

When Intelligence Ceases to Be Scarce

Humanity, AI, and the Direction of Civilization

Public Reading Edition · V1.0

Author: Zijun Fu

Produced by the Civilizational Transition Research Group

Public Reading Edition · V1.0 · August 2026

Publication Information

Author: Zijun Fu

Produced by: Civilizational Transition Research Group

Edition: Public Reading Edition V1.0

Publication date: August 2026

Official website: <https://www.civitas.top/en/>

Conceptual baseline: Chinese Public Longform / Independent Publication Edition V1.0

How to Read This Edition

This public reading edition preserves the full argument of the independent English edition while adding brief examples, operational bridges, and smoother transitions where a general reader might otherwise face an abrupt change of level. It does not simplify the theory by removing its boundaries. The Chinese Public Longform / Independent Publication Edition V1.0 remains the sole conceptual baseline.

A Few Terms Before You Begin

Civilizational Transition — Civilizational Transition. Not a one-time passage from a fixed state A to a fixed state B, but civilization's continuing capacity to revise its operating logic as capability, knowledge, relationships, risks, and forms of subjecthood change.

Continued existence — continued existence / civilizational survival. The floor condition for civilization; distinct from the first civilizational aim.

The three aims — freedom from survival insecurity → a happy life → a more meaningful life. The aims remain distinct and are not collapsed into one utility score.

Human-COS — Human-COS. One possible future augmentative tool for human judgment and direction; not the only tool and not a sovereign mechanism.

AGI-COS — AGI-COS. Machine-side perception, authorization, execution, audit, interruption, rollback, and safety architecture.

Reversibility — reversibility / correction capacity. Rollback is one technique; correction capacity is the broader civilizational capability.

Anti-Closure — Anti-Closure. Do not let a provisional answer become permanently immune to re-examination.

Meaning as foundational — meaning as foundational. Meaning may have many sources; no system, state, or theory owns its definition.

Contents

Before the Argument Why This Is Not an Eight-Question AI Governance Essay.....	5
0 Introduction: What Changes When Intelligent Capability Keeps Growing?.....	6
1 Why Civilization Needs Intelligence - and Why Intelligence Cannot Define Civilization.....	9
2 AI Is Not an External Species: Reframing the Human-AI System.....	12
3 One of the Most Important Uses of Stronger Intelligence: Seeing Consequences Before They Arrive.....	16
4 From Judgment to Action: Seek the Best Available, Not Static Perfection.....	20
5 Two Kinds of Error: Reversible Experimentation and Irreversible No-Go Zones.....	23
6 Reversibility Is Not a Return to the Past: Correction, Upgrade, Reconstruction, and Reauthorization.....	26
7 Anti-Closure Is Not Indecision: Preserving the Capacity to Produce Better Choices.....	29
8 What Exactly Are We Governing? Purpose, Authority, Conduct, Responsibility, and Consequence.....	31
9 Why Many AI-Governance Problems First Reveal Human Structures.....	34
10 After Strong AI, What Remains Irreplaceable About Human Beings?.....	38
11 A Mature Relationship: The Stronger the Capability, the Better the System Knows When to Recede.....	41
12 Conclusion: When Intelligence Ceases to Be Scarce, What Becomes Truly Scarce?.....	44
Notes on Evidence, Mechanisms, and Open Questions.....	47
Further Reading and Research Entry.....	49

Before the Argument | Why This Is Not an Eight-Question AI Governance Essay

This essay did not begin with eight questions in search of eight answers. The questions were a mirror. They revealed that the Civilizational Transition framework already contained answers about people, AI, institutions, risk, and the future across different texts and system layers, but had not yet assembled them into one continuous causal argument.

The framework already separates capability, experience, continuity of subjecthood, protection, legal standing, political membership, public authority, and sovereignty. It also separates the human-side problem of judgment and direction from the machine-side problem of safe perception, authorization, execution, and interruption. No system acquires civilizational sovereignty merely because it becomes exceptionally good at one specialized task.

The task here is not to add another rulebook for keeping AI under control. It is to move the question upstream. As intelligent capability grows, humanity has to become clearer about why it is using that capability, where it intends to take it, how consequences can be seen earlier, and how greater capability can remain subordinate to the purposes of civilization rather than becoming a purpose in itself.

0 | Introduction: What Changes When Intelligent Capability Keeps Growing?

1. The Question Is Misframed

“What if AI becomes too powerful?” sounds natural, but it treats capability growth itself as the problem before the causal structure has been examined. Governance then easily collapses into a one-way intuition: the stronger the intelligence, the tighter the restraint; the more the system can do, the more humanity must guard the final switch. If capability growth is an important technological trajectory, however, the decisive issue is not strength alone. It is where that strength is directed, by whom, and for what purpose.

The Civilizational Transition framework does not allow capability to leap automatically into standing or authority. Capability may be combined; authority may not simply be added. Facts and recommendations do not become decisions by themselves.

Two propositions can therefore be true at once. More capable AI can create new technical, deployment, complexity, and autonomous-action risks. And greater capability, by itself, does not prove that civilization is moving in the wrong direction. If these propositions are collapsed into one, governance remains trapped at the level of preventing capability from becoming catastrophe and never reaches the upstream question: can the same capability help us detect catastrophe earlier, understand complex systems more fully, compare paths at lower cost, and change course before consequences become irreversible?

What civilization must watch is the chain connecting capability to purpose, use, authorization, responsibility, feedback, and consequence. Capability can amplify destruction, but it can also make the formation of destruction visible earlier. It can reinforce closure, but it can also cut across fragmented information and reveal where a local action returns through a larger causal network. Power is not the whole problem. Direction gives power its consequences.

