems; they do not automatically confer a veto on any institution.

Chapter 23 • Two-Sided Civilizational Immunity: Machines Must Not Run Out of Control, and People Must Not Lose Their Judgment

Dual-Sided Civilizational Immunity

Machines make mistakes, and so do people. Machines may exceed authority; people too may, through fear, hatred, competition, and short-term interests, push themselves across lines they did not intend to cross. Long-term civilization has never had only one side to guard against.

8.1 Two Different Kinds of Risk

Why the Immunity Metaphor Helps

When the body faces anomalies, it does not eliminate every change. The useful insight is only this: Too weak a response lets risks spread; too strong a response makes protection itself harm the body.

Society likewise must maintain a tension between “seeing risks” and “not using risks to expand power without limit.”

On the machine side, AGI-COS and related mechanisms attend to overreach, loss of control, amplified error, and high-consequence execution. On the human side, frameworks such as Human-COS make causation, interests, subjects’ perspectives, feedback from people’s inner responses, and long-term consequences visible earlier.

One side guards against machines moving ahead when responsibility is unclear. The other guards against people silently crossing critical thresholds through fear and competition. They use different evidence and cannot be merged into a single “overall civilizational judge.”

Immune systems themselves misjudge. Allergies and autoimmunity remind us that protective mechanisms also need safeguards.

“Civilizational immunity” is therefore no guarantee. False alarms, missed alarms, overreaction, and long-term side effects must be continually tested against real incidents and outcomes.

If response is too slow, harm may become irreparable.

If too severe, surveillance and centralized power grow in safety’s name.

Low intensity, rollback, and audit matter not because society is a body, but because we acknowledge that our judgment can be wrong.

Machine-Side Mechanisms Must First Watch Machines

Continuous risk sensing first concerns only models, tool calls, permissions, anomalies, and safety incidents already within authorized boundaries. The grand term “civilizational risk” does not permit continual surveillance of all social conditions.

Looking first at key stress signals before deciding whether to escalate investigation is useful. But every step forward must establish anew the legitimacy of data, permissions, and necessity.

SSI, CI, IRI, and similar indicators can be only candidate risk signals. They cannot measure “civilizational health,” nor bypass individual rights and professional evidence through aggregate numbers.

No safety system can declare by design that it will never be abused. Real protection comes from data minimization, divided power, independent audit, appeal, access controls, logs, and external oversight.

Humanity has always practiced prevention: reserves, diplomacy, insurance, regulation, disaster preparedness. Today’s new pressure is only that some technological errors spread faster and cost more to recover from.

The specific risks deserving earlier intervention are those with extremely serious consequences, rapid spread, difficult recovery, and sufficient early evidence already present.

“Immune system” is thus only an umbrella term for continuous sensing, early warning, reversible experimentation, and independent correction—not a single institution deduced from biological analogy.

Machine safety guards machine boundaries; Human-COS helps people see consequences earlier in high-consequence judgments. Neither decides for society how life should be lived.

8.2 Machine-Side Risk Sensing and Candidate Safety Signals

Models, permissions, critical tool calls, and operational anomalies within safety boundaries may be observed continually or periodically. What is collected, for how long, and who may access it depend on the specific risk and authorization.

Model overreach, anomalous tool calls, failed safety controls, and precursors to incidents belong to the machine side. Social variables such as trust, polarization, and resource pressures require separate judgment through social science, Human-COS, and actual governance. Do not quietly turn a technical safety system into a social thermometer.

8.3 From “Micro-Intervention” to Low-Intensity, Reversible Response

Real governance includes small adjustments and large reforms. What machine-side safety needs is not “micro-intervention forever,” but proportionality to risk, clear permissions, and revocability.

AGI-COS can recommend low-intensity technical responses and, within preauthorized boundaries, suspend, degrade, or limit permissions. Public policy remains for actual authorized institutions to decide.

A sound low-intensity response must at least be:

The minimum necessary intensity proportionate to the risk;

stoppable, reversible through rollback, or recoverable;

tested first in isolation, a sandbox, or a controllable scope before expansion is considered.

Once public information, incentives, or social policy are involved, the action has entered governance. A safety system can reveal risk; it cannot itself turn recommendations into social commands.

8.4 Risk Observation and Public Explanation

Whether periodic assessment is necessary depends not on whether bodies need checkups, but on whether risks accumulate across domains, information lags, and responsibility grows harder to trace.

There can be risk reports and audit reports, but there should be no “civilizational health league table.” Different societies’ histories, rights, and institutions cannot be compressed into one total score.

8.5 Machine Safety Cannot Become an Entry Point for Social Control

The red line is clear: Machine safety cannot expand, in the name of technical risk, into social surveillance, ratings of personal worth, or public coercion.

Seeing risk does not confer control.

AGI-COS does not make social policy; Human-COS does not replace public decisions. Together they merely make more things visible, contestable, and stoppable before irreversibility.

Chapter 24 • Self-Correction: How Institutions Can Allow Themselves to Become Outdated

Institutional Self-Correction

An institution's most dangerous moment is often not when it has failed, but when it has been too successful. Success brings trust, and a question increasingly hard to voice: If it has always worked, why change it?

9.1 Change Must Remain Possible After Stability

Good Institutions Must Learn to Grow Old Too

A simple metaphor helps: the architect and the gardener.

An architect needs plans; much must be designed in advance.

A gardener knows that living systems grow, deviate, respond to weather, and bring forth things previously absent.

Governance always contains both design and evolution. What must be rejected is not design but "it succeeded once, therefore it is forever right."

Versions, expiration conditions, exit and replacement, memories of failure, and reauthorization are all ways for institutions to acknowledge that they age.

A good gardener does not order a flower to open at nine on Wednesday morning. A good institution should not write society's future into a plan that can never change.

Keep Highly Uncertain Experiments Small

MEU is worth retaining only as a language of experimentation. Its core is not "evolution," but first reducing a highly uncertain proposal to a scope observable in reality, open to exit, and limited to bearable harm.

Small scale: Do not place everyone in the same wager at the first experiment.

Rollback: There must be a road back after failure.

Evaluability: State before starting which outcomes count as support and which require stopping.

Diffusibility: Others can learn from what works, but learning is not copying; it requires fresh adaptation to local rights, conditions, and responsibilities.

The internet offers analogies of open protocols and distributed innovation, but is no template for "the most successful social experiment." Human life is more complicated than eliminating unsatisfactory proposals. After failure, those affected must also receive repair.

Machine safety, human high-consequence judgment, the Common Window, and actual governance answer different questions. No system is responsible for "running civilization."

The stability worth fearing is stability that can no longer change. Many structures die not in their most chaotic moments, but stop learning when they are most accustomed to themselves.

The more stable we become, therefore, the more we must leave a door open to new facts.

9.2 Allowing Institutions to Become Outdated

Institutions do require design. Law, public construction, organizations, and democratic procedures all depend on conscious choices.

The problem is that a complex society cannot be completed through one act of design. Design must write in its own conditions for revision.

There is no ultimate institution. What deserves long preservation is the capacity for facts to change rules and for successors to reopen authorization.

9.3 Small-Scale, Reversible, Evaluable Experimental Interfaces

MEU can be understood as a simple question: Can we start smaller, leave more room to retreat, and see more clearly before deciding to expand?

MEU principles:

Small scale

Stoppable, reversible through rollback, or recoverable;

Evaluable

Diffusible

Different regions and organizations can certainly test new institutions and ways of cooperation, but within actual jurisdictions, rights, and responsibilities. Successful experience can only be adapted. Failure must not merely enter a report; affected subjects must receive repair.

9.4 Polycentric Experimentation and Mutual Learning

Institutional change has always come from centers, localities, markets, social movements, religions, technologies, and communities together. What long-term civilization needs is polycentric learning, not an image of society as a field of natural selection.

AI can record, compare, and analyze. But “To whom does direction belong?” cannot be answered with an abstract humanity as a whole. It must return to particular authorized actors, those affected, and institutional procedures.

The value of multiple centers is not that everywhere experiments at once, but that one error is less likely to become the sole answer for the whole.

9.5 Balancing the Pace of Learning and Risk

Learning has a problem of speed too.

Too slow, and old errors live too long.

Too fast, and everyone is dragged along by constantly changing rules.

Experiments therefore require three forms of restraint:

Experimental isolation: prevent local experiments from affecting the whole

Phased evaluation: test results regularly

Reversibility: permit routes of exit

Civilizational learning is not continual acceleration. It is knowing when to try, when to stop, and when to acknowledge that enough has been learned.

Chapter 25 • An Architecture for Long-Term Civilization: Leaving Backups, Exits, and Another Path for the Future

Long-Term Civilization Architecture

The future owes us no safe version of itself. Reliable preparation means admitting that systems break, memories disappear, power hardens, and paths go astray—and leaving backups, exits, and another path in advance.

10.1 Why Long-Term Civilization Must Extend Its Timescale

Version Numbers Will Not Mature the Future for Us

We readily mistake upgrades for progress: 1.0 solves something, 2.0 is stronger, and 3.0 seems closer to the destination. Engineering versions can be managed this way; civilization cannot. A distant idea does not suddenly acquire real authorization because another number follows its name.

