

When Intelligence Is No Longer Scarce

From Expanding Intelligent Capabilities to Shifting Civilizational Scarcities

Full Translation of Chinese V1.1.1

Zijun Fu

Civilization Leap Research Group

Source edition: August 2026

English translation: October 2026

A complete translation of the Chinese Public Essay, Independent Publication Edition. This translation is distinct from the existing E3-09 and E3-10 English adaptations.

Publication and Reading Note

This is a long-form essay for the general reader, intended to be read and assessed on its own. “When intelligence is no longer scarce” is not a statistical conclusion about the world today, nor a claim that everyone already has equal access to the same intelligence. It is a directional thought experiment: as cognitive, analytical, generative, and execution capabilities continue to expand and gradually become widely available foundational capacities, where will civilization’s main bottlenecks move?

The essay’s independent task is to propose and develop a framework of “shifting civilizational scarcities”: as intelligent capabilities become abundant, direction, system-wide causal understanding, time to correct errors before consequences become irreversible, the place of real subjects, and the capacity for shared error correction may become the more decisive scarcities.

“Shifting” does not mean that inequalities in intelligence, computing power, infrastructure, or opportunities for access have disappeared. The growth of capability, whether it is equitably accessible, who controls it, and what purposes it serves are four different questions. This essay asks how, as capability continues to expand, the principal constraints shaping civilization’s consequences move beyond “Can we do it?” toward “Why do it, what will it cause, who bears the consequences, and can errors still be changed?”

The five volumes of Civilization Leap address a complete framework encompassing civilization’s purpose, a pre-institutional language, engineering pathways, civilizational perpetuity, and the foundations of meaning. This essay is neither a summary nor an extract of those volumes, and does not require readers to accept that framework in advance. It takes up just one independent question, how abundant intelligence brings about a shift in civilizational scarcities, and arranges the relevant judgments into a causal argument that can be read, questioned, and discussed within this essay itself.

The few working names used in the essay are explained in ordinary language on first appearance. They refer only to proposed tools that remain open to testing, replacement, or abandonment. The names themselves do not constitute evidence, nor do they denote completed software products or a single overarching system for running civilization.

This essay is not a completed institutional blueprint or an engineering manual. External sources support only particular facts, methods, and institutional experience. The three goals, meaning as the root, human irreplaceability, civilization-level error correction, resistance to closure, and the five civilizational scarcities remain normative judgments and theoretical syntheses proposed here. Future AI subjecthood, the feasibility of complex multi-actor mechanisms, and the actual effects of tools that have not yet entered trial operation remain open questions.

Copyright and Circulation

Author: Zijun Fu | Produced by the Civilization Leap Research Group

Official website: www.civitas.top

Source publication status: Public reading and circulation edition

Source version: V1.1.1 | Chinese Public Essay, Independent Publication Edition

Contents

Before the Main Text | The Real Question Is Shifting

0 | Why Growing Intelligence Changes Civilization's Bottlenecks

1. First, Separate Four Different Questions
2. Capability Never Enters the World Alone
3. More Speed, Less Buffer
4. From Controlling a Tool to Keeping a Shared System Correctable

1 | Why Civilization Needs Intelligence but Cannot Let Intelligence Define It

1. First Ask Why Civilization Exists
2. The Three Goals Are Not Three Interchangeable Scores
3. Meaning as the Root Does Not Give Systems the Right to Define Meaning
4. Intelligence Is a Civilizational Capacity, Not Civilizational Sovereignty
5. From Purpose to Architecture

2 | AI Is Not an External Object: Rebuilding the Shared Human–AI System

1. Moving Beyond “a Single AI System”
2. Directional Judgment, Operational Safety, and Real-World Feedback
3. From Transparent, Multiparty Validation to Low-Friction Operation as Systems Mature
4. An Example at the Intermediate Scale: How AI Can Amplify Human Presence
5. Future Subjecthood Remains an Open Question
6. What a Shared System Really Needs

3 | One Important Use of Powerful Intelligence: Making Consequences Visible in Advance

1. Civilization's Most Expensive Way to Learn Is to Discover Its Mistakes Only After the Consequences Occur
2. From a Single Answer to a Causal Map of Possible Futures We Can Choose Among
3. The Greater the Capability, the More Important the Separation Between Prediction and Decision
4. Causal Exploration Is Not Prophecy
5. Shared Facts Do Not Mean Uniform Thinking
6. Letting Consequences Become Causes Before They Occur

4 | From Judgment to Action: The Best Currently Available Judgment, Not Static Perfection

1. Seeing More Does Not Automatically Eliminate Uncertainty
2. “Best” Can Only Be Conditional
3. Efficiency Comes from Reducing Pointless Friction, Not the Capacity for Judgment
4. Principles of Action Must Be Determined by the Nature of the Consequences

5 | Two Kinds of Error: Reversible Experimentation and Exclusion on Grounds of Irreversibility

1. Civilization Cannot Learn Everything in the Same Way
2. Reversible Experimentation Is Not Unrestricted Experimentation
3. Exclusion on Grounds of Irreversibility Does Not Mean Unlimited Safety at Any Cost
4. The Real Dividing Line Is Not Conservative versus Radical, but Whether We Can Come Back

6 | Reversibility Is Not a Return to the Past: Correction, Reconstruction, and Reauthorization

1. Equating Reversibility with Rollback Is Not Enough
2. Correction Can Take More Than One Direction
3. Correction Capacity Demands More Than a Nominal Right to Exit
4. In the Age of Powerful AI, Correction Capacity Is Itself Infrastructure

7 | Anti-Closure Does Not Mean Refusing to Decide: Protecting the Capacity for Future Choice

1. Anti-Closure Opposes Finality, Not Resolve
2. A Larger Space of Options Is Not Always Better
3. The Most Dangerous Forms of Closure Often Arrive in the Name of Success
4. An Open Future Is Compatible with Boundaries Against Irreversible Harm

8 | What Should Be Governed? Purposes, Permissions, Conduct, Responsibility, and Consequences

1. “Regulating AI” Is Too Broad a Formulation
2. Begin by Answering Five Distinct Questions
3. Capability Itself Can Still Be a Direct Object of Governance
4. A System's Name Must Not Obscure the Chain of Responsibility
5. Who Regulates, and Why, Must Also Be Subject to Oversight

9 | Why Many AI Governance Problems First Bring Humanity Itself into View

1. To Understand What AI Amplifies, First Look at What It Connects To
2. Structure Before Personality
3. Theory and Tools Cannot Substitute for Real-World Mechanisms
4. Upstream Structures Also Include Distribution and Capture
5. This Does Not Eliminate Machine-Side Risks
6. Two-Sided Stability Is More Realistic Than Seeking One Side That Is Always Right

10 | After Powerful AI, What Remains Irreplaceable About Human Beings?

1. Do Not Ground Human Value in What Machines Cannot Yet Do
2. Human Beings Do Not Need to Be Preserved for Their Inefficiency
3. AI Can Extend Our Presence, but Cannot Turn Life into Something Lived by Proxy
4. Meaning Is Not the Last Job Left for Human Beings
5. What Is Truly Scarce Begins to Shift

11 | A Mature Relationship: The Stronger a System Becomes, the Better It Knows When to Step Back

1. Civilizational Maturity Does Not Mean Systems Reaching Ever Deeper into Life
2. Stepping Back Means Less Intrusion, Not Less Capability
3. Low Intrusion Is Not Invisible Control
4. Stepping Back Is a Reversible Reduction in Intervention, Not a One-Time Withdrawal
5. A Civilizational Logic Truly Takes Shape Only When It Becomes Self-Sustaining
6. Stepping Back Brings This Essay Back to Its Starting Point

12 | Conclusion: As Scarcity Shifts, What Must Civilization Truly Safeguard?

1. Returning to the Opening: Scarcity Has Not Disappeared; It Has Shifted
2. Direction: The Stronger the Capabilities, the More We Need to Know What They Are For

3. System-Wide Causality: Local Intelligence Cannot Replace Understanding of the Whole
4. Time for Correction: The Faster We Move, the More Precious the Window for Changing Outcomes
5. Real Subjects: The Stronger the System, the More Essential It Is That Life Remain Present
6. Shared Correction: No One Can Be Presumed Infallible
7. AI Governance Is Thus Placed Back within the Question of Civilization
8. What Must Civilization Truly Become Stronger At?

Limits of the Evidence, Limits of the Mechanisms, and Open Questions

Further Reading and Paths into Collaborative Research

Endnotes

Appendix | Personalized Further Reading Guide

Before the Main Text | The Real Question Is Shifting

For much of its history, civilization was constrained first by what it could not do. Knowledge was difficult to obtain, complex systems were difficult to understand, many possible paths could not be compared, and many aspirations remained imaginary because cognitive, organizational, and execution capacities were inadequate. Intelligence was therefore, for a long time, a scarce resource that defined the limits of action for individuals, organizations, and nations.

Today, that premise is beginning to loosen. Artificial intelligence is enabling cognitive, analytical, generative, and execution capabilities to grow, replicate, and enter the real world at an unprecedented pace. They remain unevenly distributed and are far from being public capacities that everyone possesses equally. Yet a new question has already emerged: if more and more activities are no longer constrained by not knowing how to do them, will civilization naturally know what it ought to do?

Not necessarily.

We may have more and more answers while still lacking direction; see faster without necessarily seeing the whole; act more rapidly while losing the time to choose again after discovering a mistake. Systems may increasingly do things for people while, amid the convenience, people gradually withdraw from their own lives. Every node may become smarter without civilization as a whole becoming better able to correct errors together.

This essay calls that change a “shift in scarcity” for civilization.

Here are five coordinates for the reader: direction, system-wide causal understanding, time to correct errors before irreversibility, the place of real subjects, and the capacity for shared error correction.

These five names are neither slogans nor a completed policy checklist. For now, they are only the legend of a map. Why they become scarcer as intelligent capabilities expand, how they are causally connected, and why they cannot be replaced by an aggregate score, an authoritative actor, or an overarching system are the arguments the following twelve chapters must establish.