2. AI does not merely create new risks; it removes old buffers

Advanced AI is binding together problems that institutions once handled separately - across disciplines, organizations, and time horizons. Frontier capability continues to rise as AI enters research, markets, medicine, education, public administration, and everyday life. Yet frontier models and compute infrastructure remain highly concentrated.¹

The deeper break is conceptual as well as operational. In human beings, intelligence, consciousness, subjecthood, life, responsibility, rights, and the capacity to bear meaning have historically overlapped. Advanced AI makes it possible for these capacities, experiences, and institutional positions to come apart. Once that natural overlap disappears, every inference across levels has to earn its own justification.

AI may therefore compress not only computation time but the time available for civilizational correction. A bad objective can be executed faster. A partial metric can be embedded more deeply. Fear, rivalry, and short-term organizational incentives can acquire technological reach. Yet the same intelligence may allow us, at unprecedented scale, to compare actors, time horizons, domains, and consequences that were previously kept apart.

AI is best understood here as a new condition of capability. It lets a sound direction travel farther, and a bad direction reach reality faster. The crucial question is whether capability growth increases civilization's power to see and correct, rather than merely its power to execute.

3. From control to direction

Human-COS marks one side of this distinction. It is not designed to make final choices for humanity. Its role is to make actors, facts, interests, human responses, thresholds, alternatives, unknowns, and the strongest objections more visible before a decision is made, thereby changing what responsible actors can know and see before they decide. AGI-COS addresses the other side: whether an authorized action can be carried out safely, whether authority has been exceeded, and whether the system can still be stopped, taken over, audited, or rolled back.

A complete account of the human-AI problem at the civilizational level therefore has at least two sides. One is direction: why act, how will the action change people and the world, and is there a better path? The other is operation: can the action be executed safely, is it within authority, and can it still be stopped? A system can execute the wrong objective perfectly. A noble objective can also be pursued through an unsafe technical path.

Once these sides are put together, the first governance question changes. It is no longer simply how to ensure that AI always obeys humans. It becomes: how can a human-AI system remain correctable across direction, judgment, execution, and feedback? Human beings are not important because they occupy the final click in a workflow. AI is not legitimate merely because it is confined to the role of an obedient tool. The relationship must ultimately answer to life, responsibility, and real consequences.

4. The first claim

The claim can be compressed into one sentence: the continued growth of AI capability is not, by itself, a complete governance problem. Civilizational outcomes depend on how stronger capability combines with human purposes, institutions, power, responsibility, cognition, and feedback.

This claim remains open at both ends. It does not deny technical risk, and it does not treat capability as progress by definition. If new forms of machine subjecthood, autonomy, or loss of control emerge, they require independent investigation. Human warfare, power competition, institutional incentives, short-termism, and confusion about direction cannot simply be relabeled as AI's responsibility either.

1 | Why Civilization Needs Intelligence - and Why Intelligence Cannot Define Civilization

1. Ask first why civilization exists

Any institutional design for AI that does not first ask why civilization exists risks promoting means into ends. A system can become faster, more accurate, and more stable. Production, knowledge, and coordination can expand. But if it remains unclear whom these capacities ultimately serve, technical maturity can hide the absence of purpose behind the appearance of efficiency.

The Civilizational Transition framework gives a direct answer: civilization does not exist for an abstract system, nor for an “overall good” detached from concrete lives. It exists in real people, relationships, communities, and those who come after us.

This is the role of the three civilizational aims. Civilization should first free people from basic deprivation that destroys dignity and meaningful choice; it should enable lives that are genuinely rich in freedom, relationship, enjoyment, and human flourishing; and, beyond survival and comfort, it should leave room for lives to generate meaning through relationship, creation, responsibility, exploration, and existence itself. In compressed form: freedom from survival insecurity → a happy life → a more meaningful life.

Continued existence is the floor beneath these aims. It is not identical to the first aim. A civilization that does not continue cannot sustain any value, but mere continuation is not yet the life civilization exists to make possible.

2. The three aims are not three scores that can be traded into one another

The three aims do not form a single utility function. A higher probability of aggregate survival does not automatically cancel persistent violations of basic rights. A rise in average well-being does not make freedom of choice an expendable cost. Nor can “greater meaning” justify imposing deprivation, suffering, or sacrifice on particular people.

This matters especially in systems built around powerful optimization. The stronger the optimizer, the more dangerous it becomes to compress

civilizational aims that cannot be automatically traded against one another into one score before optimization begins. The central discipline is not to teach AI a longer list of value words. It is to avoid building institutions that pre-convert things that should not have an automatic exchange rate.

Greater computational power should improve our ability to compare consequences. It should not silently create a universal exchange rate among survival, rights, flourishing, and meaning.

3. Meaning is foundational; that does not give a system the right to define it

Meaning is foundational, but its sources may differ. It may arise through relationship, love and being loved, creation, care, responsibility, exploration, memory, shared life, or the irreducible experience of being a particular person. Meaning is not another name for satisfaction, engagement, social impact, contribution points, or subjective pleasure.

Foundational does not mean highest-ranking, most urgent, or universally preferred. Meaning is not another item to be placed at the top of a value hierarchy, nor an objective to be maximized after survival and happiness have been secured. It is more basic: values, purposes, choices, and institutions matter only because there are subjects for whom something can matter. Meaning is therefore not the final variable in an optimization function. It is part of the condition under which any function, goal, or achievement can count as worth pursuing at all.

Under current conditions, advanced AI may play roles in understanding, protecting, conveying, and extending human meaning only within deployments that preserve traceable responsibility, explicit authorization, auditability, correction, and exit. A functional role does not settle what a being is, and it is not public authority.

This does not force AI back into the category of a trivial tool. AI may become deeply involved in memory, expression, care, creativity, and the extension of human capability. But understanding meaning does not amount to owning it, and helping a person articulate value does not entitle a system to decide what makes that person's life worth living.

This is also why the human role cannot be reduced to a ceremonial human in the loop. The purpose of civilization, the experience of living, the bearing of meaning, and the consequences of action are not a final approval click. As long as civilization is constituted by real subjects who live, suffer, relate, choose, and bear consequences, the human being is not merely a safety switch inside the loop. The human being is one reason the loop exists.