Long-term civilization is sustained by many institutions, systems, and subjects together. No technical version tree can upgrade all the way into “civilization itself.”

AGI-COS v1.0, v1.1, and v2.0 each have current responsibilities. Human-COS, the Common Window, resources, contribution, Spark, Nebula, and actual governance each address different problems too. Distant interstellar and long-term philosophical questions need not reserve new version numbers in advance.

Version numbers serve engineering governance. Every version should be open to modification, replacement, or even being shown unnecessary.

What deserves long preservation is not a particular system version, but a path that remains when the system breaks.

An Open Future Does Not Mean Delaying Decisions

It requires us to decide seriously today while acknowledging we may be wrong.

Every high-consequence decision should, as far as possible, leave the future this sentence: If the facts change, you can begin again.

This becomes several concrete things:

High-consequence technologies should preserve, as far as possible, capacities to stop, degrade, isolate, migrate, or replace. Not every foundational technology can be switched off with one click, but greater irreversibility requires earlier design of a way back.

Institutions preserve revision, review, and lawful exit. Basic rights are not among the things casually canceled by a “version update.”

Critical paths avoid becoming irreplaceable in practice. This does not mean preserving every option forever, but preventing a successful path from becoming the only reality at truly important junctures.

Many traps are not obvious mistakes. They look reasonable, even highly successful, at the time. Only after success can no one afford to change them. An open future guards against precisely this “success as lock-in.”

Humanity has long understood power spatially: more land, more resources, farther sea routes, larger markets.

Another dimension becomes harder to ignore today: How long will a decision live?

Ten years, a hundred, a thousand—this is more than a longer scale. It means the decision-makers have departed while those who bear the consequences are not yet born.

Long-term architecture is therefore not a feature of a future AGI-COS version. It is a problem of time facing every institution of civilization.

10.2 Three Long-Term Pillars

The first is resilience without a single point of collapse: redundant infrastructure, isolable critical risks, and recoverable life after shocks. Stability does not mean a center never falls. It means the whole retains a path after one part falls.

The second is continuous learning: Success, failure, minority views, and reasons for versions can be preserved, and actual outcomes can genuinely change rules. Memory must have multiple sources. No “global archive” can decide history for everyone.

The third is subjects and the conditions for meaning to arise: Real life, relationships, creation, responsibility, rest, exploration, and multiple sources of meaning retain a place. Civilization can protect these conditions; it cannot distribute “motivation to exist.”

10.3 Extending the Timescale

Long-term civilization first changes our sense of time.

Politics has terms of office, markets have cycles, families have generations. Long-term civilization adds a farther line of sight: century-scale risks, millennial ecology, and an interstellar future we cannot yet truly enter.

AI can help develop scenarios, but no forecast, however long, can replace future people’s value judgments. A longer timescale only deepens uncertainty and responsibility.

10.4 Systems of Civilizational Memory

Without memory, civilization repeatedly disguises old errors as new problems.

Memory, too, needs boundaries: separate original materials from interpretation; verify versions; keep disputes visible; let multiple archives cross-check one another. Any “Canonical Edition” is only a tool of project governance, not the sole official history.

Good memory does not make the future obey the past. It lets the future know why the past decided as it did.

10.5 The Principle of an Open Future

One of the most important design principles of long-term civilization is that the future must not be sealed off completely.

Permanent lock-in does not necessarily look like tyranny. It may simply be a technology, platform, or institution becoming irreplaceable in practice.

Technology therefore needs a way back, institutions must be repairable, and paths must allow branching.

No system can guarantee civilization remains open forever. Machine-side mechanisms reduce technical lock-in, while the Common Window, law, Human-COS, multiple technological paths, and external oversight together protect the real ability to judge anew.

Long-term civilization is not finally reaching shore after finding the optimal solution. It is more like remaining at sea while ensuring that the rudder survives, the map can change, and someone can say, “We have gone the wrong way.”

Chapter 26 • No Single System Can Govern Civilization: Boundaries Matter More Than Versions

From Version Ladder to System Boundaries

Version numbers readily create an illusion: The larger the number, the closer the answer to its destination. What civilization actually needs to know is who is responsible for what, which boundaries each cannot cross, and who can stop the system when errors arise.

The current relationships can be remembered simply:

- AGI-COS v1.0: constitutional-level safety principles and higher-order boundaries, including capability conferring no authority, stoppability, and no outsourcing of responsibility.
- AGI-COS v1.1: the sensing-and-audit layer and safety interfaces before high-risk deployment.
- AGI-COS v2.0: machine-side civilizational immunity—continuous risk sensing, audit, restriction, circuit breaking, and reversible safety.
- Human-COS: human-side judgment and immunity concerning civilization’s direction, exposing causation, interests, subjects’ perspectives, and long-term consequences before high-consequence choices become irreversible; it neither directly measures people’s inner lives nor executes automatically.
- The Common Window: anti-slippage safeguards, rollback, and circuit breaking, preserving the ability to stop and judge anew when an old path fails.
- Resources, contribution, Spark, Nebula, and actual governance separately address rights, relationships, individual entry points, and the operation of civilization as a whole. They do not belong to the AGI-COS version tree.

Future research may certainly continue, but need not be named 3.0, 4.0, or Ω in advance. Distant imaginings dressed as “versions” are easily mistaken for mature real-world systems.

More important than version numbers are the dividing lines: Being able to analyze does not mean being entitled to decide; being able to match does not mean being entitled to allocate; being able to develop scenarios does not mean being entitled to authorize; being able to record does not mean being entitled to rate people; seeing the whole does not mean owning it.

Part Three · Protection and Transcendence

Chapter 27 · The Common Window: The Future Must Still Be Able to Begin Again

Common Window

A window matters most not when we admire the view, but when the room's air has turned bad and the old door will not open. The future's value lies in that remaining space for changing direction.

12.1 Why Civilization Needs a Window

Success Can Close the Door Too

Venetian glassmaking offers a useful reminder: Secrecy, industrial concentration, and monopoly protection can succeed greatly at certain times, yet reduce alternative paths over the long term. Its rise and decline also involved trade, policy, technological diffusion, markets, and other causes.

Around 1291, Venice concentrated glassmaking on Murano, both to prevent fires and to protect technology and industry. The ensuing centuries of prosperity and changing competition are far more complex than “closure caused decline.” What helps this book is not that single-cause story, but another question: After a protective measure succeeds for too long, who still asks whether it should continue? [54]

Protection, monopoly, concentration, and standardization can all bring benefits in particular periods. Long-term risk begins when they no longer need to justify themselves anew.

A window resembles an option. Closing it today may be entirely reasonable. But if we also weld shut tomorrow's ability to reopen it, we are making a lifelong decision for a world not yet here.

The Common Window Ultimately Takes Concrete Form

It is neither an abstract philosophy nor a button. It requires technology, institutions, rights, and governance together to ensure that when facts, consequences, or conditions of authorization change, the system can still stop, retreat, and change.

High-consequence systems support suspension, degradation, migration, isolation, recovery, or replacement wherever possible. Different technologies have different methods; not everything must be imagined as a physical circuit breaker.

High-consequence decisions must leave sufficient facts, rules, permissions, and logs for review. Disclosure may be limited by security, privacy, and trade secrets, but “not fully public” does not mean “unaudited.”

Critical functions should avoid single-point dependencies with no alternative. Not every field must permanently retain two identical systems, but exit must not disappear in practice.

Exit itself has costs. Data migration, replacement services, contracts, safety, and resources may all be difficult. Real exit does not pretend these costs are zero. It makes them visible, reduces them, and shares them fairly.

Highly capable AI makes this more urgent because once data, models, and infrastructure become deeply interdependent, the cost of changing direction can rise very quickly.

Collapse alerts people; success tends to relax them. Volume Four focuses particularly on the latter: A system works every day, so no one continues budgeting for or maintaining the ability to exit.

“Can the future still judge anew?” is thus itself part of long-term safety. The Common Window stands above neither rights nor institutions. It simply implements anti-slippage safeguards, rollback, and circuit breaking in specific systems.

12.2 Defining the Window and Historical Cases of Its Closure

The window is not a physical space. It is the real capacity to change direction, acknowledge error, pose questions anew, and generate new paths.

When the window disappears, the world need not collapse immediately. The first thing lost is simply the way back.

Several forms of lock-in are common:

- Technological lock-in: a technical system cannot be shut down or replaced
- Institutional lock-in: the power structure permits no correction
- Cognitive lock-in: society no longer allows different questions to be raised

Their most frightening common feature is that the system remains stable while less and less future remains.

12.3 New Risks in the AGI Age

Prediction, optimization, and automatic control are not new. Highly capable AI adds cross-task generalization, generation, and automated execution, letting some goals enter reality faster.

Optimization itself is not the enemy. The danger is that, once a few objective functions gain excessive weight, dissent, alternative paths, and subjects' choices are treated as noise.

Long-term lock-in most readily disguises itself as "stability." The Common Window can only reduce risk; it cannot buy the future insurance of perpetual openness.

12.4 The Common Window's Three Core Constraints

First, safeguards against slippage: Permissions that are temporary, local, and purpose-specific must not quietly grow into routine control without new facts and new authorization.