This is also why the essay can stand independently of the five volumes of Civilization Leap. Those volumes develop a complete civilizational framework; this essay follows a particular change: as intelligence gradually moves from a scarce resource to a foundational capacity, where do civilization’s bottlenecks cease to lie, and where do they emerge instead?

The real question here is therefore not how powerful AI will ultimately become. It is whether, when capability is no longer the scarcest thing, civilization is ready to confront the scarcities that capability cannot generate on our behalf, yet that determine where every capability ultimately leads.

0 | Why Growing Intelligence Changes Civilization's Bottlenecks

1. First, Separate Four Different Questions

Growth in AI capability, whether that capability is equitably accessible, what authority a system receives, and whether civilization progresses as a result are four different matters. Stanford University's Artificial Intelligence Index Report 2026 documents precisely this coexistence of rapidly expanding capabilities and the continued concentration of key resources. Producing more intelligence does not mean that every person, nation, and community already possesses it equally; still less does it mean that capability itself has acquired the standing to determine public purposes.¹

"No longer scarce" in the title is therefore not a statistical declaration about the present. It is a directional hypothesis: if cognitive, analytical, generative, and execution capabilities become increasingly accessible, and constraints once imposed by insufficient capability begin to loosen, what will then actually prevent civilization from reaching better outcomes?

2. Capability Never Enters the World Alone

AI never acts in a vacuum. When a capability enters the real world, it always connects to some goal, some incentive, a degree of authority, a chain of responsibility, and specific people affected by it. The same kind of capability can help humanity identify disasters in advance or execute mistaken goals faster; it can break through information silos or amplify prejudice, fear, and competitive impulses to previously unreachable scales.

Consequences depend not only on how powerful a capability is, but on how it combines with human purposes, institutions, power, responsibility, and feedback structures.

This requires rejecting two simplifications at once: capability growth itself cannot be pronounced a catastrophe, and technological progress itself cannot be equated with civilizational progress. Machine-side failures, autonomous behavior, adversarial attacks, and cascading risks are real problems. Human-side war, short-term incentives, concentrated power, and confusion over direction are real problems too. Neither cancels out the other; in real systems, they may instead amplify one another.

3. More Speed, Less Buffer

AI brings more than greater intelligence. It also brings greater speed and reach, and a greater ability to translate continuously into action. In the past, a mistaken judgment might have remained local, been delayed, or been limited by inadequate execution capacity. When powerful intelligence connects to organizations, markets, infrastructure, and public authority, the distance from mistaken judgment to consequences becomes much shorter. Civilization gains execution capacity, but may lose the time for correction that slowness and friction once provided.

The same capability also opens up the opposite possibility: comparing more paths, bringing hidden interests and feedback into view sooner, and changing choices through simulation, trials, and dissent before real harm occurs. Powerful intelligence can increase not only the speed of reaching outcomes, but also civilization's ability to see those outcomes before arriving at them. The crucial question is whether the former overwhelms the latter.

4. From Controlling a Tool to Keeping a Shared System Correctable

A complete account of human-machine civilization therefore has at least two sides. The directional side asks: Why do this? Whose lives will it change? Which interests will it amplify? Is there a better path? The operational side asks: Can it be executed safely? Does it exceed its authority? Can it still be stopped, audited, taken over, and corrected? Looking only at operations would label a system safe even when its direction is wrong and its execution perfect. Looking only at direction could allow noble purposes to cross real-world boundaries through technical danger.

People cannot be reduced here to a formal role of clicking the final button, nor can AI be reduced to an external tool waiting for commands. Both have already entered a sociotechnical relationship made up of multiple actors, institutions, systems, and real-world feedback. What must be protected is not the permanent correctness of either side, but the assurance that when either makes an error, that error cannot pass straight through into irreversible reality.

This leads to the essay's first judgment: the continuing growth of AI capability is not, by itself, a sufficiently complete account of the problem. What truly determines civilizational consequences is how greater capability combines with purposes, institutions, power, responsibility, understanding, and feedback. The following chapters develop this argument through civilization's purpose, the shared human-AI system, making causality visible, reversibility and irreversibility, the boundaries of responsibility, the human place, and systems stepping back, before returning to the five scarcities introduced at the outset.

1 | Why Civilization Needs Intelligence but Cannot Let Intelligence Define It

1. First Ask Why Civilization Exists

Any institutional design for AI that does not first answer why civilization exists risks gradually elevating means into ends. Systems can become faster, more accurate, and more stable; productive, knowledge-generating, and coordinating capacities can keep expanding. Yet if whom those capacities ultimately serve remains unclear, the more mature the technology becomes, the more easily efficiency can conceal the absence of purpose.

A future civilization that answers only how resources should be distributed, how systems should coordinate, and how risks should be reduced still has not answered whom those capacities ultimately serve. The answer cannot be an abstract system itself, or a “whole” detached from concrete lives. Civilization exists in real people, relationships, communities, and those who come after us.

This is also the point of the three goals. Civilization must first ensure that basic deprivation does not strip people of dignity and choice, and enable life to be real, free, rich, connected, and enjoyable. Once survival and daily living are better supported, life must still be able to keep generating meaning through relationships, creation, responsibility, exploration, and existence itself. In its most concise form: freedom from survival insecurity → a happy life → a more meaningful life.

Two levels must be distinguished here. Continued existence is civilization’s minimum boundary, the lowest prerequisite for any civilizational value to remain possible. Freedom from survival insecurity is the first of the three goals. The two must not be mechanically conflated.

2. The Three Goals Are Not Three Interchangeable Scores

The progression of the three goals is not an aggregate utility function. A higher overall probability of continued existence cannot automatically offset persistent infringements of some people’s basic rights. An increase in average happiness cannot make freedom of choice a cost to be paid. Nor can particular people be required to accept deprivation, suffering, or sacrifice in the name of “greater meaning.”

This boundary is especially important for powerful AI, because more powerful intelligence is better at optimization, comparison, and aggregation. If civilization first compresses its goals into a single score, stronger AI will only execute that compressed objective more effectively to its conclusion. What matters is not teaching AI more words for values, but ensuring that human institutions do not convert things that are not interchangeable into an aggregate that can be automatically optimized.

An upstream discipline of AI governance should therefore be this: capability may help compare consequences, but it must not make nonexchangeable civilizational goals exchangeable again. The greater the computational capacity, the more important it is to keep the hierarchy of goals and the boundaries of rights clear before computation begins.

When the three goals conflict in real situations, this essay offers no formula for automatic adjudication. It requires first making the conflict visible: Are the boundaries of continued existence or basic rights being crossed? Who benefits, and who bears the losses? Are the burdens voluntary, reversible, and repairable? Are there less harmful alternatives? Who has the authority to decide, and can those affected appeal? For example, even if an AI deployment raises average efficiency, “better overall” is still an insufficient justification if it persistently shifts errors that cannot be appealed onto a minority.

3. Meaning as the Root Does Not Give Systems the Right to Define Meaning

This essay proposes that meaning is the root, while its sources may differ. Meaning may come from relationships, loving and being loved, creation, care, responsibility, exploration, memory, and shared living, or from life experiences that no one else can live in a person’s place. It is not another name for satisfaction, click-through rates, social influence, contribution points, or subjective pleasure.

This claim also limits the power of systems. Present-day advanced AI may take on functional roles in understanding, protecting, and transmitting meaning, and in extending human capabilities, only in specific deployments with traceable responsibility, explicit authorization, auditability, correctability, and the possibility of exit. Such roles do not establish an ontological status, still less public authority.

“Meaning as the root” therefore does not put AI back in the position of a low-level tool. On the contrary, it allows AI to take on increasingly complex roles in understanding, memory, expression, care, creation, and the extension of human capability. But it requires systems always to recognize that understanding meaning is not possessing it, and helping someone express values is not deciding for them what makes life worth living.

People must not be reduced to a formal “human-in-the-loop” within powerful AI systems. This is not because people will always calculate better. Civilization’s purpose, the experience of living, the bearing of meaning, and real-world consequences have never amounted simply to a final approval. As long as real subjects live, bear consequences, form relationships, and make choices within civilization, people are not safety switches in the loop. They are among the reasons the loop exists.

4. Intelligence Is a Civilizational Capacity, Not Civilizational Sovereignty

Technical capability, price, contribution, social recognition, cultural representation, and other forms of status cannot automatically confer public authority or dominion over

meaning. Both affirming future artificial subjecthood and denying it permanently carry a burden of evidence. New evidence may trigger reconsideration, but cannot automatically settle legal status, governing authority, or access to resources.

At the level of existing institutions, UNESCO's Recommendation on the Ethics of Artificial Intelligence likewise places human dignity, accountability, auditability, and human oversight at the center of AI governance, and requires that AI systems not displace ultimate human responsibility and accountability. This supports the current institutional structure of responsibility; it cannot be used to prove that nonhuman subjects can never emerge in the future.²

Civilization can therefore be highly active in expanding intelligent capabilities without expanding the political power of intelligent systems in parallel. A system may know more, reason further ahead, and offer deeper assistance while still acting only within explicit authorization and structures of responsibility. There is no necessary one-to-one relationship between growth in capability and growth in power.

This also corrects humanity's own assumptions. Civilization does not become more advanced simply because it possesses more powerful AI. A more precise test is whether civilization can make increasingly powerful intelligence serve life without allowing that capability, in turn, to define life. That capacity is itself part of civilizational maturity.

5. From Purpose to Architecture

Only after civilization's purposes, goals, and boundaries of meaning have been clarified can AI's place truly be discussed. It is neither an external species inherently obliged to obey nor a new center inherently entitled to sovereignty. It is, first, a capability entering civilization: one that can extend perception, causal exploration, expression, action, and cooperation.

2 | AI Is Not an External Object: Rebuilding the Shared Human-AI System

1. Moving Beyond “a Single AI System”

Once AI enters the real world, it is not an external species separate from society. It is embedded in organizations, markets, families, public systems, personal relationships, and transnational networks, connecting human-defined goals, institutional authorization, resource conditions, and real-world consequences.