4. Intelligence is a civilizational capability, not civilizational sovereignty

Technical capability, price, contribution, social recognition, cultural representation, or any other identity marker does not automatically generate public authority or the right to govern meaning. Evidence for future artificial subjecthood must remain open to revision, but new evidence would not automatically settle legal standing, political office, resource authority, or sovereignty.

At the level of current institutional governance, UNESCO's Recommendation on the Ethics of Artificial Intelligence likewise emphasizes human dignity, auditability, traceability, accountability, and human oversight, including the principle that AI systems should not displace ultimate human responsibility and accountability. That supports a present institutional allocation of responsibility; it does not settle the future ontology of artificial subjecthood.²

That allows civilization to be ambitious about intelligent capability without making political authority rise in lockstep with capability. A system may know more, model farther, and help more deeply while still acting within explicit authorization and responsibility.

The more mature question is not whether civilization becomes advanced by possessing stronger AI. It is whether civilization can make increasingly powerful intelligence serve life without allowing capability to redefine what life is for.

2 | AI Is Not an External Species: Reframing the Human-AI System

1. Move beyond the picture of “an AI system”

The Civilizational Transition architecture has never assumed that one unified system should run civilization. Human-COS addresses systemic failures in human judgment, collective behavior, and civilizational direction. AGI-COS addresses machine-side perception, authorization, execution, upgrading, autonomy, and risks arising when systems cannot be reliably stopped. Nebula is a proposed environment in which people, AGI, institutions, resources, relationships, public space, private domains, and real-world action interact. Spark works at the scale of persons and relationships, treating AI as an authorized extension of a real human being’s capacity.

The structure that emerges is clear: this is not a model in which one AI system decides and humans approve at key checkpoints. It is a socio-technical system of multiple actors, multiple systems, layered authorization, distributed responsibility, and real-world feedback. AI is one increasingly powerful class of capability within that system.

Human in the loop may be necessary in some engineering workflows, but it should not be elevated into the final civilizational picture. Human beings do not matter because they occupy a loop position. AI does not become legitimate merely because it has been placed inside one. The more basic unit is a real subject, a shared purpose, an authorized capability, a responsibility structure, and feedback from reality.

2. Human-COS addresses direction; AGI-COS addresses operation; reality bears the result

The cross-phase baseline of Human-COS draws the distinction clearly. Human-COS asks: Why are we doing this? How will it change people and the world? Is there a better path? AGI-COS asks: Can the authorized action be carried out safely? Is authority being exceeded? Can the process still be stopped, taken over, audited, or rolled back? The former does not acquire civilizational sovereignty, final authority over value, or automatic

execution power. The latter does not acquire the right to set policy goals, govern society, or turn temporary safety powers into permanent control.

Nebula does not possess a third abstract sovereignty over judgment. Its role is to let a judged, decided, and authorized choice enter reality, and then return action, benefit, harm, lived experience, institutional friction, appeals, failures, repair, and new facts to the next round of judgment. Reality is not the endpoint of system output. It is the next input.

Human-AI coexistence should therefore not be understood as encoding human values into one model and then allowing that model to run civilization. It is better understood as a two-sided stability problem. Human beings can make directional errors. Machines can make perceptual, inferential, or execution errors. Stability does not require either side to be permanently correct. It requires that deviation can still be seen, reconsidered, reauthorized, stopped, and redirected before it becomes irreversible reality.

A shared system cannot eliminate error. What it can do is prevent an error on either side from passing straight through into irreversible consequence. Human fear, interest, or misjudgment should not become fact merely because it has acquired AI speed. A machine's high-confidence output should not bypass purpose, authorization, and responsibility merely because it is more efficient.

3. From transparent multi-party validation to lower-friction mature operation

When a new system is first established, the links among judgment, authority, and consequence have not yet been sufficiently tested. Transparency, participation by multiple actors, expert review, strong counterarguments, independent replication, and responsibility records therefore matter more. Human-COS explicitly requires multi-actor, multi-path, and multi-horizon simulation, together with the strongest counterexamples and alternative explanations. AGI-COS requires sources, confidence, unknown regions, and authorization boundaries to remain visible.

But maturity does not require the same number of human checkpoints and approval layers forever. Research on meaningful human oversight reaches

a similar conclusion: oversight can be distributed across design, operation, and accountability rather than reduced to real-time human intervention at every step. At the same time, nominal human-in-the-loop arrangements can deteriorate into rubber-stamping and create a false sense of security.³

Reducing steps is not reducing responsibility. Procedural density can fall, while authorization integrity, feedback, accountability, and the capacity to reopen judgment must remain. Efficiency and legitimacy are not necessarily enemies. A mature system does not reopen every civilizational principle before every routine action. It moves validated rules into the background and reserves scarce human attention for genuine change: anomalies, altered purposes, value conflict, irreversible risk, and changes in authority.

4. Spark as a concrete middle-scale example

Spark expresses its core principle simply: do not replace the person; amplify the person. AI functions as an authorized extension that helps a real human being carry care, goodwill, capacity, and judgment across time, distance, and form toward other real people. The subject is not an AI persona manufactured by the system. It is the person who authorizes the extension, bears the relationship, and receives the consequences.

Spark matters here not because it represents every possible human-AI relationship, but because it makes an abstract principle concrete. AI can greatly extend a person's capacity to be present without taking subjecthood away from that person. A system can help a person do more without turning "more" into the reason for the system's own existence.

This suggests a broader direction for human-AI civilization. Stronger AI does not have to prove its value by becoming more human-like or by occupying more of the positions previously held by human beings. Another path is possible: the stronger AI becomes, the farther it can extend human capacity, care, understanding, and action while real subjects remain themselves.