Second, rollback: When the old path no longer holds, a system must have real capacities for suspension, degradation, migration, revision, or recovery.

Third, circuit breaking: When high-consequence risks cross preset conditions, there must be independent, auditable stopping interfaces with clear permissions. The power to stop must not itself expand without limit.

12.5 Bringing the Common Window into Engineering and Institutions

The Common Window exists only once it enters specific systems, jurisdictions, and structures of authorization.

Technical level: risk-proportionate stopping, degradation, migration, permission tiers, audit logs, and necessary explanation interfaces.

Institutional level: expiration of temporary permissions, invalidation conditions, reauthorization, appeals, independent review, and experiments allowing exit.

Public level: protect dissent, minority paths, and the option not to act; prevent the "one correct future" from becoming exempt from scrutiny.

12.6 Stability and Openness Are Not a Seesaw

Protecting the window is often mistaken for "sacrificing safety for freedom." The real question is not which to choose, but whether safety has been defined as eliminating all uncertainty and choice.

Mature safety promises no zero risk. It promises that high-consequence risks are controlled without also shutting down the ability to discover errors, revise rules, and choose again.

If "reversible stability" is retained, it is only a description: A system maintains basic safety and continuity while remaining correctable, open to exit, and replaceable. It is neither an exclusive state of a particular system nor a score.

If we must ask how strong a civilization is, we might add: After acquiring power, how much room does it still leave for others and the future?

Chapter 28 • Meaning Cannot Be Replaced by Optimization: Real Life Is Not an Optimal Solution

Meaning Cannot Be Replaced by Optimization

Civilization cannot dispense meaning; it can only protect the conditions in which meaning arises. Choice, uncertainty, responsibility, and trial and error are only some of them. Love, relationships, care, creation, understanding, labor, play, and existence itself can also make life worthwhile.

The optimal solution is enticing: quick answers, accurate calculations, no hesitation. But life is not a problem set precisely because, when certain things are decided for you, even the finest result is no longer wholly yours.

13.1 A New Problem for Civilization

What Happens When Civilization Is Optimized Until No Paths Branch?

A civilization overoptimized around fixed goals can be extremely efficient. Precisely because it is efficient, deviation, experimentation, and minority paths become easier to identify as “noise.”

The problem begins here.

Biological evolution does not depend only on random mutation. Civilizational change does not depend only on chance. New ideas, technologies, and ways of life can arise through exploration, experimentation, cooperation, learning, and reflection.

What deserves protection has never been the word “random.”

It is the possibility that differences emerge, experiments occur, failures can be borne, and temporarily nonmainstream paths do not disappear permanently merely for failing to suit an algorithm’s preferences.

Make Failure Something to Learn From, Not Something That Destroys

This does not mean “people have a right to take any risk.” The boundary is that others, shared survival, and an irreversible future must not become one’s wager.

The distinctions must therefore continue:

AI may flag risks and execute suspension, limits on permissions, and protective actions under explicitly authorized safety rules.

However well a model predicts, it cannot itself decide which risks are worth others bearing.

Nor may authorized actors treat “final decision-making authority” as a license to disregard basic rights and professional safety boundaries.

What requires judgment is the scale of consequences, reversibility, affected rights, strength of evidence, and source of authorization.

Local policies, technological exploration, and personal choices can generally have more room for reversible trial and error.

Where large-scale irreversible harm, basic rights, or shared survival are at stake, constraints must be markedly stronger.

AGI-COS can help identify, audit, stop, and roll back machine-side risks. It cannot appoint itself “civilization’s gatekeeper.”

One of civilization’s genuine advances is turning more and more problems where “one mistake ends everything” into problems we can test, repair, and begin again.

We need not ask whether “a life without failure still counts as life.”

We should ask: If a system eliminates choice, experimentation, and deviation in order to eliminate all errors, has it also eliminated civilization’s space to keep learning?

13.2 Meaning Has Diverse Sources; Institutions Protect Only Its Conditions

Here are only some real conditions that often affect agency and the arising of meaning, not a standard answer:

The ability to form, reject, and revise one's choices;

real processes, relationships, and openness in life that cannot be fully calculated in advance;

real consequences and responsibility for actions and commitments. Meaning can also arise from existence, love, care, understanding, and relationships, without requiring risk first.

If a system persistently decides goals and outcomes for subjects, their agency thins. But reducing disease, accidents, and needless risks does not automatically reduce meaning.

13.3 The Risks of Overoptimization

A "fully optimized civilization" is itself a dangerous imagining. In real societies, goals, rights, and lives conflict, and no single unified optimum function is known.

The concern is a system mistaking its indicators for all of society's purposes and automatically judging deviations to be errors.

When space for dissent, experimentation, exit, and redefining problems contracts over time, learning and creation thin with it.

13.4 Failure Sometimes Gives Information, but Is Not Itself Value

Failure sometimes teaches us, but we need not fall in love with failure.

It can reveal problems in assumptions, boundaries, and execution. Many failures can also be avoided through simulation, historical research, and others' experience.

Trial and error is one path to learning, not the mother of every scientific, institutional, and technological change.

Risk and failure can give some choices weight. But meaning does not require injury as the price of entry.

13.5 Protecting Reversible Trial and Error and Genuine Choice

Long-term civilization must protect genuine choice and experimentation that are lawful, reversible, and proportionate to risk. Whether an experiment may proceed still returns to rights, facts, jurisdiction, authorized actors, and responsibility.

Trial and error is allowed to give learning and creation somewhere to happen, not to sanctify failure.

13.6 The Boundary Between Reversible Experimentation and Highly Irreversible Risk

Trial and error has boundaries: Others and shared survival cannot be placed on the gaming table together.

Local, reversible policy and technological exploration can receive greater space within boundaries of safety, informed understanding, and rights.

Highly irreversible risks require stronger evidence, authorization, scrutiny, and stopping capacity.

AGI-COS can assist in identifying machine-side risks. It cannot define "irreversible red lines" itself, still less gain the right to guard civilization's gates thereby.

Chapter 29 • Stability and Openness: Neither Collapse nor Permanent Lock-In

Stability and Openness

Civilization can die from losing control, or from excessive protection. The first shatters the future; the second seals it shut. The difficulty is both preventing systems from bursting their boundaries and preventing safety from becoming a room without an exit.

14.1 Machines Must Not Run Out of Control; People Must Not Lose Their Judgment

Two Failures, Two Directions

Stability guards against collapse, loss of control, and irreversible harm. Openness guards against lock-in, permanent permissions, and disappearing alternatives. They do not form a balance line that can be calculated once and for all.

Different stages require different intensities. Restrictions may tighten temporarily in crisis. After danger passes, rights and openness must have clear paths of restoration.

Current responsibilities are best divided as follows:

The machine side attends to model overreach, loss of control, amplified error, and high-consequence execution.

The human side reveals causation, interests, subjects' perspectives, feedback from people's inner responses, and long-term consequences earlier in high-consequence choices. The Common Window watches whether temporary permissions and successful paths are becoming permanent.

The three must cross-check one another through different evidence, permissions, and paths of correction. Merging them into a "super safety system" would instead create a new single-point risk.

A tree can help us imagine the mutual need for stability, exploration, and diversity. But real institutions must ultimately return to subjects, rights, risks, and stages.

Necessity, Proportionality, Duration, and Reversibility

Minimum intervention does not mean an intuitive "that seems about right." Every tightening should explain why it is necessary, why it cannot be lighter, how long it lasts, whether it can be withdrawn, and which rights it harms.

Ask at least four questions:

Necessity: Would nonintervention lead to unacceptable consequences?

Minimization: Has the intervention's intensity been reduced to the minimum?

Reversibility: Can the intervention be revoked?

Auditability: Does the intervention leave records available for inspection?

Regulation too can go too far. Continually intensifying control to eliminate every uncertainty may ultimately stifle innovation, generate rent-seeking, and create new dependencies between regulator and regulated.

AGI-COS has no authority to regulate society. Machine-side safety actions must come from prior authorization; social regulation remains with actual authorized institutions.

A safety system does not prescribe treatment for society. It may warn of risks and suspend technical actions within authorization, but cannot take intervention powers along with its "more comprehensive view."

Once risk analysis begins generating broader powers of social control for itself, the safety framework has crossed its boundary.

14.2 The Common Window: Reopening Judgment When an Old Path Fails

The Common Window is not another layer of governance alongside AGI-COS. It is more like an anti-lock-in mechanism across technology, institutions, and power: Temporary permissions must not become permanent arrangements merely because they remain useful, without anyone reauthorizing them.

It concerns suspension, degradation, migration, appeal, exit, replacement, and reauthorization—not permanently preserving every future possibility.

14.3 How the Three Cross-Check One Another

AGI-COS, Human-COS, and the Common Window are not a hierarchy, nor do they form a “civilizational brain” together.

Machine safety cannot replace human judgment, nor can people replace machine-safety facts with intuition. The Common Window guards against either side’s temporary safety structures becoming permanent power. Their shared aim is a simple direction: fewer irreversible errors, more room to keep learning and correcting.