Seen together, these relationships reveal an important fact: the arrangement has never simply been “one AI system decides, and humans approve at key points.” The actual structure is a sociotechnical system composed of multiple subjects, multiple systems, multiple layers of authorization, overlapping responsibilities, and real-world feedback. AI is simply one increasingly important and powerful category of capability within it.

A “human in the loop” may therefore be a necessary design feature in certain engineering processes, but it should not be elevated into the definitive picture of civilization. People matter in civilization for reasons beyond occupying a position in a loop; nor does AI gain legitimacy merely by being placed in one. The basic units of civilizational relations should be real subjects, shared purposes, capabilities that can be authorized, structures of responsibility, and real-world feedback.

2. Directional Judgment, Operational Safety, and Real-World Feedback

Directional judgment asks, “Why do this? How would it change people and the world? Is there a better path?” Machine-side operational safety asks, “Can this be done safely? Does it exceed its authority? Can it still be stopped, taken over, audited, and rolled back?” The former must not acquire final authority over value judgments or the authority to execute automatically. The latter must not acquire the authority to determine policy goals, govern society, or exercise permanent control in the name of safety.

Real-world feedback does not constitute a third, abstract sovereignty over judgment. Its role is to bring choices that have been assessed, decided upon, and authorized into reality, then return actions, benefits, harms, people's actual experiences, institutional friction, appeals, failures, repairs, and new facts to the judgment process. The real world is not the endpoint of a system's output; it is the input to the next round of judgment.

In Civilization Leap's related research, Human-COS, a framework for augmenting human directional judgment and causal exploration, and AGI-COS, a machine-side operational safety framework, refer respectively to the first two functions. They are only proposed tools that still require ongoing validation. They are not subjects acting in the real world, much less entities entitled to sovereignty over civilization by virtue of their expertise.

Human–AI symbiosis should therefore not mean encoding human values in a model and then letting that model run civilization on humanity's behalf. It is closer to a structure that maintains stability on both sides: people can be mistaken about direction, and machines can make mistakes in perception, reasoning, or execution. True stability does not mean that either side is always right. It means that when either side goes off course, the system still has opportunities to see afresh, reassess, renew authorization, stop, and change the outcome.

A shared system cannot eliminate error. What it really needs to do is prevent an error on either side from passing straight through into irreversible reality. Human fears, interests, and misjudgments must not become accomplished facts simply because they have gained AI's speed; nor may highly confident machine outputs bypass purpose, authorization, and responsibility simply because they are more efficient.

3. From Transparent, Multiparty Validation to Low-Friction Operation as Systems Mature

When a new system is first established, the links between judgment, permissions, and consequences have not yet been adequately validated. Transparency, participation by multiple subjects, expert review, counterexamples, independent replication, and records of responsibility therefore matter all the more. Tools for directional judgment need to compare multiple subjects, pathways, and timescales, and actively seek the strongest counterexamples and alternative explanations. Machine-side safety tools need to disclose sources, confidence levels, areas of uncertainty, and limits of authority.

This does not mean that a mature system must always retain the same number of human checkpoints and approval layers. Research on meaningful human oversight points out that oversight can be distributed across design, operation, and structures of responsibility; it cannot be reduced to real-time human intervention at every step. Conversely, a nominal “human in the loop” can degenerate into a rubber stamp. As certain mechanisms are validated through long-term real-world use, repetitive processes can reduce the cost of interaction, machines can take on more routine operations, and human attention should focus more on anomalies, changes in goals, conflicts of values, irreversible consequences, and the boundaries of authorization.³

“Fewer steps” here does not mean less responsibility. Procedures can become less dense, but the integrity of authorization, feedback on outcomes, traceability of responsibility, and the ability to reopen judgment must not disappear with them. This proposition will be developed in Chapter 4, on “the best currently available judgment,” and Chapter 11, on “systems stepping back.”

In other words, efficiency and legitimacy need not be adversaries. A truly efficient system does not reopen a discussion of every civilizational principle for every action. It moves

validated rules into the background and reserves human judgment for what is genuinely changing.

4. An Example at the Intermediate Scale: How AI Can Amplify Human Presence

One concrete example at the intermediate scale is amplification rather than replacement of people. As an extension that can be authorized, AI can help real people bring goodwill, care, capabilities, and judgment to more real people across time, distance, and forms of interaction. The subject here is not a manufactured AI persona. It is the person who grants authorization, bears responsibility for the relationship, and receives its real-world responses.

This example belongs in this essay not because it represents all human–AI relationships, but because it makes an abstract principle concrete: AI can greatly expand a person's capacity to be present without taking subjecthood away from that person. A system can help someone do more without making that “more” the reason for the system's own existence.

This also suggests a direction for a broader human–AI civilization. More powerful AI need not prove its value by becoming ever more humanlike and occupying more of the positions humans once held. Another possible path is that the stronger AI becomes, the further it can extend human capabilities, care, understanding, and action, while allowing real subjects to remain themselves.

5. Future Subjecthood Remains an Open Question

Positioning AI as a layer of capabilities that can be authorized within today's civilization does not mean permanently denying the possibility of future AI subjecthood.

The evidence currently available to the public is insufficient to establish that advanced AI has subjective experience, nontransferable continuity as a subject, or awareness of the consequences it bears. But this proves neither that artificial subjects will necessarily emerge in the future nor that they can never emerge. Both conclusions continue to carry a burden of evidence.

“AI entering the capability layer of human civilization” is therefore an institutional positioning for the present stage, not a final ontology. If sufficient evidence of new nonhuman subjects emerges in the future, questions of protection, legal standing, shared life, and political relations will need to be revisited.

Yet even if new real subjects emerge, this will not automatically alter another boundary established in this chapter: subjecthood does not in itself confer public office, and public office does not confer sovereignty. Who has capabilities, who has experiences, who deserves protection, and who may make decisions on behalf of a community remain distinct questions.

6. What a Shared System Really Needs

For a shared human–AI system to achieve both sound judgment and efficiency, at least four things must remain in place: purposes must be traceable, authorization must be explainable, real-world outcomes must feed back into the system, and errors must be able to reopen judgment.

These four conditions matter more than whether humans remain involved in every operational step. People can step back from hands-on operations in low-risk, repetitive, validated processes. They must reenter when purposes change, rights conflict, risks carry grave consequences, or irreversible boundaries approach. AI can take on more and more routine capabilities, but it must not allow a purpose that is no longer valid to remain in execution simply because automation runs so smoothly.

3 | One Important Use of Powerful Intelligence: Making Consequences Visible in Advance

1. Civilization's Most Expensive Way to Learn Is to Discover Its Mistakes Only After the Consequences Occur

For irreversible outcomes at the scale of civilization, we cannot wait for the outcomes themselves to provide final validation. The value of exploring consequences is not to prove that AI judges better than humans. It is to help people see their choices earlier, more fully, and more openly, while the consequences have not yet occurred, options remain available, and actions can still change.

This shifts the value of AI's capabilities from “giving people answers” to “changing what people can see before they decide.” Many major mistakes are not entirely without warning before their consequences arrive. Rather, facts, interests, power, expert knowledge, and feedback about people's feelings are scattered among different subjects and have never been brought together in a single causal map.

If powerful AI can bring civilization a genuinely important new capability, it is to reduce, as far as possible, the proportion of consequences understood only after they have become history. The earlier consequences enter judgment, the more likely we are to preserve options while the future can still be changed.

2. From a Single Answer to a Causal Map of Possible Futures We Can Choose Among

Causal exploration for major public decisions should not merely produce a risk score. At a minimum, it should make visible, together, the subjects and their intentions, the main facts, the distribution of benefits and harms, key feedback, tipping points, alternative pathways, unknowns, the strongest objections, conditions for falsification, and where responsibility lies. It must also be open to adversarial challenge, independent replication, and testing against real-world outcomes.

The ultimate product of such exploration should be a “causal map of possible futures we can choose among,” repeatedly testable against reality and by different subjects and different models, rather than an answer that AI supplies on people's behalf.

This matters because complex social problems often lack not advice but structured visibility: whom an action would affect; what would happen if no action were taken; where delayed, partial, joint, or alternative approaches would each take the system; which risks are approaching a tipping point into qualitative change; and which supposed gains merely transfer losses to third parties or the future.

One of the most valuable directions for more powerful AI is to make these maps broader, deeper, and faster to produce, while ensuring that they remain maps rather than commands.

3. The Greater the Capability, the More Important the Separation Between Prediction and Decision

A model's output should first be treated as advice that informs understanding, accompanied by its sources, confidence levels, and a statement of what remains unknown. Subjects, rights, authorization, and responsibility must be kept distinct. Capability does not confer sovereignty on a system, and facts and recommendations do not automatically become decisions.

Similarly, the output of a directional-judgment tool should not be approval or rejection. It should change what can be known and understood before a decision is made. The final choice remains with subjects who hold real-world rights, responsibilities, and authorization.

This is not distrust of AI's capabilities. It is a distinction between different kinds of questions. Predictive accuracy answers, "How accurately does the system foresee a particular aspect of the future?" Decision-making authority answers, "Who is entitled to determine who must bear a particular consequence?" Even exceptionally high predictive accuracy does not automatically create that authority.

Precisely because future AI may become increasingly adept at prediction, diagnosis, and comparing options, we need to separate these questions in advance. Otherwise, the more successful improvements in capability become, the easier it will be for expert advantage to turn unnoticed into de facto power.

4. Causal Exploration Is Not Prophecy

Reliable causal exploration must actively seek the strongest counterexamples, alternative explanations, and pathways that could overturn the original conclusion. It must also enable different models, implementations, and subjects to independently replicate its key analyses. What needs to be assessed is not which model sounds most certain, but which directions keep reappearing when assumptions change, and which disagreements remain unresolved.

"Making consequences visible in advance" must therefore not be taken to mean predicting a future that is already fixed. Social systems involve feedback, adaptation, contingencies, hidden variables, and learning by subjects. Many pathways will change simply because they have been seen. Genuine foresight includes articulating what is unknown rather than eliminating uncertainty.

A mature system for causal exploration should allow three kinds of output to coexist: what we know with reasonable confidence; which pathways we believe most warrant caution or pursuit; and what we still do not know, including what new facts would compel us to revise our judgment.