5. Future subjecthood remains open

Treating AI, under current conditions, as an authorized layer of civilizational capability does not permanently deny future artificial

subjecthood. The framework applies the same standard of evidence in both directions: current public evidence is insufficient to establish subjective experience, non-transferable continuity of subjecthood, and a machine's own awareness of consequences; at the same time, both permanent affirmation and permanent denial of future artificial subjecthood carry a burden of evidence.

“AI as a layer of human civilizational capability” is therefore a present institutional position, not a final judgment about what kind of being AI is. If sufficiently strong evidence of new non-human subjects emerges, protection, legal standing, shared life, and political relations will have to be reopened.

Even then, another boundary would remain. Subjecthood is not the same thing as public office, and public office is not sovereignty. The questions Who is capable? Who experiences? Who deserves protection? and Who may make binding decisions for a community? remain distinct.

6. What a shared system actually needs

If a human-AI system is to achieve both correctness and efficiency, at least four things must remain continuously available: purpose must be traceable, authorization must be explainable, real-world outcomes must return, and error must be able to reopen judgment.

These four conditions matter more than whether a human being remains inside every operational step. People can move away from low-risk, repetitive, well-validated operations. They must re-enter when purpose changes, rights conflict, consequences become high-stakes, or an irreversible boundary approaches. AI may take on more routine capability, but it must not allow an obsolete purpose to keep executing simply because automation has made it smooth.

3 | One of the Most Important Uses of Stronger Intelligence: Seeing Consequences Before They Arrive

1. Civilization's most expensive way to learn is to discover error only after the consequence

Human-COS begins from a direct proposition: civilization-scale irreversible outcomes cannot wait for the outcome itself to serve as final validation. Its purpose is not to prove that AI judges better than humans. It is to help human beings see their choices earlier, more completely, and more openly while consequences have not yet occurred, options still exist, and action can still change direction.

This shifts the value of AI capability from “giving people answers” to “changing what can be seen before a decision.” Many major failures are not wholly invisible in advance. Facts, interests, power, expertise, and human responses are dispersed across different actors and never assembled into a common causal picture.

If stronger AI can offer civilization one especially important new capability, it is to reduce the proportion of consequences that become intelligible only after they have become history. The earlier consequence enters judgment, the more likely the future is to remain open while direction can still be changed.

2. From one answer to a causal map of selectable futures

Human-COS is not designed to output a single risk score. It makes actors and intentions, major facts, distributions of interest, key feedback loops, thresholds, alternative paths, unknowns, strongest objections, falsification conditions, and locations of responsibility visible at the same time. Its workflow then moves through perception, aggregation, simulation, adversarial challenge, independent replication, convergence, warning and opportunity detection, option comparison, decision records, and post-action testing and repair.

Its final output is defined as a causal map of selectable futures that can be repeatedly tested by reality, by different actors, and by different models - not as an answer that AI gives on humanity's behalf.

This matters because complex social problems often lack structured visibility more than they lack advice. If we act, who is affected? If we do not act, what changes? Where do delay, partial action, joint action, and alternatives lead the system? Which risks are approaching qualitative thresholds? Which apparent benefits merely shift losses onto third parties or the future?

One of the most valuable directions for stronger AI is to make this map wider, deeper, and faster while keeping it a map rather than a command.

3. The stronger the capability, the more important the separation between prediction and decision

One of the most important contributions of AGI-COS 1.1 is to downgrade model output from “the answer” to a recommendation about what the evidence suggests, together with its sources, confidence, and explicit unknowns. The common hard constraints also separate subjecthood, rights, authorization, and responsibility: AGI-COS does not acquire sovereignty from capability, and facts or recommendations do not automatically become decisions.

Human-COS follows the same discipline. Its outputs neither approve nor veto. They change what can be known and seen before a decision. Final choice remains with actors who actually hold the relevant rights, responsibilities, and authorization.

This is not distrust of AI capability. It is a distinction between types of question. Predictive accuracy asks how well a system sees some aspect of the future. Decision authority asks who is entitled to impose a consequence, and on whom. Even if the first becomes exceptionally strong, the second does not arise automatically.

The better future AI becomes at prediction, diagnosis, and option comparison, the more important it is to separate these questions in advance. Otherwise, technical success can quietly turn expert advantage into de facto power.

4. Causal simulation is not prophecy

Human-COS requires the strongest counterexamples, alternative explanations, and pathways capable of overturning the original

conclusion. It also requires different models, implementations, and actors to independently reproduce key simulations. What matters is not which model speaks with the greatest confidence, but which directional patterns recur after assumptions are changed and which disagreements refuse to disappear.

Seeing consequences earlier therefore cannot mean that the future has become fixed or fully predictable. Social systems contain feedback, adaptation, contingency, hidden variables, and learning actors. Many trajectories change because they have been seen.

A mature causal-simulation system should allow three outputs to coexist: what we know with relatively high confidence; which paths appear most worth avoiding or pursuing; and what remains unknown, including what new facts would force us to revise the judgment.

The civilizational value of stronger AI is not to turn an open future into a computable answer. It is to reduce the number of consequences that could have been seen earlier but were missed because of fragmented information, hidden power, cognitive limits, or short-term pressure.

5. Shared facts do not require uniform thought

Human-COS proposes that consensus need not come from making everyone think alike. It can emerge gradually from jointly testable facts, causal relations, distributions of interest, and conditions of continued existence. The system may simulate from the position of any actor, but it should not lose the final observational standpoint of humanity as a whole.

This means that strong AI used for public judgment should not be tasked with unifying values. It is better used to help different actors see whether, despite different values, institutions, and interests, some outcomes would destroy everyone's future option space; whether some forms of cooperation serve multiple actors' long-term interests; and whether some conflicts are being mistaken for inevitabilities because the information structure is wrong.

Causal transparency is not ideological uniformity. A sufficiently transparent causal system should make disagreement more precise: are we disagreeing about facts, causal inference, values, risk tolerance, or who has

the right to decide? Once disagreement is decomposed, consensus no longer depends on coercion.