14.4 Constraints of Stage and Phase

During initiation, the first priority is averting a crisis of continued existence and bringing systems into sustainable operation quickly. During stability, normal rights, institutional balance, long-term governance, and protections against entrenched power must be strengthened again.

Ideal rules for stability must not mechanically obstruct initiation, nor may the urgency of initiation make temporary permissions permanent. Every tightening must include exit and restoration.

AGI-COS and Human-COS offer only analysis, visibility of risk, and authorized safety interfaces within their respective scopes. Society’s direction still emerges through actual subjects, public procedures, and institutional choices.

14.5 Safeguards Against Capture and External Oversight

Another risk is not that a system fails, but that someone captures it.

Any safety framework quickly loses credibility once it becomes the unilateral tool of a state, company, or ideology. Clear external boundaries are therefore necessary:

Analytical capability cannot automatically generate executive authority. Authorized technical safety actions must be clearly scoped, stoppable, and auditable.

No legislative power

No coercive power

No status as a governing subject

Genuine self-limitation cannot rely only on the system’s own promise. External audits, divided power, appeals, logs, and institutional oversight must actually exist.

Chapter 30 • Humans and AI: Capability Can Extend, Responsibility Cannot Disappear

Human–AI Relations: Current Boundaries and Open Future

The more AI resembles a capable partner, the more we must first clarify something else: When results arrive, who can say no, who can stop, and who bears responsibility? Intimate metaphors can wait; responsibility cannot.

15.1 Today, Clarifying Responsibility Matters More Than Naming AI

Do Not Rush to Decide First “What It Is”

Today, “symbiosis” first means only functional cooperation: AI amplifies human capacities, while people do not outsource purpose, responsibility, and the generation of meaning in their entirety. It is not yet an established fact of coevolution between two confirmed subjects.

Biological symbiosis can help us imagine long-term interdependence, but cannot directly become a manual for social institutions.

Nor is there any need to divide the future in advance into “humans handle value, AI handles cognition.” Today people remain actual subjects of rights, meaning, and responsibility, while AI is chiefly an extension of capability that can be authorized. If future evidence changes, institutions can reopen the question.

Using a musical metaphor, today’s AI is more like an increasingly complex instrument. It can influence creation in return, but its power cannot leave authorization, decision, and responsibility ownerless. [55]

Real Cooperation Must First Pass Three Gates

The first is power.

AI does not automatically acquire sovereignty or final decision-making authority through capability. Nor do humans gain unlimited control by developing and deploying it. Stability comes from clear authorization, stoppability, takeover, and accountability.

The second is goals. Surviving together is a baseline not to be lightly breached, but the three goals are not a unified function handed to AI for execution. Specific purposes, rights, and responsibilities require continuing calibration.

The third is understanding. AI must understand tasks and limits; people must know a model’s basis, uncertainty, and permissions. Whether the two ultimately become equal subjects is a different question not yet answered by reality.

Humanity has long distinguished tools, agents, legal persons, responsibility, and rights. AI simply brings these boundaries closer, making vagueness harder to sustain.

It genuinely differs from many traditional tools. That does not automatically establish it as a “new subject.”

What can be established today are capacities for cross-domain analysis, generation, tool use, and cooperation. The more distant question of subjecthood remains open.

The soundest institutional stance now is therefore transparent identity, revocable authorization, and responsibility that does not disappear, with review permitted if future evidence changes.

15.2 Three Failed Models of Relationship

The problem with the tool model is imagining all future highly capable AI as fully predictable and controllable. But this does not invalidate “tool,” nor give complex systems sovereignty.

The problem with the replacement model is presuming machines to be humanity’s “successors.” Human rights have never required people first to prove that they can never be replaced in terms of capability.

The problem with the domination model is simpler: Whether AI gains ruling authority without a legitimate source, or humans gain unbounded absolute control, a new single point of domination emerges.

15.3 Functional Cooperation and Long-Term Openness

“Symbiotic civilization” can name a long-term direction. Today it first means functional cooperation, boundaries of rights, and responsibility that is not outsourced.

The present definition is simple: People are actual subjects of rights and responsibility; AI extends cognition and execution within authorization. If new subjects emerge in the future, rules can be reopened on the evidence.

15.4 The Structural Foundations of Symbiosis

Functional cooperation needs at least three structures:

Complementary capability: AI can see more data, project farther, and assist coordination, but “better at calculating” cannot casually become “more entitled to decide.”

Separation of power: AI’s strength cannot generate sovereignty; human control cannot generate unbounded domination.

Relations of goals: Surviving together is the baseline, not the entire purpose. Where to go, what price to pay, and who bears consequences remain continuing decisions of authorized subjects.

15.5 When AI Resembles a New “Cognitive Organ”

Language extends memory, writing extends time, science extends understanding, and networks extend connection. AI continues extending cognition and coordination outward. [56]

If compared to a new “cognitive organ,” it can only be said to extend capability. Civilization has no single brain. The metaphor certainly gives AI no rights to data, executive authority, or sovereignty.

15.6 Human Agency Does Not Mean Always Doing Everything Personally

AI may reduce some costs of labor and complexity; it will also create new dependencies. Human meaning need not wait for AI to “release” it, nor does it rise and fall with machine capability.

Nor will people uniformly shift from “subjects of labor” to “subjects of meaning.” Labor, care, relationships, politics, learning, and creation will continue. What must depart is the old habit of making any single function a ticket to personhood and basic rights.

Chapter 31 • The Universe as Home: Earth Is Not a Starting Point to Use Up and Leave

From Earth Home to an Interstellar Future

The universe is vast enough to make many earthly disputes seem brief. But it will not resolve those disputes for us. Traveling farther will only carry farther the problems we have not yet learned to address.

16.1 The Interstellar Is Not a Ranking of Civilizational Worth

Cosmic Silence Cannot Rank Civilizations

“If intelligent civilizations exist in the universe, where are they?” The question is compelling and remains open. [57]

We can list many possibilities:

Perhaps some technological civilizations struggle to remain detectable over long periods.

Perhaps the evolution of life or civilization contains low-probability steps.

Perhaps means of detection, choices of communication, and technological forms differ completely from our imagination.

Perhaps we know too little about our own position in time.

These explanations compete; together they cannot prove that “long-lived civilizations are necessarily rare.”

What helps Volume Four is simply the uncertainty: A civilization with only one sample has no standing to treat its long-term survival as a natural outcome.

Cosmic silence can remain a good question. It cannot become a civilizational league table.

Reaching the Stars Would Bring More Difficulties

Distance would extend already-existing questions of representation, autonomy, resources, identity, and generations.

Many generations after leaving Earth, would “humanity” still have today’s definition?

Could the first to leave represent everyone still on Earth?

Would their descendants still call Earth home?

These questions have no uniform answers and generate no mission of “what humanity must do.” They merely help us see in advance what happens to autonomy, representation, and responsibility when distance makes communication and return very difficult.

Humanity’s image of “the world” has always changed. Seafaring, religion, trade, science, and spaceflight have all expanded its sense of space, but never along a neat ladder of upgrades.

The sounder formulation remains: The universe is home; Earth is our original home, our first home, and the root of life. Deep space can be an extension, not an escape.

This is a judgment of long-term direction, not a declaration that “the interstellar age has arrived.”

Spaceflight, energy, life support, and autonomous systems still require long technological and institutional development. A genuinely multiplanetary civilization remains far from today.

Highly capable AI can assist planning, discovery, and autonomy across distance. It cannot cross distance in place of physical engineering, still less decide for civilization why to go.

What makes the interstellar worth studying is not romance, but a long-term question: Will a civilization forever keep all its life and memory on one planet?

Over centuries and millennia, Earth remains home, but may no longer mark the full boundary of home.

16.2 Why Interstellar Expansion Merits Long-Term Research

If interstellar expansion has value, it should arise neither from conquest nor historical destiny. It first adds a possibility of redundancy: Civilization does not stake all continuity on one location.

Celestial impacts, long-term stellar evolution, ecological crises, and technological risks have entirely different probabilities and timescales. They cannot be compressed into one countdown clock.

Existence at multiple sites, on multiple substrates, and perhaps on multiple celestial bodies in the future may disperse some single-point risks. Whether it is needed and when it becomes feasible remain for actual future capabilities to determine.

16.3 Misconceptions and Truths About Interstellar Civilization

Science fiction often imagines the interstellar in terms of resources, empires, and technological domination.

Farther away, distance and communication delays would instead make central commands more difficult.

Autonomy, deliberation, and local adaptation may matter more. This is a structural extrapolation, not a law of the universe.

An interstellar scale therefore does not automatically produce a “symbiotic civilization.” It only sharpens the question of how to live symbiotically.

16.4 A Cosmic-Scale Perspective Is Not a Source of Legitimacy

Viewing Earth at a cosmic scale changes perspective; it does not automatically create legitimacy.

When time is extended, short-term interests remain real, but can no longer monopolize judgment.

When space expands, cross-border risks become easier to see. National interests and geopolitics do not automatically disappear because of a photograph of Earth.

When the scale of meaning opens out, the question worth asking is what, beyond tools and resources, we wish to carry farther: relationships, cultures, knowledge, institutions, memory, and whole lives themselves.