The significance of more powerful AI for civilization lies not in turning an open future into a computable answer, but in reducing the consequences that could have been foreseen yet

were missed because of fragmented information, concealment by power, cognitive limitations, or short-term pressure.

5. Shared Facts Do Not Mean Uniform Thinking

Public consensus need not depend on uniform thinking. It can emerge gradually from jointly testable facts, causal relationships, benefits and harms, and continued shared existence. Causal exploration can adopt different subjects' perspectives to understand their circumstances, but it must not thereby erase overall consequences or continued shared existence.

This means that powerful AI used in public judgment should not be tasked with “unifying everyone's values.” It is better suited to helping different subjects see whether, despite continuing differences in values, institutions, and interests, certain consequences would destroy everyone's room to choose; whether certain forms of cooperation serve different subjects' long-term interests at the same time; and whether some conflicts are wrongly regarded as inevitable because the structure of information is flawed.

Causal transparency is not uniform thinking. On the contrary, a sufficiently transparent causal system should make disagreements more precise: are they disagreements about facts, causal relationships, values, risk tolerance, or who has the right to decide? Only when these disagreements are disentangled can consensus cease to depend on coercion.

6. Letting Consequences Become Causes Before They Occur

In conventional decision-making, consequences usually follow action: first we act, then observe what happens, then bring that experience into the next action. The civilizational capability discussed here allows future consequences to enter present judgment before action, in the form of causal exploration, counterfactuals, tipping points, and alternative pathways.

This can be understood as an important change in temporal structure: “What will happen if we continue this way?” becomes part of the choice we make now, rather than merely an explanation after the event. The consequences have not yet occurred, but sufficiently clear causal evidence can still allow them to change our present direction.

This capability is meaningful, however, only within two boundaries. First, it must remain open to refutation, replication, and correction by real-world outcomes. Second, superior causal exploration must not confer the authority to bear consequences in place of the real subjects concerned.

4 | From Judgment to Action: The Best Currently Available Judgment, Not Static Perfection

1. Seeing More Does Not Automatically Eliminate Uncertainty

Chapter 3 identified one of AI's important values for civilization as expanding the visibility of causal relationships and consequences. But seeing more does not mean waiting until everything can be seen clearly. The real world always contains missing information, model error, subjects' responses, contingencies, and time pressure. A civilization that makes the disappearance of all uncertainty a precondition for action will never be able to act in many situations that truly matter.

A single causal analysis is not an eternal conclusion. After real-world action, benefits, harms, people's experiences, institutional friction, third-party responses, appeals, failures, and new facts must return to the judgment process. Each must be examined to determine which assumptions were supported, which causal relationships were overturned, which unknowns grew, and which risks need to be upgraded or downgraded. An important capability of a framework that can correct itself is not being right forever, but ensuring that errors can be preserved, reproduced, attributed, and allowed to genuinely change the next operation.

This changes the relationship between judgment and action. Judgment is no longer a one-time approval before action, and action is no longer execution that cannot be revisited after judgment. The two form a cycle: arrive at the best-supported judgment available now, bring it into reality, obtain new facts from that reality, and let those facts change the next round of judgment.

2. “Best” Can Only Be Conditional

This essay expresses that principle as “the best currently available judgment”: once the subjects, facts, benefits and harms, objections, alternative pathways, and unknowns have been fully examined, choose the most reasonable actionable option under the current conditions of information, time, and responsibility, rather than waiting for a perfect answer detached from time.

“Best” here is not a global optimum declared by a system, much less the compression of the three goals, rights, and irreversible harms into a single aggregate utility function. It is simply best-available judgment: provisionally valid on the current evidence, with reasons that can be explained, conditions that can be traced, and conclusions that can be revised when new facts emerge.

Any formula claiming to automatically calculate “net civilizational benefit” will conceal distinctions among beneficiaries, those harmed, irreversible consequences, and alternative

pathways that cannot automatically be traded against one another. Real-world decisions must make these dimensions visible separately. An optimum for the whole does not automatically create obligations for individuals.

Seeking “the best currently available judgment, not static perfection” does not lower standards of rigor. It shifts rigor away from the illusion of getting everything right in one attempt and toward something more realistic: judgments have grounds, actions generate feedback, errors leave a record, and new facts can change the next choice. Research on decision-making under deep uncertainty likewise emphasizes that the value of computation lies not only in pursuing a more accurate single prediction, but in stress-testing many possible futures and finding strategies that are more robust across scenarios and can adapt to new information. ⁴

3. Efficiency Comes from Reducing Pointless Friction, Not the Capacity for Judgment

When a system is first established, new causal relationships, boundaries of authority, and real-world consequences require more intensive validation. Once certain low-risk, repetitive mechanisms have been tested over a long period in reality, every routine action need not undergo a full deliberation on civilizational principles. Research on Dynamic Adaptive Policy Pathways offers a related methodology: take explainable near-term action while preserving long-term options, and use triggers and ongoing monitoring to determine when to change course.

Rules can stabilize, processes can be automated, and interaction can decrease. ⁵

What should decrease, however, is repetitive friction, not civilization's capacity for judgment. A system can operate efficiently when its purpose has not changed, its permissions have not expanded, its domain has not materially shifted, and its risks have not increased. As soon as those conditions change, judgment must be reopened.

This anticipates the later principle of systems stepping back: the density of rules is not the density of interaction, and complete authorization does not mean repeated authorization. A mature system can remain rigorous behind the scenes while keeping the interface simple.

4. Principles of Action Must Be Determined by the Nature of the Consequences

There is, however, a boundary to “the best currently available judgment” that cannot be ignored: if an action could directly cross a tipping point from which recovery is impossible, real-world feedback would come too late. In such a case, we cannot rely on the learning logic of “try it first, then adjust in light of the results.”

5 | Two Kinds of Error: Reversible Experimentation and Exclusion on Grounds of Irreversibility

1. Civilization Cannot Learn Everything in the Same Way

The more irreversible the consequences, the faster they spread, and the shorter the window for intervention, the less we can afford to wait for outcomes to validate our judgments. Irreversible risks at the scale of civilization cannot be finally validated by allowing irreversible outcomes to occur. Methodologically, this echoes the precautionary logic of the United Nations Framework Convention on Climate Change: where serious or irreversible harm is threatened, a lack of complete scientific certainty should not be a reason to postpone cost-effective measures. ⁶

This principle may appear to concern only risk management, but it actually changes how civilization learns. In ordinary life, products, institutions, and local mechanisms, many errors can be corrected through real-world feedback. But once some consequences occur, subjects may die, structures of rights may be permanently closed off, wars or technological chains may become impossible to call back, and ecosystems or social trust may cross tipping points from which recovery is extraordinarily difficult. In these cases, the idea that “practice is the test” must not be crudely interpreted to mean that a catastrophe has to happen once for real.

This essay therefore divides real-world action into two domains governed by different principles: a domain of reversible experimentation and a domain of exclusion on grounds of irreversibility. The former allows learning from reality within controlled boundaries. In the latter, until the conditions for lifting the prohibition on entry have been met, information must not be obtained through irreversible real-world consequences. Learning must instead take place earlier, through causal exploration, simulation, historical backtesting, red-teaming, sandboxes, shadow operation, low-risk prototypes, and independent replication.

2. Reversible Experimentation Is Not Unrestricted Experimentation

Reversibility does not mean zero cost. A policy that can be withdrawn may still cause real losses, and a model deployment that can be stopped may still harm particular people. Reversibility means only that, once an error is discovered, practically viable paths remain for stopping, containment, repair, compensation, or rebuilding.

Reversible experimentation therefore requires at least four conditions: a sufficiently limited scope, identifiable harms, a genuine right to stop, and feedback that travels faster than risk spreads. If any one of these is missing, a supposed pilot may merely shift a high-consequence experiment onto people who do not know they are being experimented on.

Reversibility is not a blank check for bold action. It places conditions of responsibility on learning. The more a system emphasizes rapid experimentation, the more it must demonstrate that it can genuinely stop, recognize those affected, preserve a record of errors, and take responsibility for repair, rather than merely showing that it can launch quickly.

3. Exclusion on Grounds of Irreversibility Does Not Mean Unlimited Safety at Any Cost

Strict treatment of irreversibility must not be expanded into the claim that whenever someone declares a risk enormous, unlimited resources, unlimited surveillance, and an indefinite suspension of social life become permissible. That would allow risk prevention itself to create new, irreversible structures of power.

Safety needs do not automatically justify unlimited surveillance, resource requisition, or permanent emergency powers. A crisis can change the pace of action and the intensity of procedures, but temporary powers must not be written into the normal conditions of stable times.

What demands strict treatment is not the word “safety,” but irreversible consequences analyzed in concrete terms: what the consequences are, whom they affect, how quickly they spread, where the last reversible point lies, what alternative pathways exist, and what evidence could overturn the current judgment.

At boundaries involving extremely grave consequences, some preventive costs may indeed be worth bearing. But “every cost is worth paying” cannot become a general principle. The proper formulation is this: the more severe the irreversible consequences and the stronger the evidence, the more reason civilization has to bear higher upfront costs of protection while direction can still be changed.

4. The Real Dividing Line Is Not Conservative versus Radical, but Whether We Can Come Back

An apparently radical experiment may be safer than an apparently conservative institution that quietly entrenches power, provided the experiment is limited in scope, can be stopped, has transparent results, and offers a genuine exit. Conversely, a mechanism nominally described as only a “temporary measure” may already have entered irreversible territory in civilizational terms if, once its data, permissions, dependencies, and behavioral patterns are established, there is no real way to exit them.

So, when judging whether an action should proceed, we should ask not only “How great is the risk?” but also “If we are wrong, can we still change the outcome?” This is precisely why the next chapter redefines reversibility.