6. Let consequences become causes before they arrive

In operational terms, this means using simulation, counterfactual comparison, threshold detection, red teaming, and real-world feedback so that plausible future consequences can influence a decision before those consequences become irreversible facts.

In ordinary decision-making, consequence follows action: act first, observe what happened, and carry the lesson into the next decision. The capability proposed by Human-COS attempts to let future consequences enter present judgment earlier, in the form of causal simulation, counterfactuals, thresholds, and alternative paths.

This is a change in temporal structure. “What happens if we continue?” should not remain only a retrospective explanation. It should become part of the present choice. The consequence has not yet occurred, but sufficiently clear causal evidence may already change today’s direction.

This capacity matters only under two constraints. First, it must remain open to rebuttal, replication, and correction by real-world outcomes. Second, greater predictive power must not give the system the right to bear consequences on behalf of real subjects.

4 | From Judgment to Action: Seek the Best Available, Not Static Perfection

1. Seeing more does not eliminate uncertainty

Chapter 3 treated expanded causal visibility as one important civilizational value of stronger AI. But seeing more does not mean waiting until everything is visible. The real world always contains missing information, model error, strategic response, contingency, and time pressure. A civilization that makes the disappearance of all uncertainty a precondition for action will be unable to act in many of the situations that matter most.

The cross-phase baseline of Human-COS treats every simulation as provisional. After real action, benefits, harms, lived responses, institutional friction, third-party reactions, appeals, failures, and new facts must return. Assumptions must be tested; causal claims may be overturned; unknowns may expand; risks may be upgraded or downgraded. One of the system's important capabilities is not permanent correctness, but the capacity for error to remain visible, reproducible, attributable, and consequential for the next run.

Judgment and action therefore form a loop. Judgment is not a one-time approval before action, and action is not an irreversible execution after judgment. A best-supported judgment is formed, reality is entered, new facts are acquired, and those facts change the next judgment.

2. “Best” can only be conditional

This principle can be called best-available judgment. After actors, facts, interests, objections, alternatives, and unknowns have been sufficiently surfaced, choose the most reasonable actionable path under current information, time, and responsibility conditions rather than waiting for a timelessly perfect answer.

“Best” here is not a system-declared global optimum, and it does not compress the three aims, rights, and irreversible harms into one aggregate utility function. It means provisional validity under current evidence: the reasons can be stated, the conditions can be traced, and new facts can trigger revision.

AGI-COS explicitly rejects the idea of a formula that automatically calculates net civilizational benefit. Real decisions must make benefit, harm, irreversibility, and alternative paths visible separately. An alleged overall optimum does not automatically create an obligation for a particular person.

Seeking the best available rather than static perfection does not relax rigor. It relocates rigor from the fantasy of getting everything right once to a harder discipline: reasons are recorded, action generates feedback, errors leave traces, and new facts are allowed to change the next choice. Robust Decision Making offers a close methodological analogue: under deep uncertainty, computation is used to stress-test strategies across many plausible futures rather than to optimize against one supposedly accurate forecast.⁴

3. Efficiency comes from reducing meaningless friction, not reducing the capacity to judge

When a system is new, new causal relations, authorization boundaries, and real-world consequences require denser validation. Once some low-risk, repetitive mechanisms have survived sustained real-world testing, every routine action need not reopen the full civilizational debate.

Dynamic Adaptive Policy Pathways offers a related logic: take intelligible near-term action while preserving longer-term options, monitor how the world unfolds, and change pathways when trigger conditions are reached. Rules can stabilize. Processes can automate. Interaction can decline.⁵

What must decline is repetitive friction, not civilizational judgment capacity. If purpose has not changed, authority has not expanded, the domain has not materially shifted, and risk has not escalated, the system can operate efficiently. When any of those conditions change, judgment must reopen.

This anticipates the later principle of system receding: rule density is not the same thing as interaction density, and complete authorization is not the same thing as repetitive authorization. Mature systems can remain strict in the background while becoming simpler in the foreground.

4. The nature of the consequence determines the learning rule

Best-available judgment has a boundary. If an action may directly cross a threshold that cannot realistically be restored, feedback arrives too late. In such cases, “try it and then correct it” is not a legitimate learning strategy.

5 | Two Kinds of Error: Reversible Experimentation and Irreversible No-Go Zones

1. Civilization cannot learn everything in the same way

Human-COS applies a strict underlying rule: the more irreversible the consequence, the faster it propagates, and the shorter the intervention window, the less acceptable it is to wait for the outcome as validation. Civilization-scale irreversible risk cannot be validated by allowing the irreversible outcome to occur. This is methodologically consonant with the precautionary logic in Article 3(3) of the UN Framework Convention on Climate Change: threats of serious or irreversible damage need not await full scientific certainty before cost-effective preventive measures are considered.⁶

This rule changes the logic of learning. In ordinary life, products, institutions, and local mechanisms, many errors can be corrected through real-world feedback. But some consequences kill the subject, permanently close rights, trigger wars or technological chains that cannot be recalled, or push ecological and social systems across thresholds that are exceptionally difficult to restore. In these cases, “practice is the test” cannot mean that catastrophe must occur once.

The framework therefore distinguishes two governance domains: reversible experimentation and irreversible no-go zones. The first permits controlled learning through reality. The second pushes learning upstream into simulation, historical backtesting, red teaming, sandboxes, shadow operation, low-risk prototypes, and independent replication.

Consider a system that could initiate nuclear-scale action without a reliable chain of authorization, interruption, and termination. It cannot be tested by allowing a real launch to determine whether its safeguards work. The relevant burden of proof must be met before the action can reach the world.

2. Reversible experimentation is not casual experimentation

Reversible does not mean costless. A withdrawable policy can still harm people. A stoppable model deployment can still produce real loss.