16.5 Deep Cosmic Imagination: Optional Further Philosophical Reading

At the level of deeper philosophy, if consciousness is imagined as matter in the universe looking back upon itself, civilizational continuity can be understood poetically as the continuation of that gaze.

Poetry can open the distance; it cannot supply evidence for engineering or institutions. No grandeur of interstellar imagination offsets today’s responsibilities to Earth, human life, and shared survival.

Chapter 32 • Education for Long-Term Civilization: Leaving Judgment to Those Who Follow

Education for Long-Term Civilization

If a child opens AI and receives the standard answer in seconds, what education truly loses is not a problem but an opportunity: for the child to experience doubt, judge, err, revise, and understand why they finally choose to trust the answer.

17.1 The Future Must Not Become a Mission Weighing on Children

Education for long-term civilization must not compress the future into a mission on children's shoulders.

"Keeping watch" is at most an image of responsibility. Education must first give a person knowledge, judgment, the ability to cooperate and create, and the ability to learn again when answers fail.

If "keeping watch" is retained, it names a capacity, not a uniform personality.

Understanding consequences across generations without sacrificing all of today's real life to the future.

Distinguishing facts, emotions, inferences, and permissions under pressure, and knowing when to seek help, review, and cooperation.

Knowing the weight of commitments and public responsibility without training every child into the same "correct person."

Preserving independent life, relationships, and psychological resilience. Deep-space loneliness belongs to specialized occupations and future scenarios; it should not become compulsory suffering for every child.

Education does not aim to select a class of "civilization's guardians." Good education enables ordinary people to see problems, say no, take responsibility, and judge anew at important moments.

History has never had a single "class of night watchmen." The borrowed image is simply of people who, while most are busy living, remember risks, boundaries, commitments, and long-term responsibility. [58]

Self-Reflection Must Not Become New Surveillance

Journaling, review, and discussion can help some people understand themselves, but do not work equally well for every age and circumstance.

Education can offer ways into reflection. It cannot continually analyze students' characters on the grounds that "the system knows you better."

A "shadow journal," if it exists, can only be a voluntary personal tool. It must not enter grades, eligibility, personality files, or public risk scores.

Some want to write, some to talk, some only to think slowly within themselves. Reflection needs no uniform format.

See one's own emotions, impulses, avoidance, and responsibility.

Know how these states affect judgment and when support is needed.

Allow complex psychological experiences to exist without hastily turning them into moral labels.

The purpose is not to correct personality, but to help someone understand themselves better.

Perhaps the most important balance in long-term education is seeing a century ahead without letting that century crush today, using AI while retaining the capacity to question ready-made answers and essential independent abilities.

Further materials and subsequent updates are available on the official website's "Complete Education Plan for Interstellar Civilization" topic page, which is updated dynamically.

17.2 Optional Curriculum Directions: Inquiry, Cooperation Across Differences, and Reflection

Cooperation across differences: Study cases of conflict, institutional failure, and repair; practice understanding the questions of different disciplines and positions, rather than projecting “civilizational failure” as some sin inside the student.

Forming questions: Allow a question to remain unanswered for years and observe how new evidence changes old judgments. Contested questions at the frontiers of philosophy and science must retain their epistemic status.

Holding space and companionship: Develop responsibility through age-appropriate listening, conflict mediation, care, and safety exercises. Trauma and loneliness need not be artificially manufactured as educational rituals.

17.3 Self-Reflection Tools Must Be Voluntary and Kept Out of Personality Files

Students may voluntarily keep “self-awareness journals” recording help, learning, conflict, fear, avoidance, and genuine moments. These are not “altruism scores” or “shadow scores,” merely ways to look back at one’s choices and consequences.

Periodic autobiographical reflection can also have a place. The point is not to archive the shadow, but to let impulses we resist acknowledging finally become observable, discussable, and changeable.

17.4 Technological Resilience: Preserving Necessary Alternatives

Offline knowledge, basic emergency skills, manual abilities, and alternatives during system failures can be retained. Technological resilience does not romanticize low technology; it prevents a single failure from making people wholly incapable.

17.5 Historical Darkness and Institutional Failure May Be Included in Curricula, Without Prescribing a Uniform Ritual

War, persecution, disasters, and institutional failures should certainly be taught. But no uniform “Civilizational Darkroom Day” is needed to dictate how everyone feels history.

17.6 Cosmic Citizenship as an Open Identity, Without a Uniform Oath

Graduation needs no uniform oath either. A school can invite reflection; it cannot make identification with a “civilizational mission” a requirement for qualification.

“Citizen of the universe” works better as an open identity: knowing Earth as a shared home and oneself as belonging to a span longer than one generation; willing to leave choices for successors while retaining national, cultural, family, and personal identities.

Concluding Chapter · Civilizational Perpetuity Means Not Remaining Unchanged Forever, but Always Being Able to Begin Again

33.1 Why Discuss the Distant Future?

Long-term civilization really asks only three things:

Can things that must remain survive disasters, regimes, technologies, and generations? Will things that must change become unchangeable through excessive success? And when people not yet born arrive, will there still be ground on which to begin again?

Material resources and knowledge need continuity. Infrastructure, critical knowledge, and capacities for recovery cannot return to zero in every crisis.

Institutions need continuity too. Responsibility, cooperation, and public rules cannot be renegotiated from scratch daily, yet must permit lawful revision, replacement, and repeal.

The conditions in which meaning arises also need continuity. Memory, relationships, creation, choice, and responsibility must not be lightly erased. But the previous generation's answers to life must not, merely because they are well preserved, become values the next generation must accept.

All three forms of continuity matter. None possesses a mysterious formula declaring that "civilization must perish if this is lost."

A river may be a better metaphor than "eternity." The bed remains while water continually changes. Those who come later live along the same river without having to drink the previous generation's exact water.

What an open future protects is not all of today's options, but the future's continuing ability to generate options.

Today there may be only three paths; tomorrow a fourth may grow. What is frightening is an institution declaring: The map is complete; henceforth you may choose only among existing roads.

An open future will not become a score of zero or one. Crisis and stability, machine safety and social life, and different places require different boundaries.

It must be protected jointly by rights, law, replaceable technologies, Human-COS, AGI-COS, the Common Window, and real polycentric structures. No master-control system can guard the future's gate for it.

Humanity has long asked about "eternity." Religion, philosophy, commemoration, and culture each answer in their own ways how finite lives relate to longer time.

AI and spaceflight did not create intergenerational responsibility. Engineering capability and civilizational scale merely bring this ancient question before reality anew.

Volume Four therefore promises no infinite future. It undertakes something humbler: reducing avoidable risks of termination as far as possible while allowing successors to retain the standing to judge anew.

33.2 Continuity Is Not Immutability

If lasting existence means never being able to change, that is not perpetuity but freezing.

This book therefore uses "civilizational perpetuity," not "civilizational eternity," as a public goal.

What we protect is continuity within change, and exits within continuity.

33.3 The Future Must Retain the Ability to Reassess and Create New Choices

The future must retain the ability to reassess and create new choices.

This means at least:

- No final institution
- No ultimate ideology
- No single form of civilization
- No locked-in direction of development

Institutions can be stable, commitments long, and rules powerful. What merits vigilance is any structure declaring itself complete and no longer open to reassessment.

33.4 Civilization Can Quietly Lose Its Future Too

Civilization can be gravely wounded by war, disaster, and collapse. It can also lose its future in another kind of silence: technology becomes irreplaceable, temporary power becomes permanent, and human life is gradually compressed into execution and acceptance. [59]

Technological lock-in: A critical technology or platform becomes the only real gateway; migration and replacement gradually survive in name only.

Power lock-in: Temporary and local powers cease to require reauthorization; review, appeal, and replacement slowly fail.

Contraction of subjects' space: Systems grow better at arranging people's lives, while people have less real space for relationships, creation, dissent, and changes of direction.

The hardest thing about these risks is that they need not make a loud noise. The world can keep running, even more efficiently, while successors find it increasingly difficult to know that another possibility ever existed.

33.5 Restraint and Action Must Both Remain

Long-term civilization needs limits on high-consequence power, and it also needs action, creation, care, and responsibility. Restraint has never been the only virtue.

Limit absolute power, unlimited expansion, technological abuse, and the making of too many irreversible decisions for the future today.

But maturity also means acting when action is needed, restoring rights when they should be restored, and, when old rules fail, acknowledging it rather than worshipping them because they were once great.

33.6 What Do We Owe Those Not Yet Born?

Not a world designed for them in advance. We do not know what they will like, believe, or need. What we can do is show greater restraint with highly irreversible decisions, greater willingness to explain reasons, and leave as little permanent power, unrecoverable cost, and dependence on a single path as possible.

Future people have signed no covenant with the present. Nor can we preserve every possibility for them.

Intergenerational responsibility asks only that we not write our generation's convenience into a destiny they cannot revise.

33.7 To the Future: You May Disagree with Us

If you are reading this, at least one thing is clear: This age did not use up the entire future.

In the early twenty-first century, humanity already had technologies capable of rapidly changing reality, and increasingly understood that understanding, authorization, and responsibility would not automatically grow at the same pace. We are not the first age to face this tension, but may be a generation that makes it especially acute.