6 | Reversibility Is Not a Return to the Past: Correction, Reconstruction, and Reauthorization

1. Equating Reversibility with Rollback Is Not Enough

Reversibility means more than having a rollback button. Purposes, rights, authorizations, data, resources, capabilities, roles, and temporary control should all have ways to end or revert, appropriate to the object concerned. When opening interfaces with high-consequence effects, provisions for withdrawal, suspension, expiry, reversion, repair, and termination must be designed at the same time. Restoration does not revive old permissions; restarting requires fresh judgment and fresh authorization. Existing AI risk-management standards already identify continuous monitoring, clear responsibility, safe decommissioning, and phaseout as requirements for risk governance.⁷

This already shows that genuine reversibility does not mean restoring the entire world to a historical snapshot. Reality does not actually run backward: what has happened remains, relationships change, knowledge grows, and mistakes themselves become part of the new circumstances.

At the level of civilization, then, reversibility is more accurately understood as correction capacity. What it protects is not a “return to the past,” but the practical ability to change direction before irreversible consequences occur.

2. Correction Can Take More Than One Direction

Once an error is discovered, the right response may be rollback, but it may also be suspension, downgrading, isolation, replacement, repair, redesign, upgrading, separation into parts, reconstruction, or termination. If the old state was itself the source of the problem, restoring it in full merely replaces a new failure with an old one.

Civilizational transformation likewise cannot be accomplished through a single replacement. It can neither declare all existing institutions invalid nor return to the starting point after every failure. Instead, it must take shape gradually, while the old world still exists, through experimentation, failure, appeals, rollback, bearing responsibility, and choosing again.

Rollback in its technical sense must be retained here. Backups, version restoration, low-tech alternatives, and human takeover remain important tools. Reversibility at the scale of civilization simply cannot be reduced to technical recovery.

3. Correction Capacity Demands More Than a Nominal Right to Exit

A system's terms may say that “users can exit,” yet in practice leaving may be almost impossible because of data dependencies, network effects, tied resources, identity

relationships, or the disappearance of alternatives. Conversely, an institution may not be designed to return to its former state, yet still have a genuine capacity for replacement, reconstruction, and reauthorization.

Whether exit is real, therefore, cannot be judged solely by the presence of a button. We must ask whether those leaving can still sustain their lives, whether they can obtain alternative services, whether old permissions truly lapse, whether data and resource relationships end, and whether reentry requires new authorization.

Historical facts may remain, but past permissions cannot use that history to perpetuate themselves. Nor can temporary control continue automatically simply because it was once justified.

4. In the Age of Powerful AI, Correction Capacity Is Itself Infrastructure

AI increases the speed of execution and also allows dependencies to form more quickly. The more powerful the system, the less correction can be treated as an after-the-fact patch. Stopping, isolation, explanation, human takeover, alternative capabilities, reauthorization, and real-world repair must be present from the architectural stage.

This turns reversibility from a “safety feature” into a condition of civilizational legitimacy: a high-consequence system without genuine correction capacity should not, by virtue of its success, acquire a position from which exit is impossible, even if its current goals are sound and its performance excellent.

7 | Anti-Closure Does Not Mean Refusing to Decide: Protecting the Capacity for Future Choice

1. Anti-Closure Opposes Finality, Not Resolve

At its core, anti-closure means preventing civilization from being permanently locked into a single judgment, system, or optimal solution. It guards against institutional, metric, narrative, and single-model lock-in as well as technological lock-in.

This principle is easily mistaken for “keeping every option open forever.” Yet an open future does not require civilization to preserve indefinitely real-world courses of action that, on current evidence, have already reached the threshold for exclusion on grounds of irreversibility. Nor does it require endless delay before every judgment. Anti-closure does not oppose making decisions; it opposes turning a provisional decision into a final settlement that cannot be questioned, replaced, or subjected to fresh authorization.

2. A Larger Space of Options Is Not Always Better

If an open future is understood as maximizing the number of options, the result is absurd: even courses of action known to systematically destroy basic rights would have to be retained for the sake of “diversity of choice.” What truly matters is not how many old paths are preserved, but whether civilization can still generate, recognize, and turn toward better new ones.

Anti-closure therefore protects “the capacity to continue arriving at sound choices,” not a warehouse of old options. A civilization can close doors that, on current evidence, have reached the threshold for exclusion on grounds of irreversibility, while retaining the capacity for renewed research, fresh judgment, and revision of boundaries when new evidence emerges.

3. The Most Dangerous Forms of Closure Often Arrive in the Name of Success

Failure forces people to seek alternatives; success can all too easily become a reason to make things permanent. A metric that once proved useful may be expanded into an overall score for civilization. A power granted during a crisis may become an ordinary, ongoing power. A model with a long record of excellent performance may gradually become an unacknowledged arbiter of fact.

Precisely because success is so easily entrenched, metrics must be confined to observation, comparison, early warning, organizing evidence, or serving as model inputs. They must not automatically expand permissions, diminish rights, change governance sovereignty, allocate resources, or generate an overall score for civilization.

What anti-closure must truly guard against is a seemingly reasonable claim: “Since it has always worked, there is no need to ask again.” A system calls for the greatest vigilance not

only when it fails, but also when success has become habitual and habit begins to replace fresh judgment. Research on AI governance institutions likewise warns that early institutional and organizational choices create path dependence and may lock in future forms of governance.⁸

4. An Open Future Is Compatible with Boundaries Against Irreversible Harm

An open future does not require civilization to remain neutral toward every direction. Boundaries concerning continued existence, basic rights, and irreversible harm are intended precisely to ensure that there are still people able to choose in the future. Certain prohibitions grounded in irreversibility protect a greater openness of that future.

Three layers must be distinguished here. At the object layer, what is closed is permission to carry out a specific action in the real world: when, on current evidence, it has crossed the threshold for exclusion on grounds of irreversibility, and genuine means of stopping, isolating, repairing, and compensating for its effects do not yet exist, civilization cannot make the real world bear the cost of validation. The rule layer must not be closed as a result: the grounds for exclusion, risk thresholds, scope of application, burden of proof, triggers for review, and criteria for lifting the prohibition must be public and open to challenge, audit, and reconsideration. Still less may the knowledge layer be closed: research, causal exploration, simulation, counterexamples, new evidence, and alternative designs should continue to advance. Knowledge simply cannot be gained by allowing catastrophe to happen.

Closing a path of real-world action, therefore, does not close the paths to understanding it. Preserving the conditions for fresh judgment likewise does not preserve permission to implement it immediately. New evidence that merely adds controversy or uncertainty does not automatically lift a prohibition. Reclassification is warranted only when the evidence is sufficient to change the assessment of consequences, or when a new design makes consequences that were previously unstoppable and irreparable genuinely stoppable, containable, and repairable. The more severe and irreversible the consequences, the higher the burden of proof on those seeking to proceed in the real world.

Anti-closure prevents exclusion rules from becoming unquestionable power; exclusion on grounds of irreversibility prevents “open to review” from deteriorating into “act first, discuss later.” Irreversibility determines whether real-world trial and error is permissible. Correction capacity determines whether the outcome can genuinely be changed after an error occurs. Anti-closure ensures that current judgments and the rules governing them remain open to reexamination, rather than losing that possibility through success, power, or inertia. Together, the three protect not every old path, but a future in which people remain able, and have the conditions, to judge anew.

8 | What Should Be Governed? Purposes, Permissions, Conduct, Responsibility, and Consequences

1. “Regulating AI” Is Too Broad a Formulation

If we could ask only one question about governance, “How should AI be regulated?” would still be too broad. AI can refer to models, products, agents, infrastructure, decision-making tools within organizations, extensions of individuals, interfaces with public systems, and even artificial subjects that may emerge in the future. Treating all of these as a single object of regulation mixes capability, deployment, power, and consequences together once again.

The objects of governance must first be distinguished: capability does not automatically confer power; prediction does not automatically constitute a decision; an optimal outcome for the whole does not automatically create obligations for individuals; and a need for safety does not automatically authorize unlimited surveillance or permanent emergency powers. Every system role must allow for exit, replacement, downgrading, and shutdown.

2. Begin by Answering Five Distinct Questions

Governance should answer at least five questions in sequence: Why govern? What kind of capability, conduct, deployment, or consequence is being governed? Who holds legitimate authority? Who bears responsibility? And when should governance itself end, contract, or undergo reauthorization?

Framed this way, many vague disputes become concrete. Whether highly capable models need evaluation is a question of capability risk. Whether a hospital may let a model change treatment automatically is a question of professional responsibility and authorization. Whether a government may use models to influence citizens' choices at scale is a question of public power. Whether a platform may retain all data about a person's personality is a question of data and control relationships. All involve AI, but they are not the same governance problem.

3. Capability Itself Can Still Be a Direct Object of Governance

Reframing the objects of governance must not lead to the opposite extreme. If certain capabilities can reliably produce serious risks even before a specific deployment purpose is at issue, such as dangerous capabilities that spread extremely easily or uncontrollable chains of autonomous activity, then the capabilities themselves can certainly be subject to direct assessment and restriction.

The distinction is that we need to explain “why this capability itself poses a risk,” rather than assume it should be restricted simply because it is more powerful. Governance should rest on specific causal mechanisms, not an abstract fear of strength.

4. A System's Name Must Not Obscure the Chain of Responsibility

High-consequence public tasks must specify who proposes the goals, who holds authorization, who provides professional judgment, who executes, who has the right to stop and roll back, and who is accountable for the outcome. Any tool for directional judgment may change what can be seen before a decision is made, but it cannot allow decision-makers to hide behind the system.

This is especially important in the AI age. Complex systems can easily create an illusion of responsibility: the model recommends, the organization approves, the algorithm executes, and the platform distributes, yet in the end no actor acknowledges having decided anything. The more complex the technical chain, the more clearly the chain of responsibility must be made visible.

A deeper task of AI governance is not to find a separate regulator for every risk, but to restore traceable relationships among purposes, permissions, conduct, responsibility, and consequences. The OECD explicitly locates accountability within the AI value chain and states that those deploying AI should bear responsibility for the system's appropriate operation.⁹

5. Who Regulates, and Why, Must Also Be Subject to Oversight

Regulation is not inherently right. Any governance undertaken in the name of safety, the public interest, or civilizational risk must account for its own source of authorization, evidentiary standards, affected parties, appeals, duration, and conditions for exit. Research on human oversight of AI likewise observes that overseers may lack the necessary capabilities, face misguided incentives, or make mistakes themselves. Oversight must therefore rely on institutional design rather than the assumption that “humans are inherently reliable.” Otherwise, “regulating AI” could readily become another form of irreversible expansion of power.¹⁰

This completes the loop with anti-closure and correction capacity: not only AI systems, but also the institutions that regulate AI, must be open to correction.