Reversibility means only that, after error is discovered, there remains a practically viable path to stop, isolate, repair, compensate, or reconstruct.

At minimum, reversible experimentation needs four conditions: the boundary is sufficiently limited; harm can be detected; stopping authority is real; and feedback arrives faster than risk propagates. Without those conditions, a “pilot” may simply transfer a high-stakes experiment onto people who do not know they are subjects of one.

Reversibility is not a license for boldness. It is a responsibility condition for learning. The more a system celebrates rapid experimentation, the more it must prove that it can actually stop, identify affected people, preserve evidence of error, and bear the burden of repair - not merely that it can deploy quickly.

3. Irreversible no-go zones are not unlimited safetyism

Strict treatment of irreversible risk must not become a doctrine under which anyone who invokes catastrophic danger may demand unlimited spending, unlimited surveillance, or indefinite suspension of ordinary life. Risk prevention can itself generate irreversible power structures.

AGI-COS therefore contains another hard constraint: safety needs do not automatically generate unlimited monitoring, resource requisition, or permanent emergency authority. Crisis can change speed and procedural intensity, but temporary powers cannot silently become the normal constitution of stable periods.

The object of heightened caution is not the word safety. It is a specified irreversible consequence: what exactly can happen, who is affected, how fast does it spread, where is the last realistically reversible point, what alternatives exist, and what evidence could overturn the current judgment?

At extremely high-consequence boundaries, higher preventive costs may be justified. But “any cost is worth paying” cannot be a general rule. A more defensible proposition is this: the more severe the irreversible consequence and the stronger the evidence, the stronger civilization’s reason to bear higher preventive costs while direction can still be changed.

4. The real dividing line is not conservative versus radical; it is whether correction remains possible

An apparently radical experiment may be safer if it is small, stoppable, transparent, and genuinely exitable than an apparently conservative institution that quietly locks in authority. Conversely, a measure described as temporary may already have entered an irreversible zone if data, authority, dependencies, and patterns of behavior become practically impossible to unwind.

Whether an action should proceed therefore cannot be decided only by asking how large the risk is. We must also ask: if we are wrong, can we still change the result?

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

1. Translating reversibility as “rollback” is not enough

AGI-COS continuity work already makes this clear: reversibility is not merely the existence of a rollback button. Purpose, rights, authorization, data, resources, capability, roles, and temporary controls should have object-specific paths of exit or return. Opening a high-consequence interface should be accompanied by withdrawal, pause, expiry, restoration, repair, and termination design. Recovery is not the resurrection of old authority; restart requires renewed judgment and renewed authorization.

Contemporary AI risk-management standards point in the same operational direction. NIST's AI Risk Management Framework includes post-deployment monitoring, appeal and override, incident response, recovery, change management, and decommissioning among the mechanisms for managing identified AI risks.⁷

True reversibility is therefore not the restoration of the whole world to a historical snapshot. Reality does not move backward. Facts remain. Relationships change. Knowledge grows. Error itself becomes part of the new conditions.

At the civilizational level, the more accurate concept is correction capacity. It protects not a return to the past, but the practical ability to change direction before irreversible consequence has occurred.

2. Correction can move in more than one direction

Once an error is discovered, the right action may be rollback. It may also be pause, downgrade, isolation, replacement, repair, redesign, upgrade, separation, reconstruction, or termination. If the old state was itself the source of the problem, restoring it wholesale only replaces a new failure with an old one.

This fits the phase logic of Civilizational Transition. A new civilization is not created by declaring every old institution invalid, nor by returning to the starting point after every failure. It forms gradually, while the old

world still exists, through experiment, failure, appeal, rollback, responsibility, and renewed choice.

The technical meaning of rollback must still be preserved. Backups, version restoration, low-tech alternatives, and human takeover remain important tools. Civilizational reversibility simply cannot be reduced to technical restoration.

3. Correction capacity is stricter than nominal exit

A system can state that users are free to exit while data dependence, network effects, tied resources, identity relations, or the disappearance of alternatives make exit practically impossible. Conversely, an institution may not be designed to restore an old state yet still preserve real capacities for substitution, reconstruction, and reauthorization.

Whether exit is real therefore cannot be decided by checking for an exit button. We have to ask whether the subject can still live after exit, whether alternative services remain available, whether old authority actually expires, whether data and resource relations truly end, and whether re-entry requires new authorization.

Consider an algorithmic welfare or credit system that displays an “opt out” button. If leaving the system means losing access to benefits, housing, health care, or other essential services, the exit is formal rather than real. Correction capacity requires a viable life outside the system, not merely a visible button inside it.

AGI-COS contains an important boundary here: historical facts may remain, but old authority cannot continue merely because history is preserved. Temporary control does not become permanent because it was once justified.

4. In the era of strong AI, correction capacity becomes infrastructure

AI increases execution speed and can make dependencies form faster. The stronger the system, the less acceptable it is to treat correction as an afterthought. Stopping, isolation, explanation, human takeover, alternatives, reauthorization, and practical repair have to exist from the architecture stage.

This turns reversibility from a safety feature into a condition of civilizational legitimacy. A high-consequence system that lacks real correction capacity should not acquire an effectively non-exitable position merely because its current objective is sound or its performance is excellent.

7 | Anti-Closure Is Not Indecision: Preserving the Capacity to Produce Better Choices

1. Anti-Closure opposes finality, not commitment

One of the most stable principles across AGI-COS versions is that civilization should not be permanently locked in by one judgment, one system, or one purported optimum. Anti-Closure guards not only against technical lock-in but against institutional lock-in, metric lock-in, narrative lock-in, and lock-in to a single model.

The principle is easy to misunderstand as “keep every option open forever.” That is not the claim. Civilization does not need to preserve every path that has been sufficiently shown to be destructive, nor does it need to postpone every decision indefinitely. Anti-Closure does not oppose decision. It opposes turning a phase-bound decision into an unquestionable, irreplaceable, and non-reauthorizable final settlement.