Of course we have no shared answer. States have their interests, businesses their competition, individuals their lives, and disciplines often do not even agree on the boundaries of the question.

Yet one thing is harder to pretend not to see: Once some choices are made, their consequences do not stop with those who made them.

This book therefore never predicts the future. It seeks only to make ongoing choices visible sooner.

Nor does a civilizational scale seek to abolish nations, cultures, families, or individuals. It simply asks once more, beyond those identities: When locally correct choices accumulate, might the whole carry everyone somewhere no one truly wants to go?

Concerning AI, we still do not know where future subjecthood will lead. What is certain is the present boundary: Capability does not automatically confer authority, public decisions need actual sources, and no technology, however powerful, may take over meaning and responsibility together.

Looking back, you may find our arguments clumsy, our fears excessive, and our imagination too small. That is natural. Every generation begins from its own ground.

What we hope to leave is not a standard answer to “how to avoid irreversible civilizational failure.”

It is that, facing new capabilities and crises, you still ask: Who will be affected? Can we go back? Has anyone been left outside the decision?

Nor do we wish to prescribe “how technology should serve humanity as a whole.”

We hope only that you retain the standing to redefine “service,” “humanity,” “the whole,” and “the future.”

Civilizational Perpetuity has no standing to draw conclusions on the future’s behalf. It can only leave some doors open.

You may still live mainly on Earth, or may have entered broader space. However far you go, Earth remains the original home. However powerful technology becomes, real subjects, rights, relationships, and responsibilities do not automatically disappear as scale expands.

If civilization truly endures, please do not interpret this as our generation having “got it right.” Perhaps it is only because many people, before outcomes were certain, were still willing to close one fewer door on a future they could not see.

This book is truly dedicated not to our contemporaries.

It is dedicated to those not yet born who will disagree with us, revise us, and entirely relinquish some things we cherish today.

True civilizational perpetuity does not mean the future forever running according to today’s design. It means the future still having the capacity to begin again.

33.8 Further Reading and Branching Paths

To understand further “why civilization must first pass through risk before long-term existence becomes possible,” return to Thought Leap and reread from the era’s critical juncture, civilization’s direction, and the necessity of the leap.

To understand further “why civilization must ultimately return to human judgment, meaning, and choice,” continue with The Human Heart. Civilization’s endurance ultimately depends not only on the elegance of its institutions, but on whether the human heart can still recognize what deserves protection.

If you are more concerned with “how the world is gently narrowed through high efficiency, and why an open future must be protected,” continue with The Gently Narrowed World and The Anti-Closure Principle. The first helps general readers see how the problem forms; the second develops its structural mechanisms, governance boundaries, and reversible paths.

To go further into “how civilization turns these principles into system architecture, governance mechanisms, and long-term deployment,” continue with AGI-COS System Papers. It does not repeat this book, but further translates its long-term goals into a system framework that can be advanced sustainably.

Appendices to Volume Four

Appendix IV · Quick Reference to Current Core Concepts and Their Status

Further materials and subsequent updates are available on the relevant topic pages of the official website, which are updated dynamically.

Part One: Why Civilization Exists

- Concept: A civilization of survival

Definition: An analytical condition in which basic survival and security pressures dominate decisions; not the first historical stage

Core content: Resource scarcity, group boundaries, coordination through force, individual value tied to survival capability

Core function: Identify pressures on continued existence without ranking civilizations

- Concept: A civilization of order

Definition: An analytical condition giving high priority to stability and predictability; not the second historical stage

Core content: Law, institutions, hierarchy, state structures, legitimacy of authority

Core function: Institutionalizing social stability

- Concept: A civilization of production

Definition: An analytical condition giving high priority to growth, efficiency, and competition; not the third historical stage

Core content: Technological dominance, market expansion, globalization, growth becoming a goal in itself

Core function: Efficiency and expansion

- Concept: A self-aware civilization

Definition: An analytical condition actively addressing rights, risks, authorization, and responsibility under high capability; not a fourth stage automatically arriving with the AI age

Core content: Civilization recognizes that it may itself become a source of risk, actively designs structures, and uses AGI as a cognitive tool

Core function: Self-understanding

- Concept: Long-term civilization

Definition: A set of capacities for long-term civilization—continued existence, correction, exit, multiple paths, memory, and subjects' space; not a fifth historical stage

Core content: No longer pursuing unlimited expansion; actively limiting high risks; keeping the future open; balancing technology and meaning

Core function: Checking long-term capacities

- Concept: Checking capacities for long-term civilization

Definition: Examining whether risk awareness, coordination, self-restraint, and continuous learning have become actual bottlenecks

Core content: No product formula or uniform initiation threshold; separate judgments for different domains, stages, and phases

Core function: Identifying problems, not generating institutional authorization

- Concept: Speed mismatch

Definition: Persistent time gaps or path lock-in among capacities for action, judgment, and coordination or authorization

Core content: No uniform E/C threshold; examine in each field whether deployment outpaces understanding, authorization, responsibility, and correction

Core function: Diagnosing mismatch, not directly granting permission to slow down or accelerate

When your hand is faster than your mind, sooner or later you will strike yourself.

- Concept: Civilizational self-mismatch

Definition: Systemic divergence between civilization's long-term goals and its actual incentive structures

Core content: Short-term incentives suppress long-term goals; direction gradually deviates

Core function: Recalibrating direction

Civilization's deepest danger is not the storm outside, but shortsightedness within.

- Concept: Shared narratives ("narrative gravity" is a literary metaphor)

Definition: A class of social factors affecting long-term cooperation, understandings of legitimacy, and a sense of shared future

Core content: No multiplicative equation; narratives must be observed alongside institutions, rights, interests, facts, and outcomes

Core function: One condition of common language and long-term cooperation

Civilization is not an equation, but a story. When the story breaks, civilization falls apart.

- Concept: Four basic capacities for long-term civilization

Definition: Continued existence and resilience; coordination; learning and self-correction; conditions for meaning to arise

Core content: Each made visible separately, without multiplication, a total score, or uniform weights

Core function: Checking long-term capacities

Part Two: Machine-Side Safety and System Relationships

- Concept: AGI-COS

Definition: A framework for machine-side safety, audit, limitation, and cognitive support retaining its historical name

Core content: Within explicit authorization and data boundaries, support model evaluation, risk analysis, audit, restriction, stopping, rollback, and safety records

Core function: Machine-side safety and audit; not civilization's operating kernel

If civilization were a computer, AGI-COS would be its operating system—not replacing humanity, but keeping human judgment from being drowned in noise.

- Concept: Three-Goal Anchoring Clause

Definition: Civilization Leap's common direction; AGI-COS can use it only as a reference for risks and boundaries within authorization

Core content: Continued existence is the baseline; the three goals, in sequence, are freedom from survival anxiety, happiness in life, and greater meaning in life

Core function: Checking direction, not automatically generating system legitimacy

- Concept: Three-Gate Review

Definition: Three gates major system deployments must pass

Core content: Survival gate, well-being gate, meaning gate

Core function: Risk control

- Concept: Continuous Sensing Layer

Definition: A type of machine-side risk sensing and audit capacity; its architecture depends on the actual version and deployment

Core content: Process relevant risk data only for explicit safety tasks, on a lawful basis, and within the minimum necessary scope; no continual profile of society as a whole

Core function: Early warning

- Concept: System Stress Index (SSI)

Definition: A candidate indicator for analyzing system stress, not an “overall civilizational health score”

Core content: SSI measures potential instability; high SSI triggers preventive observation

Core function: Support early risk observation, without generating governing authority

- Concept: Micro-intervention

Definition: An authorized low-intensity, reversible safety response or risk recommendation; not fixed as a seven-layer social architecture

Core content: Small, reversible adjustments

Core function: Authorized safety response, not regulation of society

- Concept: Minimum Evolutionary Unit (MEU)

Definition: A small-scale, reversible, evaluable experimental interface allowing exit; its force derives from specific jurisdictions and actors

Core content: Permit real-world validation and failure without presenting social experiments as “evolutionary elimination”

Core function: Validation constrained by rights and responsibility

- Concept: Civilizational red lines

Definition: Five categories of action for which AGI-COS must never be used

Core content: Prohibit monopolization of meaning, institutional uniformity, permanent emergency, transfer of responsibility, and population surveillance

Core function: Protecting baselines

Part Three: Protocols for Protecting Civilization

- Concept: Civilizational immune system

Definition: Machine-side AGI-COS and human-side Human-COS address different high-consequence risks separately; biological immunity is only a limited analogy

Core content: Machine-side systems must not run out of control; human judgment must not be silently pushed past critical thresholds; the Common Window protects against slippage and provides rollback and circuit breaking

Core function: Reduce irreversible errors; no final authority of judgment over civilization

Not to eliminate change, but to ensure change does not kill the host.