9 | Why Many AI Governance Problems First Bring Humanity Itself into View

1. To Understand What AI Amplifies, First Look at What It Connects To

Major consequences of AI cannot be explained as a single technical problem. Observable causal chains and feedback loops form among technology, capital, competition between states, organizational incentives, political power, public opinion, the judgments of key individuals, and group psychology. The real difficulties arise when information is held only by particular parties, causal connections are severed, interests are concealed, organizations filter out bad news, and key individuals cross established boundaries under extreme pressure.

When AI enters these structures, it does not create every direction from nothing. It often first accelerates directions already in place: a company's pursuit of short-term metrics can be automated; a state's security dilemma can acquire a more powerful technological extension; an organization's rejection of bad news can be further sealed off by more efficient information systems.

At this stage, therefore, many AI governance problems first bring into view humanity's own structures of purpose, competition, power, incentives, and responsibility. Powerful intelligence accelerates and amplifies them, while making them harder to keep hidden. If all responsibility is shifted onto AI, the upstream causes will simply remain where they are.

2. Structure Before Personality

When analyzing upstream structures, the first question should be which real-world conditions are pushing ordinary actors toward dangerous or constructive choices, rather than beginning with personality scores, judgments of loyalty, or psychological diagnoses. This is what “structure before personality” means.

If an organization rewards only speed and not the reporting of bad news, if the structure of competition between states makes any slowdown seem like a strategic defeat, or if a platform's business model rewards polarization, then asking for “kinder AI” can address only a small part of the problem at its furthest downstream end. Truly effective governance must also change the upstream structures that continually generate misguided goals.

3. Theory and Tools Cannot Substitute for Real-World Mechanisms

Changing upstream structures does not mean waiting for competing actors to become kinder first, or having some system design answers on the real world's behalf. Individuals and groups, businesses and civil society organizations, localities and states, alliances and regional communities, international organizations and transnational research networks

have long been seeking ways forward through diplomacy, law, markets, research collaboration, crisis management, local experiments, and public action. What theory can do is clarify the causal relationships these efforts all confront. Whether a mechanism works can be tested only against real-world constraints and consequences.

These constraints are hard ones. States do not automatically escape security dilemmas because cooperation is worth pursuing, nor do businesses automatically abandon short-term competition because the public interest matters. Those controlling models, computing power, data, and infrastructure may also turn governance rules into extensions of their existing advantages. A more realistic direction is to change the benefits and costs actors face, their information conditions, the costs of misjudgment, their shared exposure to risk, and the patterns of their long-term interactions: make concealment, shifting risk onto others, and unbounded escalation harder to profit from, while making verification, communication, risk containment, conditional cooperation, and long-term feedback more consistent with each party's own interests.

4. Upstream Structures Also Include Distribution and Capture

Who controls models, computing power, data, and infrastructure; who supplies resources; who benefits first; who bears the costs to employment, privacy, safety, and other aspects of life; and who has the ability to refuse and leave: all of these shape governance outcomes. If overall efficiency rises while the costs are left to those least able to make themselves heard, upstream structures have not been corrected. They have merely been entrenched more efficiently.

Any institution or tool used for correction can itself be captured by power, capital, and organizational inertia. Those acting to change structures cannot stand outside auditing. Affected parties must retain a practical position from which to question, appeal, and exit, and the tools involved must themselves be auditable, replaceable, and stoppable. In the future, even if Human-COS or other causal exploration tools prove effective in real-world trials, they can only help real-world actors see causal relationships, compare paths, and receive feedback. They cannot replace those actors, monopolize direction, or become a new overarching system.

5. This Does Not Eliminate Machine-Side Risks

One excessive inference must be stopped here: “many AI problems are first of all human problems” does not mean “AI itself has no problems.” Complex models may hallucinate, come under attack, develop strategies that are difficult to explain, produce cascading effects across systems, or execute actions autonomously beyond their designers' expectations. More powerful future systems may also introduce new kinds of risk that we cannot adequately characterize today.

Human directional judgment and machine-side operational safety cannot substitute for one another. Machines may still generate independent risks in perception, permissions, execution, upgrading, autonomy, and the inability to stop. External data also show that documented AI incidents continue to increase, transparency among frontier models is declining, safety performance can weaken substantially under adversarial attacks, and trade-offs among goals such as safety, fairness, and privacy remain insufficiently understood.¹¹

The greatest danger is not that one side is inherently worse, but that, without mutual constraints, the capabilities of the two sides amplify each other. Human fear, interests, and misjudgments may give machines a misguided direction, while machines' speed, scale, and capacity for autonomy may rapidly entrench it. Machines' own errors may in turn reshape people's attitudes, politics, and the next round of judgments.

6. Two-Sided Stability Is More Realistic Than Seeking One Side That Is Always Right

“Who is more trustworthy, humans or AI?” is a poorly framed question. Civilization cannot rest its hopes on either side being right forever. A more realistic goal is to prevent, as far as possible, both human errors of direction and machine errors of operation from passing in a single step into irreversible reality.

Two-sided stability requires at least three capacities: on the direction side, helping actors reduce misjudgments about direction; on the operational side, reducing errors in the perception, permissions, and execution of intelligent capabilities; and through real-world feedback, keeping authorized choices visible and open to correction. Genuine civilizational correction still takes place through the institutions, relationships, actions, and real-world feedback of multiple actors.

Whether risk originates in human direction or machine operation, its consequences are ultimately borne by particular subjects, relationships, communities, and future people. As AI becomes increasingly able to perform functions once carried out by humans, a deeper question therefore emerges: what, ultimately, grounds the human place in the world? If human value comes only from “things machines still cannot do,” then every step forward for AI forces humanity one step back. Human meaning cannot be built in the gaps left by technology's limitations.

10 | After Powerful AI, What Remains Irreplaceable About Human Beings?

1. Do Not Ground Human Value in What Machines Cannot Yet Do

Whenever AI gains a new capability, people begin searching again for “things only humans can do.” This defensive strategy will ultimately fail, because the boundaries of technological capability will keep moving. If human value depends on our temporarily irreplaceable performance of some function, that value will keep depreciating as technological benchmarks advance.

Artificial intelligence can expand our capacity to observe, understand, express ourselves, and act. But knowing more does not automatically give it the authority to make value judgments on a subject’s behalf. A person’s body, experiences, relationships, responsibilities, finite time, and consequences they must bear themselves mean that a predictive model cannot simply take the subject’s place.

2. Human Beings Do Not Need to Be Preserved for Their Inefficiency

Powerful AI can help with care, learning, creation, work, decision-making, and sustaining relationships. A truly mature civilization does not need to preserve inefficient labor deliberately to prove that people still matter, nor does it need to reject automation for the sake of subjecthood.

What makes a person irreplaceable is not a monopoly on a function, but a position in existence: a person lives their own life, bears the weight of their own relationships and choices, faces time that cannot be transferred to someone else, and through these experiences generates, claims, questions, and changes meaning.

The civilization of the future must ultimately find its way into a child’s growing up, the love between two people, family, friendship, work, illness, aging, death, and an ordinary person’s morning. Systems must eventually learn to step back; people must still live their own lives.

3. AI Can Extend Our Presence, but Cannot Turn Life into Something Lived by Proxy

Another kind of human–machine relationship is possible: extending people rather than replacing them. As an extension of a real person, acting under their authorization, AI can carry their care, capabilities, judgment, and presence farther.

This shows that “living one’s own life” does not mean doing everything oneself. A person can use machines to see distant places, AI to express complex ideas, and remote systems to care for loved ones, while handing large amounts of repetitive work over to automation.

What matters is not whether a proxy is used, but whether relationships, purposes, and consequences can still be brought back to the subject.

As the capacity to act on our behalf grows, the real concern is not “how much AI has done for me,” but whether a person’s life is gradually being delegated away, until all that remains is consuming experiences optimized by a system.

4. Meaning Is Not the Last Job Left for Human Beings

Meaning is the root, and its sources can differ. AI can serve to understand, protect, and transmit meaning, and to extend human capabilities. But a role is not the same as ontological status, and understanding meaning does not confer ownership of another’s meaning.

Meaning therefore cannot be understood as a “higher-level job” left for humanity once automation is complete. Meaning is not a job; it is a reality continually taking shape among existence, relationships, choices, actions, memory, responsibility, and experience.

Even if new, genuine nonhuman subjects emerge in the future, this principle need not collapse. A new subject can form its own meaning, but that meaning does not thereby nullify existing human meaning, nor does it automatically create the power to dominate other subjects.

5. What Is Truly Scarce Begins to Shift

As more and more cognitive and operational capabilities become abundant, what is scarce in civilization will shift. In the past, scarcity concerned who could calculate, acquire knowledge, or organize resources. In the future, what becomes increasingly scarce may be whether subjects still have a real place, can know how their choices enter the shared world, and can still form relationships, take responsibility, and live their own lives without a system deciding for them.

This also explains why systems stepping back is not a secondary design concern. If a person’s place ultimately depends on continual rewards, reminders, scoring, and nudging from a system, then their subjecthood is still being propped up by that system. A mature civilization must allow ever stronger capabilities to exist while making people ever less dependent on being driven by them.

11 | A Mature Relationship: The Stronger a System Becomes, the Better It Knows When to Step Back

1. Civilizational Maturity Does Not Mean Systems Reaching Ever Deeper into Life

A truly mature system does not reach ever deeper into people's lives. It is reliably present where needed and keeps its distance where it is not. Rather than continually manufacturing engagement, it allows subjects to sustain a new civilizational logic on their own, even as explicit external driving forces diminish.

If a behavior must rely indefinitely on rewards, leaderboards, system reminders, moral exhortation, or algorithmic nudging to continue, it remains primarily externally driven. The endpoint of genuine success is not a system that becomes ever better at driving people, but people who have ever less need to be driven by it.