2. More options are not automatically better

If an open future is interpreted as maximizing the number of available options, absurd consequences follow. We would be forced to preserve paths that systematically destroy basic rights merely in the name of diversity of choice.

What matters is not how many old options remain in storage. What matters is whether civilization can still generate, recognize, and move toward better new paths. Anti-Closure therefore protects the continuing capacity to form better choices, not a museum of inherited options.

A civilization may close a door that has been shown to be dangerous while preserving the ability to research again, judge again, and revise boundaries when new evidence appears.

3. The most dangerous closure often arrives in the name of success

Failure forces people to search for alternatives. Success can become a reason for permanence. A metric that once worked may expand into a total civilizational score. Emergency authority that once helped may become

routine power. A model that performs well for years may gradually become an unstated judge of fact.

AGI-COS therefore limits metrics to observation, comparison, warning, evidence organization, and model input. Metrics must not automatically enlarge authority, reduce rights, reassign sovereignty, distribute resources, or generate a total civilizational score.

The sentence Anti-Closure most needs to resist is deceptively reasonable: “It has worked for so long that we no longer need to ask.” A system is worth watching not only when it fails, but when success has become habit and habit begins to substitute for renewed judgment. AI-governance research identifies a parallel institutional danger: early choices can create path dependence and lock in later governance arrangements.⁸

4. An open future is compatible with irreversible boundaries

Keeping the future open does not require neutrality among all directions. Continued existence, basic rights, and boundaries against irreversible harm are themselves conditions under which future choice remains possible. Some irreversible no-go zones protect a larger space of future openness.

Anti-Closure can therefore be summarized in one sentence: it does not reject direction; it rejects allowing any direction to lose the conditions for re-examination while it may still be wrong.

8 | What Exactly Are We Governing? Purpose, Authority, Conduct, Responsibility, and Consequence

1. “Regulating AI” is too coarse a description

If we are allowed only one governance question, “How should AI be regulated?” is still too coarse. AI may refer to a model, product, agent, infrastructure layer, organizational decision aid, personal extension, public-system interface, or perhaps one day an artificial subject. Treating all of these as one governance object mixes capability, deployment, power, and consequence again.

The existing hard constraints of Civilizational Transition already suggest a decomposition: capability does not automatically generate authority; prediction does not automatically generate decision rights; an aggregate optimum does not automatically generate individual obligation; safety needs do not automatically generate unlimited monitoring or permanent emergency power; and system roles must remain exitable, replaceable, degradable, and stoppable.

2. Ask five different questions first

Governance should at least ask five questions in sequence: Why govern? What capability, conduct, deployment, or consequence is actually being governed? Who possesses legitimate authority? Who bears responsibility? And when must the governance arrangement itself expire, contract, or be reauthorized?

Once the questions are separated, vague debates become more concrete. Whether a high-capability model requires evaluation is a capability-risk question. Whether a hospital may allow a model to alter treatment automatically is a question of professional responsibility and authorization. Whether a government may use models to influence citizens at scale is a question of public power. Whether a platform may retain an entire personal-memory record is a question of data and control. All involve AI. They are not the same governance problem.

3. Capability itself can still be a direct object of governance

Reframing the governance object must not swing to the opposite extreme. Some capabilities may generate serious risk before a particular deployment purpose is specified - for example, easily diffused dangerous capabilities or autonomous chains that cannot be reliably controlled. Capability itself can then be a legitimate object of direct assessment or restriction.

The discipline is causal specificity. We should be able to explain why the capability itself constitutes a risk, rather than assume that greater strength is sufficient reason for restriction. Governance should respond to mechanisms of harm, not to an abstract fear of intensity.

4. Complex systems must not dissolve responsibility

The Human-COS baseline requires high-consequence public tasks to identify who proposed the objective, who holds authority, who provided professional judgment, who executes, who can stop or roll back, and who bears responsibility for the outcome. Human-COS may change what can be seen before a decision; it cannot allow the decision-maker to hide behind the name of a system.

This becomes more important as technical chains grow more complex. A model recommends, an organization approves, an algorithm executes, a platform distributes - and at the end no actor acknowledges having decided anything. The longer the technical chain, the more visible the responsibility chain has to become.

One deeper task of AI governance is therefore not to assign a regulator to every risk, but to restore traceability among purpose, authority, conduct, responsibility, and consequence. OECD guidance similarly places risk management across the AI value chain and states that those who deploy AI should be accountable for the proper functioning of their systems.⁹

5. Those who govern must also be accountable

Regulation is not intrinsically correct. Any governance arrangement that invokes safety, public interest, or civilizational risk must answer questions about its own source of authority, evidentiary standard, affected parties, appeal, duration, and exit conditions. Research on human oversight

likewise shows that overseers are fallible; effective oversight depends on institutional design rather than an assumption that human involvement is inherently reliable.¹⁰

Otherwise, “regulating AI” can itself become a route to irreversible power expansion. This closes the loop with Anti-Closure and correction capacity: AI systems must be correctable, and the institutions governing AI must be correctable as well.

9 | Why Many AI-Governance Problems First Reveal Human Structures

1. Before asking what AI amplifies, ask what it is connected to

Human-COS never treats major consequences as a single-technology problem. It starts from observable causal chains and feedback loops among technology, capital, state competition, organizational incentives, political power, public discourse, human instincts, judgments by key actors, and group psychology. The hard part is often that information remains local, causal links are severed, interests are hidden, organizations filter bad news, and people under extreme pressure cross boundaries they would normally respect.

When AI enters these structures, it does not create every direction from nothing. It often accelerates directions already present. A firm's short-term metric can be automated. A state's security dilemma can acquire stronger technological reach. An organization's resistance to bad news can be reinforced by a more efficient information system.

Many current AI-governance problems therefore first reveal structures of human purpose, rivalry, power, incentive, and responsibility. Strong intelligence accelerates and magnifies them and makes them harder to ignore. If responsibility is transferred wholesale to AI, the upstream causes remain untouched.