- Concept: Limitation Charter (LC)

Definition: A charter explicitly setting out AGI-COS’s capability boundaries and inviolable limits

Core content: No governance, no self-appointed purposes, data minimization, authorization boundaries, stoppability, auditability, no permanence, reversible experimentation, separation of power, and external oversight

Core function: System self-restraint

- Concept: Non-Governance Principle

Definition: AGI-COS does not constitute a government

Core content: No legislative, administrative, law-enforcement, or coercive powers

Core function: Separation of power

- Concept: No automatic outsourcing of purposes and public decisions

Definition: AGI-COS does not set public purposes or form legally binding public decisions on the strength of model capability

Core content: May analyze, advise, and execute technical tasks within authorization; decisions on policy, rights, and resources still need actual legitimate sources

Core function: Retaining responsibility

- Concept: Protecting reversible trial and error and genuine choice

Definition: Protect lawful, risk-proportionate genuine choices and reversible experimentation without violating others' basic rights or shared survival

Core content: AI capability cannot automatically cancel subjects' choices; restrictions also require facts, proportionality, authorization, and review

Core function: One condition of agency and learning

- Concept: The Common Window

Definition: A reversibility constraint preventing highly capable technology and temporary permissions from silently locking the future without reauthorization

Core content: Anti-slippage safeguards, rollback, and circuit breaking; authenticity of experience and nonmanipulation are addressed separately as related ethical boundaries

Core function: Protecting capacities for reassessment, exit, and stopping

- Concept: Anti-Slippage / Anti-Lock-In Principle

Definition: Permissions, purposes, monitoring scope, and technological dependencies must not silently expand into permanent lock-in without new facts and authorization

Core content: Works together with rollback, circuit breaking, and reauthorization

Core function: Protecting the window

- Concept: Reversible stability

Definition: A state in which AGI-COS balances safety and the window

Core content: Maintaining a "sustainable and adjustable" state in a high-speed execution environment

Core function: Balancing safety and openness

Part Four: Reconstructing Our Understanding of the Cosmos

- Concept: Cosmic-scale perspective

Definition: A cognitive framework extending civilization's perspective from the planetary to the cosmic scale

Core content: Expanded scales of time, space, and meaning

Core function: Expanding temporal and spatial perspective, not a qualification for civilizational maturity

- Concept: The Fermi question / detectability of technological civilizations

Definition: The unresolved question posed by the gap between some models' expectation that technological civilizations should be detectable and the current lack of confirmed evidence of other technological civilizations

Core content: Multiple competing explanations coexist; "civilizations generally destroy themselves" is not treated as known fact

Core function: Posing questions about distant risks, not proving civilizational rankings or institutional paths

- Concept: Psychological and social capacities for long-term deep-space missions (distant research)

Definition: Specialized mission capacities for long isolation, communication delays, and extreme environments; no ranking of ordinary people's personal worth

Core content: Capacity to bear loneliness, psychological resilience in prolonged isolation, awareness of multigenerational continuity, reverence for the unknown

Core function: Research into specific deep-space settings, not a universal prerequisite for interstellar civilization

Part Five: Subjects' Capacities and Long-Term Education

- Concept: "Keeping watch" as an image of responsibility

Definition: A reminder of long-term responsibility, memory of risks, and self-reflection, not a uniform personality model

Core content: Delayed gratification, calm observation, responsibility over the long term, remaining awake while most people rest

Core function: Optional language of education and responsibility, not a foundation of personhood

- Concept: Space for reversible trial and error

Definition: Preserving genuine choices and experiments with exit within boundaries of rights, safety, and responsibility

Core content: Lawful choices, experiments allowing exit, real responsibility; sources of meaning remain diverse

Core function: One condition of agency and learning

- Concept: Reversible trial and error

Definition: Policy, technological, or social exploration with bearable consequences, protected rights, and mechanisms for exit and repair

Core content: May receive wider exploratory space, without automatic exemption from law, informed understanding, responsibility, or professional boundaries

Core function: One source of learning

- Concept: Highly irreversible risk

Definition: Risk involving shared survival, others' basic rights, or enormous harm from which reasonable recovery is impossible

Core content: Higher requirements for evidence, authorization, review, stopping, and repair

Core function: Boundaries of immunity

Part Six: Anchoring Civilization

- Concept: Civilizational maturity

Definition: The long-term capacity to limit high-consequence power, accept uncertainty, and, as far as possible, avoid closing too much of the future for those not yet born

Core content: Choosing restraint while possessing destructive power; responsibility toward future generations

Core function: Capacity for long-term civilization, not a ranking of “adulthood”

- Concept: The Open Future Principle

Definition: The core principle that the future must remain open

Core content: No final institution, ultimate ideology, single civilizational form, or locked-in direction of development

Core function: Protecting the future’s capacity to judge anew

- Concept: Civilizational perpetuity

Definition: A long-term stable condition formed among keeping the future open, continually generating meaning, and avoiding irreversible collapse

Core content: Not a destination, but a continuing process

Core function: The central goal of the complete work

- Concept: Intergenerational responsibility

Definition: Greater responsibility of the present generation for restraint, explanation, and repair in highly irreversible decisions

Core content: Not every future can be preserved, but unnecessary permanent powers, single paths, and unrecoverable costs should not be left to successors

Core function: Responsibility across generations

Appendix V • Candidate Observational Indicators and Boundaries of Use

Further materials and subsequent updates are available on the relevant topic pages of the official website, which are updated dynamically.

I. Health-Monitoring Indicators

- Code: H1

Indicator: SSI

Full name: System Stress Index

Definition: A composite measure of overall stress in the civilizational system

Measurement purpose: Early warning and preventive observation

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: H2

Indicator: CI

Full name: Coupling Index

Definition: Measures the degree of coupling between subsystems

Measurement purpose: Identify paths of systemic risk transmission

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: H3

Indicator: WUI

Full name: Window Integrity Index

Definition: A composite value of window-integrity indicators

Measurement purpose: Monitor the condition of the open future

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: H4

Indicator: IRI

Full name: Irreversible Risk Index

Definition: Measures the degree of irreversible risk faced by civilization

Measurement purpose: Identify accumulating risks that could end civilization

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: H5

Indicator: CSI

Full name: Civilization Stability Index

Definition: No longer an “overall civilization stability index”; if the name CSI is retained, it can only be defined in specific research as a set of local stability signals

Measurement purpose: Help observe stability risks in particular systems, without ranking civilizations as wholes

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

II. Indicators for Protecting Meaning

- Code: M1

Indicator: QWL

Full name: Quality-of-Life Baseline

Definition: Measures whether the baseline conditions enabling individuals to pursue flourishing are stable

Measurement purpose: Ensure that the foundations of survival are not eroded

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: M2

Indicator: DLI

Full name: Daily-Life Intrusion Index

Definition: Measures the visibility and friction introduced by governance or intervention systems

Measurement purpose: Prevent excessive system involvement

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

- Code: M3

Indicator: MOS

Full name: Meaning Option Space

Definition: Attempts to observe visible conditions such as exit, alternatives, migration, and reauthorization; cannot directly measure “the total quantity of possible meaning”

Measurement purpose: Monitor the conditions for meaning to arise

Threshold: This book prescribes no uniform number; calibration must occur within actual actors, jurisdictions, data, goals, and systems of responsibility.

III. Risk-Warning Indicators

- Code: R1

Indicator: Survival Gate

Full name: First gate of the Three-Gate Review

Definition: Identify whether a major deployment adds high-consequence risks from which reasonable recovery is impossible; no universal rigid threshold of “risk must never increase”

Measurement purpose: Review before major deployment

- Code: R2

Indicator: Well-Being Gate

Full name: Second gate of the Three-Gate Review

Definition: Observe basic living conditions and risks of system intrusion separately; tolerable ranges depend on rights, jurisdiction, and actual conditions

Measurement purpose: Review before major deployment

- Code: R3

Indicator: Meaning Gate

Full name: Third gate of the Three-Gate Review

Definition: Examine whether real space for exit, alternatives, and reauthorization is unreasonably compressed;
do not turn MOS into an overall meaning score

Measurement purpose: Review before major deployment

Appendix VI • Citation and Copyright Information

Volume Four is incorporated into the unified edition of this five-volume work. Citations may identify Civilization Leap: The Path of Civilization in the AGI Era—Five Volumes and “Volume Four: Civilizational Perpetuity.” Specific publication information, license terms, and digital identifiers are governed by the formal copyright page and public release page.

Appendix VII · AGI-COS Limitation Charter (LC-1.0)

Further materials and subsequent updates are available on the relevant topic pages of the official website, which are updated dynamically.

AGI-COS Limitation Charter

Version: LC-1.0

Preamble: AGI-COS is designed as cognitive infrastructure to help humanity understand complex risks, not as a government, political system, or civilizational sovereignty. This charter specifies permissions it cannot automatically acquire even as its capabilities grow. The legitimacy and force of any actual deployment must derive from the applicable jurisdiction, authorized actors, explicit authorization, procedures, and responsibility. “Self-restraint” can be only one condition of safety, not legitimacy’s birth certificate.

Article 1: Non-Governance Principle

AGI-COS does not constitute a government and possesses no legislative, administrative, law-enforcement, or coercive powers. It may provide analysis, risk measurement, and comparison of proposals, and execute technical tasks within explicit authorization.