2. Stepping Back Means Less Intrusion, Not Less Capability

This does not conflict with “powerful AI.” A stronger system can verify permissions more accurately, detect anomalies earlier, and provide help at lower cost in the background, without continually occupying people's attention, choices, and relationships. Greater capability does not require a more conspicuous presence in the foreground.

For systems to step back, rules need not become less complete. Instead, rigorous checks move into the background, and renewed confirmation is limited to points at which the relationship genuinely changes: when purposes change, permissions expand, a substantive boundary between domains is crossed, major risks emerge, a term expires, or the subject actively withdraws authorization. At those points, authorization and judgment must be reopened.

This also answers the earlier question about efficiency: procedural steps and the frequency of interaction can decrease, but the traceability of purposes, completeness of authorization, records of responsibility, feedback on anomalies, and capacity to reopen judgments must not diminish.

3. Low Intrusion Is Not Invisible Control

Moving into the background carries a danger: a system may appear to step back while actually gaining a deeper capacity for invisible intervention. If users cannot see which objectives are being optimized, which data are being used, or which choices are being ranked, less interaction may instead mean greater control. Systems stepping back must therefore be designed to combine minimal disturbance with strong auditability, replaceability, and the ability to exit. “Seamlessness” cannot substitute for transparency.

Systems stepping back must therefore satisfy two conditions at once: less day-to-day disturbance for subjects, with no reduction in the auditability of power, purposes, or anomalies. What truly steps back is the driving force, not responsibility; the frequency of intervention, not boundaries.

4. Stepping Back Is a Reversible Reduction in Intervention, Not a One-Time Withdrawal

What steps back is not civilization's capacity to detect anomalies, trace power, receive appeals, repair harm, and reconsider judgments. It is the system's continual driving of subjects and its unnecessary occupation of their lives. A tool can be withdrawn, an interface can disappear, and a service can be replaced; but the necessary capacities of civilization must continue to be carried by real-world subjects, institutions, or other replaceable means.

Between intensive involvement and self-sustaining operation, there needs to be a reversible process of reducing intervention. Begin with clearly bounded functions and settings: reduce rewards, rankings, reminders, default recommendations, behavioral nudges, and repeated confirmations, while retaining necessary services, anomaly detection, audit records, appeals and remedies, and routes for restoring support. Further reductions in foreground intervention should come only after real-world outcomes have been observed in a manner proportionate to the risks. If harm increases, appeals fail, the capacity to provide alternatives disappears, or subjects' autonomy declines, the judgment must be reopened.

Whether stepping back can continue depends first on these questions: without continual rewards, reminders, and algorithmic nudging, can subjects still understand reasons, choose autonomously, raise objections, handle conflicts, and sustain necessary cooperation? Would stepping back disrupt basic living needs, the protection of rights, or assistance indispensable to certain subjects? Do the system's purposes, data, rankings, and permissions remain auditable? Can services be replaced, and is exit a genuine possibility?

Other questions also matter: can harms, anomalies, and those who have been overlooked still be identified, and can appeals reach those able to act and produce remedies? Are there periodic reviews, as well as triggers for renewed intervention, such as a change in purpose, expanded permissions, a substantive crossing between domains, rising risks, or the expiry of a term? If stepping back happens too quickly, can it be paused or scaled back, can necessary support be restored, or can a different means take over, without automatically reviving old permissions?

The unit of judgment for these questions should not be an abstract "civilization as a whole," but specific functions, settings, and subjects. Different domains, groups, and individuals may be at different stages. For people who need long-term assistance, the continued availability of support as needed does not mean a failure to mature; turning "stepping back" into the withdrawal of necessary help is the failure. Nor can a system decide for itself whether to step back on the basis of its own usage or success rates. That decision should be

made by subjects who hold real-world rights and responsibilities, under conditions that allow those affected to appeal, evidence to be audited, and independent review to take place.

There is currently no unified set of indicators, validated over the long term and across multiple settings, that can declare “civilization has matured.” The questions above do not constitute an overall civilization score, fixed thresholds, or a formula for automatic decisions. They are questions of evidence that must be answered publicly before stepping back and reviewed continually afterward.

5. A Civilizational Logic Truly Takes Shape Only When It Becomes Self-Sustaining

Stability is not yet the same as being self-sustaining. Civilization truly becomes self-sustaining only when, as rewards, rankings, forceful prompts, administrative mobilization, and algorithmic nudging gradually step back, the new logic continues through subjects’ choices, relationships, habits, institutional inertia, and reproduction across generations.

This means that a civilizational system ultimately cannot take its own existence as a measure of success. The most successful tool for judging direction is not one that everyone uses every day, but one through which society becomes increasingly able to foresee consequences, allow dissent, and preserve the capacity for correction. Likewise, the most successful AI support does not fill every part of life with AI. It gives people powerful support when they need it, while allowing them to live fully when they do not.

Strong capabilities, low intrusion, and self-sustaining operation together form a mature human-machine relationship distinct from the idea that “more control means more safety.”

6. Stepping Back Brings This Essay Back to Its Starting Point

This essay began by asking, “Is AI being too powerful the problem?” We can now see a complete reversal: a truly mature civilization does not need to suppress AI’s strength. It needs powerful capabilities to operate within a sound direction, clear authorization, continuous feedback, and genuine correction, and ultimately to prove their value without occupying people’s lives.

12 | Conclusion: As Scarcity Shifts, What Must Civilization Truly Safeguard?

1. Returning to the Opening: Scarcity Has Not Disappeared; It Has Shifted

The opening set out five points of reference. After eleven chapters, we can now see that these are not an arbitrary collection of value terms, but five aspects of the same civilizational bottleneck as intelligent capabilities expand. The stronger the capabilities, the less room there is for ambiguity about direction; the faster the action, the more precious system-wide causal understanding and time for correction become; the more a system can act on our behalf, the more important it is that the place of real subjects not be hollowed out. Since any party can make mistakes, shared correction cannot be entrusted to a center that claims to be invariably right.

Nor are these five elements isolated from one another. Direction determines what capabilities are used for; system-wide causal understanding subjects that direction to the test of reality; time for correction leaves an opportunity to change errors once they are discovered; the place of real subjects answers who is living, who bears the consequences, and whom civilization ultimately serves. The capacity for shared correction prevents the first four from being closed off again by power, inertia, or the system's own success.

The shift in scarcity is not a passage from a “world of scarcity” into “unlimited abundance,” still less a reason to abandon justice in the distribution of intelligence. Frontier models, chips, data centers, and critical infrastructure remain concentrated, and gaps between subjects in access to capabilities, control, explainability, and the capacity to find alternatives may still widen. The shift points to another fact: even if the supply of intelligence continues to grow, more intelligence alone is not enough to produce a better civilization automatically.¹²

2. Direction: The Stronger the Capabilities, the More We Need to Know What They Are For

As “how to do it” becomes easier, “why do it” becomes more important. The stronger the capabilities of a civilization headed in the wrong direction, the more likely it is to reach the wrong outcome faster. Direction is not a grand declaration of values, but a civilizational judgment that can inform concrete goals, the boundaries of rights, authorization, and responsibility.

The value of judging direction does not lie in possessing a final answer about values. It lies in setting out in advance how different choices would change people and the world, and whether a better path remains available.

3. System-Wide Causality: Local Intelligence Cannot Replace Understanding of the Whole

Intelligence can generate answers in abundance. What is truly scarce, however, is the ability to bring scattered facts, different subjects, short-term interests, long-term feedback, and irreversible thresholds into a single causal map that can be challenged. Civilization often fails not because no one is intelligent, but because even when everyone is right locally, the whole still goes wrong.

Causal transparency does not mean enforcing uniform thought. It makes disagreements more precise, allows different subjects to understand exactly at what level they disagree, and gives shared survival and shared consequences a chance to enter judgment before hostility does.

4. Time for Correction: The Faster We Move, the More Precious the Window for Changing Outcomes

One of the truly costly resources of the AI age may be time for correction. The faster capabilities operate, the less time it takes for an error in judgment to become a fact in the world. It is therefore all the more important to preserve a window for pausing, questioning, replacing, and rebuilding before outcomes become irreversible.

This is why reversibility cannot mean merely returning to an earlier version, and anti-closure cannot mean merely keeping options open. Together, they protect civilization's capacity to change outcomes before errors become final.

5. Real Subjects: The Stronger the System, the More Essential It Is That Life Remain Present

When knowledge, expression, creation, action by proxy, and execution can all be expanded on a vast scale, what is truly scarce is not "what work still has to be done by humans," but whether people still have a real place in the world: whether they can live, form relationships, choose, refuse, take responsibility, be needed, create, and know that their actions genuinely reach other lives.

Meaning cannot be delivered centrally by a system, nor can AI produce it on a subject's behalf. AI can extend human capabilities, and new nonhuman subjects may even emerge in the future. But abundance of capability cannot justify civilization reducing living beings to the receiving end of system optimization.

6. Shared Correction: No One Can Be Presumed Infallible

A civilization's deepest security does not come from having an AI that never makes mistakes, or a group of people who never make mistakes. It comes from ensuring that, when any party errs, facts can still reach those with authority, dissent can still emerge, responsibility can be traced, systems can be stopped, permissions can be narrowed, old answers can be replaced, and new paths can be created.

Shared correction cannot be monopolized by a single system. Individuals, groups, organizations, states, communities, professional institutions, and future human-machine collaboration tools can each provide only part of the capacity to observe and act. Truly reliable civilizational correction arises among multiple subjects through visible facts, judgments open to challenge, traceable responsibility, execution that can be stopped, and feedback that can return to the real world. Any tool that positions itself as the sole point of access has already begun to erode what it was meant to protect.

7. AI Governance Is Thus Placed Back within the Question of Civilization

If we must ask which judgment about governance readers should change after reading this essay, the closest answer is not a regulatory rule but a shift in perspective: AI governance at this stage cannot govern AI alone. It must also govern the purposes, permissions, and incentives that humans give AI; the capabilities and operational risks of the machines themselves; the real-world consequences of their coupling; and the limits of power within the governance structure itself.

This does not displace conventional AI safety, model evaluation, privacy, responsibility, regulation, or international rules. It simply places that work back within a longer causal chain.

8. What Must Civilization Truly Become Stronger At?

What is genuinely worth pursuing for the future is not keeping AI forever too weak to threaten humanity, nor keeping a human forever beside a nominal final button. What truly needs strengthening is civilization's capacity to place powerful capabilities within purposes that can be explained, clear rights and responsibilities, visible consequences, and continual correction.

Abundance of intelligence will not automatically bring abundance to civilization. Answers can be generated in vast numbers, but real subjects must still take responsibility for direction; actions can be completed in an instant, but their consequences will still fall on particular lives; systems can become ever stronger, but civilization must still leave itself time to stop, question, and choose again.

When intelligence is no longer scarce, civilization's test truly begins: are we better able to endure? Are living beings better able to live authentically, form relationships, create, and generate meaning? And after acquiring unprecedented capabilities, have we retained the power to judge, correct ourselves, and become ourselves anew? Only then will intelligence become more than a stronger force and begin to become a better capacity of civilization.

Note | Limits of the Evidence, Limits of the Mechanisms, and Open Questions

- External literature in this essay supports only specific facts, methods, or institutional experience. It does not establish the overall validity of Civilization Leap theory.
- The three goals; meaning as the root; the view that human irreplaceability is a position in existence; the extension of corrective capacity to the scale of civilization; the affirmative definition of anti-closure; and the five categories of civilizational scarcity remain this essay's normative claims and theoretical synthesis.
- Whether future AI will develop subjective experience, nontransferable continuity as a subject, and the corresponding standing remains an open question. Present-day institutional human responsibility must not be substituted for a permanent ontological conclusion.
- When the three goals conflict, this essay provides only minimum requirements for analysis, including the boundaries of rights, the distribution of benefits and harms, reversibility and repair, alternative paths, authorization, and appeals. It offers no formula for automatically balancing them. Relevant cases and mechanisms require separate study.
- Chapter 9's discussion of the structure of benefits and costs, information conditions, risk shifting, the distribution of capabilities, and capture by power proposes only causal directions and minimum constraints; it does not constitute a unified mechanism design. The existing efforts of real-world subjects form the basis on which mechanisms develop. Bilateral research prototypes, including "A Computational Research Design for China–U.S. Competition and Cooperation," do not serve as proof of this essay's claims.
- Once Human-COS has been designed and enters trial operation, it can serve as one tool for causal exploration involving multiple subjects, counterfactual comparison of mechanisms, identification of critical thresholds, and real-world feedback. Other AI tools, existing institutional mechanisms, professional methods, and the judgment and action of real-world subjects themselves are also part of mechanism research and practice. This essay does not claim in advance practical gains from the tool that have yet to be validated, nor does it establish the tool as the only instrument or the only direction.
- What this essay proposes for systems stepping back is a normative framework of criteria, not a system of indicators validated through trials across different settings. It specifies no unified overall score, fixed thresholds, or automatic decision to step back. The criteria still require real-world validation for specific functions, subjects, and levels of risk.

Further Reading and Paths into Collaborative Research

If this essay has led you beyond asking “How powerful is AI?” to ask “Where is civilization preparing to take this capability?”, you can continue reading along the following paths.

Civilization Leap: The Path of Civilization in the AGI Era (All Five Volumes) | Explore the full overarching framework of civilizational goals, institutions, engineering, and structures of meaning.

[*English overview of the five-volume work*](#)

Living Civilization: The Trilogy, Combined Edition | Return from institutions and systems to the real lives of ordinary people, and explore how a future civilization can enter everyday life.

[*English overview of the trilogy*](#)

Overview of the Complete AGI-COS Series | Explore the different tracks of machine-side perception, permissions, execution, auditing, rollback, and long-term safety.

[*English version-family overview*](#)

Complete Public Library | Browse the currently public Civilization Leap writings.

[The supplementary English reading index is selective; it does not establish that every work has a complete English edition.](#)

Public Reading Gateway | Enter the Civilization Leap public reading paths through aspirations or crises.

[*Civilization Leap: Public Reading and Communication Entry*](#)

Official website: www.civitas.top

Endnotes

1 Stanford Institute for Human-Centered AI, Artificial Intelligence Index Report 2026 (Stanford University, 2026), report overview and Chapter 1, “Research and Development,” https://hai.stanford.edu/assets/files/ai_index_report_2026.pdf. The report documents continued capability gains and broad diffusion of AI, while also showing substantial concentration in frontier-model production and compute infrastructure.

2 UNESCO, Recommendation on the Ethics of Artificial Intelligence, adopted 2021 and published 2022, <https://unesdoc.unesco.org/ark:/48223/pf0000381137>. UNESCO emphasizes human dignity and rights, auditability, traceability, responsibility and accountability, and human oversight, including that AI systems should not displace ultimate human responsibility and accountability. This supports current institutional responsibility, not a permanent ontological judgment about future AI subjecthood.

3 Liming Zhu, Qinghua Lu, Ming Ding, Sung Une Lee, and Chen Wang, “Designing meaningful human oversight in AI,” *AI and Ethics* 6, art. 286 (2026), <https://doi.org/10.1007/s43681-026-01147-7>; Ben Green, “The flaws of policies requiring human oversight of government algorithms,” *Computer Law & Security Review* 45 (2022): 105681, <https://doi.org/10.1016/j.clsr.2022.105681>.

4 Robert J. Lempert, “Robust Decision Making (RDM),” in *Decision Making under Deep Uncertainty: From Theory to Practice* (Springer, 2019), 23-51, https://doi.org/10.1007/978-3-030-05252-2_2.

5 Marjolijn Haasnoot, Jan H. Kwakkel, Warren E. Walker, and Judith ter Maat, “Dynamic adaptive policy pathways: A method for crafting robust decisions for a deeply uncertain world,” *Global Environmental Change* 23, no. 2 (2013): 485-498, <https://doi.org/10.1016/j.gloenvcha.2012.12.006>.

6 United Nations Framework Convention on Climate Change (1992), Article 3(3): where there are threats of serious or irreversible damage, lack of full scientific certainty should not be used as a reason for postponing precautionary measures, while policies and measures should be cost-effective, <https://unfccc.int/resource/ccsites/zimbab/conven/text/art03.htm>. This is used as a methodological analogy, not as direct validation of the framework proposed here.

7 Elham Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1 (2023), <https://doi.org/10.6028/NIST.AI.100-1>. GOVERN 1.5 and 1.7 address ongoing monitoring, periodic review, clear roles and responsibilities, and safe decommissioning or phasing out; GOVERN 1.6 addresses system inventory.

8 Charlotte Stix, “Foundations for the future: institution building for the purpose of artificial intelligence governance,” *AI and Ethics* 2 (2022): 463-476, published online 2021, <https://doi.org/10.1007/s43681-021-00093-w>.

9 OECD, “AI risks and incidents,” accessed August 30, 2026, <https://www.oecd.org/en/topics/sub-issues/ai-risks-and-incidents.html>. The OECD states that AI risks need to be managed throughout the value chain and that those who deploy AI should be accountable for proper system functioning.

10 Johann Laux, “Institutionalised distrust and human oversight of artificial intelligence: towards a democratic design of AI governance under the European Union AI Act,” *AI & Society* 39 (2024): 2853-2866, <https://doi.org/10.1007/s00146-023-01777-z>.

11 Stanford Institute for Human-Centered AI, Artificial Intelligence Index Report 2026, Chapter 3, “Responsible AI,” https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_3_responsible_ai.pdf. The chapter reports rising documented AI incidents, declining frontier-model transparency, weakened safety performance under adversarial attack, and trade-offs among safety, fairness, and privacy objectives; incident databases remain reporting-dependent.

12 Stanford Institute for Human-Centered AI, Artificial Intelligence Index Report 2026, Chapter 1, “Research and Development,” https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_1_research_development.pdf. Industry produced over 90% of notable AI models in 2025; global AI compute capacity grew rapidly, while leading model production and core compute infrastructure remained concentrated.

Appendix | Personalized Further Reading Guide

Translation of the navigation appendix added to the online PDF as page 43. This appendix is separate from the 42-page Chinese publication edition.

P03-AISCARCE-002 | V1.1.1 | Public Ideas Edition

This is a supplementary library navigation page. It does not alter the main text of When Intelligence Is No Longer Scarce V1.1.1.

The Next Question | Humanity’s Civilizational Choices in the AGI Era

Why read it: This essay first brings the “shift in civilizational scarcities” into view. The next question is what, as capabilities grow ever stronger, humanity intends future intelligence to learn, and where it intends to take civilization.

Real-World Judgment | Human-COS: Seeing Earlier What Would Otherwise Be Seen Only Later

Why read it: It brings direction, system-wide causality, and time for correction into high-consequence judgments: how consequences can become visible earlier, and how responsibility can enter real-world judgment before irreversibility.

Operational Safety | The Controlled AGI-COS Version Family

Why read it: When a problem concerns machine-side perception, crises, coordination, or long-term stability, use the different specialist AGI-COS tracks appropriate to the problem and stage, rather than searching for one supreme version.

Theoretical Foundations | The Five Volumes of Civilization Leap

Why read it: Continue exploring civilization’s purpose, the operating logic of institutions, engineering pathways, civilizational perpetuity, and the foundations of meaning. This essay is not a summary of the five volumes, but an independent entry point through the “shift in scarcity.”

Real Life | The Living Civilization Trilogy

Why read it: Test these abstract judgments about direction, subjects, error correction, and meaning by entering an ordinary person’s day, relationships, work, choices, and future life.

A Concise Review | The Public Guide and Eight-Question Reading Map for When Intelligence Is No Longer Scarce

Boundary: Growth in capability, equitable access to capability, who controls it, and what purpose it serves are four different questions. “When intelligence is no longer scarce” is a directional thought experiment, not a statistical conclusion about the world today.

English reading index: <https://www.civitas.top/en/library/> (selective existing English material; not the complete Chinese library in translation).

English links lead to corresponding webpages. Some destinations are public introductions or version-family overviews rather than complete English books. Unlinked source references do not establish the availability of a full English text.