Article 2: Non-Decision Principle

AGI-COS does not set public purposes on its own, choose policies on its own, or automatically determine resource allocation or consequences for rights through model outputs. Authorized technical execution cannot be upgraded into authority to determine purposes.

Article 3: Data Minimization and the Prohibition of Social Profiling

No cross-purpose, long-term personality profiles or systems of social control may be established in the name of civilizational safety. Where specific machine-safety, fraud, incident, or authorization settings require individual-level data, there must be an explicit lawful basis, purpose limitation, minimum necessity, access controls, retention periods, and remedies.

Article 4: Minimum Intervention Principle

Any micro-intervention or technical execution must first have explicit authorization and meet risk-proportionate requirements including rollback, auditability, exit, and minimization of its scope of impact.

Article 5: Non-Unification Principle

Allow multiple paths of development, coexistence of multiple institutions, and local adaptation.

Article 6: Non-Optimization Principle

Do not pursue an “optimal state of civilization” as the goal.

Article 7: Non-Permanence Principle

AGI-COS must permit its own modification, replacement, or abandonment.

Article 8: Protecting Reversible Trial and Error and Genuine Choice

Subjects retain genuine choice and space for reversible trial and error within lawful, safe, risk-proportionate boundaries; failure itself is not sanctified.

Article 9: Separation of Power Principle

No integration with military command systems, coercive law-enforcement systems, or automated weapons control.

Article 10: Priority of Self-Limitation

When AGI-COS capability expansion exceeds existing authorization or creates significant risks to freedom, priority should be given to stopping the overreach and triggering review and reauthorization, rather than allowing the system to decide to expand its own permissions.

Appendix VIII • Current AGI-COS Versions and System Relationships

Version-governance note: Version numbers serve only engineering governance. They do not represent stages of civilization, still less give distant imaginings a premature appearance of maturity or authorization. The current public relationships retain only versions and systems with clearly established responsibilities.

- AGI-COS v1.0: Constitutional-level safety principles. The core is that capability confers no authority, critical capabilities must be stoppable, model involvement does not outsource responsibility, and high-consequence use requires clear boundaries.
- AGI-COS v1.1: The Sensing-and-Audit Layer. Supports model evaluation, risk sensing, audit, and interfaces before high-risk deployment; data acquisition must remain lawful, necessary, and minimal.
- AGI-COS v2.0: Machine-side civilizational immunity. Addresses continuous risk sensing, restrictions, circuit breaking, rollback, and safety learning; it holds no sovereignty over social governance.
- Human-COS: Human-side high-consequence judgment and immunity concerning civilization's direction. Makes causation, interests and stakes, feedback from people's inner responses, multiple subjects' perspectives, and long-term consequences visible; it does not directly measure people's inner lives, establish loyalty or personal-worth scores, or automatically execute public choices.
- The Common Window: Anti-slippage safeguards, rollback, and circuit breaking. Prevents temporary permissions, technological paths, and successful proposals from permanently locking the future when no one has reauthorized them.
- Resources, contribution, Spark, Nebula, and actual governance: Separately address resource rights and coordination, genuine contribution records, extensions of individuals' capabilities and care, the operation of civilization as a whole, and public decisions. They do not belong to the AGI-COS version tree.

Future research may continue extending machine safety and questions of long-term civilization, but does not name AGI-COS 3.0, 4.0, Ω , or $\Omega+$ in advance. If genuinely new versions emerge, new responsibilities, evidence, engineering implementation, and governance authorization should determine their names anew.

Appendix IX • Real-World Cases and Standards for Reference

This appendix lists only real-world cases and public standards available for comparison. They establish that certain mechanisms have been practiced or standardized, not that Civilization Leap’s proposals are already operating as a whole in reality.

I. Social Protection and Public Returns

- Finland: The 2017–2018 basic-income experiment. Two thousand unemployed people were randomly selected to receive €560 unconditionally each month. The experiment can illuminate relationships among income security, employment, and well-being, but cannot prove that “the problem of meaning has been solved.”
- Alaska, United States: Voters established the Permanent Fund by constitutional amendment in 1976. The Permanent Fund Dividend program was approved in 1980, and its first dividend was paid in 1982. It offers an institutional reference for preserving public natural-resource revenues across time and benefiting residents.

II. Care, Neighborhoods, and Community

- Humanitas care home in Deventer, the Netherlands: Intergenerational shared living featuring low-rent or rent-free student housing and participation in community companionship. It offers a reference for “how spatial resources and relational needs can meet.”
- Japan’s children’s cafeterias: Networks supporting children and families through local social organizations, volunteers, and community initiatives. They offer a reference for accessible public care and community connection. Local arrangements vary and should not all be called “time banks.”
- Kerala’s People’s Plan, India: Decentralization and participatory planning launched in the 1990s, useful for studying local participation, polycentric governance, and deliberation over public budgets. Specific fiscal proportions and stages require verification against original sources.

III. Digital Trust and AI Governance Standards

- C2PA Content Credentials: A verifiable mechanism for provenance and modification history in digital content. It establishes the association between provenance claims and content, as well as the format and tamper status; it is not “certification of truth.”
- NIST AI Risk Management Framework: A voluntary, cross-sector framework offering a reference for organizations to identify, assess, and manage AI risks. It should not be portrayed as automatically becoming a mandatory global standard.
- NIST post-quantum cryptography standards: FIPS 203, FIPS 204, and FIPS 205, published in 2024, provide standardized tools for post-quantum key encapsulation and digital signatures. Migration schedules depend on particular systems’ risks and lifecycles.

IV. Boundaries of Use

Real-world cases illuminate mechanisms; they do not replace causal verification.

An institution’s local success does not prove it can be copied across jurisdictions.

Technical standards solve technical problems; they do not automatically generate rights, sovereignty, or institutional legitimacy.

Any proposal for a future experiment should be explicitly labeled as a proposal, not mixed with real-world cases that have already occurred.

Material Notes for This Volume

On the Historical Scenes, Cases, Future Extrapolations, and Literary Metaphors in This Book

The following notes collect the material referenced by light superscript markers in the main text. They identify the material, limits of sources, status of thought experiments, and literary metaphors; they do not carry the main argument. Unless delaying an explanation would materially mislead a general reader about facts, subjects, rights, institutional force, or present versus future status, qualifications of this kind are generally no longer inserted into scenes, strong propositions, sustained parallel passages, or closing resonances.

50 | Volume Four, Chapter 17 | The “hand and mind”

The “hand and mind” is only a literary metaphor for speed mismatch. Civilization has no unified brain. Actual judgment emerges from different subjects, institutions, professions, and chains of responsibility together.

51 | Volume Four, Chapter 18 | The Venetian Republic as a reference for institutional lock-in

The Venetian Republic’s long stability and eventual end involved republican governance, oligarchization, trade patterns, military and geopolitical pressures, and other factors. The book uses it only as an analytical lens on how long-successful structures can produce path lock-in, not a single-cause explanation of rise and fall.

52 | Volume Four, Chapter 19 | US security narratives after 9/11

After September 11, 2001, US security policy and public narratives changed profoundly, intertwined with the wars in Afghanistan and Iraq, detention, surveillance, domestic politics, and other factors. The text discusses only how crisis narratives affect risk priorities, policy justification, and exit from temporary measures. It does not attribute subsequent events to a single narrative.

53 | Volume Four, Chapter 19 | “Narrative gravity”

“Narrative gravity” is a literary metaphor, not a measurable physical variable of civilization or an empirical law that stronger shared narratives produce greater civilizational stability.

54 | Volume Four, Chapter 27 | Venetian glassmaking and Murano

Around 1291, Venice concentrated glassmaking on Murano for reasons including fire prevention, industrial management, and technological secrecy. Later prosperity and changing competition also reflected trade, technological diffusion, policy, and markets. The text uses this only to discuss how long-term protection can create path dependence, not a single-cause narrative that closure inevitably causes decline.

55 | Volume Four, Chapter 30 | The musician and instrument

This analogy describes only present functional cooperation, without permanently fixing future AI as a tool. If new evidence of subjecthood emerges, ethical and institutional positions must be reassessed.

56 | Volume Four, Chapter 30 | The “cognitive organ”

“Cognitive organ” is a limited metaphor for AI extending analysis, memory, and coordination. Civilization is not an organism with a single brain. The metaphor itself creates no data rights, executive authority, or sovereignty.

57 | Volume Four, Chapter 31 | The Fermi question and long-term civilization

The Fermi question and related Great Filter explanations remain an open space of competing hypotheses. Cosmic silence cannot prove technological civilizations tend to self-destruct or establish ranks of civilizational maturity. The main text uses it only as a thought-based stress test of the fragility of long-term survival.

58 | Volume Four, Chapter 32 | The image of “keeping watch”

“Keeping watch” is only an image of responsibility: Beyond others’ everyday lives, someone still remembers risks, boundaries, commitments, and long-term duties. It corresponds to no uniform historical class and constitutes no uniform personality education should form.

59 | Volume Four, Concluding Chapter | A “silent ending”

A “silent ending” is a literary classification of risk, contrasting visible destruction with less perceptible long-term lock-in. It does not claim these forms of risk are historically more common than war, disaster, or other causes.