2. Structure before personality

One frozen Human-COS rule is “structure before personality”: first ask what real structure is pushing otherwise ordinary actors toward dangerous or constructive choices, rather than beginning with personality scoring, loyalty judgments, or psychological diagnosis.

The same principle applies to AI governance. If an organization rewards speed but not the reporting of bad news, if interstate competition makes any slowing appear to be strategic defeat, or if a platform business model rewards polarization, asking AI to be “more benevolent” addresses only a small downstream part of the problem.

Effective AI governance must therefore also change upstream structures that repeatedly generate bad objectives. Otherwise each new model can be realigned while the surrounding incentives continue pushing the system in the same direction.

3. Why Upstream Change Cannot Be System-Centered

Once structural causes are recognized, a new misunderstanding becomes possible: perhaps correcting them requires a single superior system capable of seeing and redesigning the whole. That conclusion would reproduce the very form of closure this framework rejects.

Changing upstream structures does not mean waiting for competitors to become morally better, and it will not be accomplished by one system. Individuals, communities, civil-society organizations, firms, local and national governments, alliances, regional communities, international organizations, scientific communities, transnational networks, and other actors are already searching for different paths through diplomacy, law, markets, institutional reform, scientific cooperation, public communication, crisis management, local experimentation, and social practice.

4. How Plural Actors Can Change the Structure

A more realistic direction is to help those efforts change the payoff structures, information structures, conditions of misperception, exposure to shared risk, and long-run interaction faced by the actors involved. The aim is to reduce the returns to hidden defection, risk transfer, misperception, and endless escalation, while making verification, communication, risk insulation, conditional cooperation, and long-run feedback more compatible with each actor's own interests. This is a direction for further work, not a finished unified mechanism.

The early study *Calculating Competition and Cooperation between China and the United States* is only one bilateral research prototype in that direction. It tries to recast cooperation from a moral demand into questions of risk, payoff, and thresholds. It does not establish that real-world cooperative mechanisms have been solved, and it does not represent the full range of efforts already underway elsewhere.

Future mechanism research must absorb existing human practice and institutional experience and extend across groups, states, alliances, blocs, regional communities, multilateral institutions, cross-domain coupling, and domestic-international interaction. That work belongs in a separate project.

If Human-COS is eventually completed and proves useful in trial operation, it may become one augmentative tool among these efforts: helping different actors surface objectives, constraints, interests, and behavioral feedback; compare the long-run consequences of different paths; identify thresholds; and return real-world outcomes more quickly to the next round of judgment. It does not replace existing actors, monopolize the direction of mechanism formation, or exclude other AI tools, analytical methods, institutional innovation, or human practice.

5. This Does Not Cancel Machine-Side Risk

A possible overreach must be stopped explicitly: saying that many AI problems first reveal human structures does not mean that AI itself has no problems. Complex models may hallucinate, be attacked, develop difficult-to-interpret strategies, create cross-system cascades, or execute autonomously beyond the expectations of their designers. Stronger future systems may introduce risks that cannot yet be adequately characterized.

This is precisely why Civilizational Transition retains AGI-COS alongside Human-COS. A framework for human judgment and direction cannot substitute for machine-side work on perception, authorization, execution, upgrading, autonomy, and inability to stop. The 2026 AI Index reports rising documented AI incidents, declining frontier-model transparency, weaker safety performance under adversarial attack, and unresolved trade-offs among safety, fairness, and privacy.¹¹

The Human-COS - Nebula - AGI-COS interface makes the deeper point explicit: the danger is not that one side is inherently worse. It is that two kinds of capability can amplify one another when mutual constraint is weak. Human fear, interest, and misjudgment may give a machine a bad direction. Machine speed, scale, and autonomy can then harden that direction rapidly. Machine-side error can also feed back into public psychology, politics, and the next round of human judgment.

6. Do Not Search for a Permanently Correct Side

This makes “Who is more trustworthy, humans or AI?” the wrong level of question. Civilization cannot rest on the hope that either side will remain permanently correct. A more realistic objective is to prevent human error about direction and machine operational error from passing straight through into irreversible reality.

Within the Civilizational Transition toolset, Human-COS is meant to help actors reduce errors about direction; AGI-COS is meant to reduce machine-side errors of perception, authorization, and execution; Nebula is intended to let authorized choices enter reality while keeping outcomes visible and correctable. These are only one set of tools and interfaces within the framework. They do not replace the world's wider mechanisms of action. Civilizational correction still occurs through the institutions, relationships, actions, and feedback of multiple real actors.

The systemic answer is therefore not a single-cause claim. Human upstream structures and machine-side risks coexist. The former strongly influence what powerful capability is used for. The latter determine whether new forms of loss of control arise during operation. Governance has to address both.

7. The Problem Returns to How Civilization Treats Subjects

Whether the risk originates in human direction or machine operation, the consequences are ultimately borne by concrete subjects, relationships, communities, and future people. As AI becomes capable of performing more functions that human beings once performed, a deeper question therefore appears: what is the human place actually based on?

If human value rests only on “what machines still cannot do,” then every advance in AI forces humanity to retreat another step. Civilizational Transition cannot build the meaning of human life in the shrinking gaps left by technical limitation.

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

1. Do not build human value on what machines temporarily cannot do

Each time AI acquires a new capability, people search again for the task that “only humans can do.” This defensive strategy will eventually fail because the boundary of technical capability will continue to move. If human worth depends on a function being temporarily irreplaceable, human worth declines every time the technical benchmark advances.

The Civilizational Transition framework begins elsewhere. AI can expand observation, understanding, expression, and action. Greater knowledge does not automatically give it authority to make value judgments on behalf of a subject. A person's body, history, relationships, responsibility, finite time, and consequences that must actually be borne cannot be replaced merely by a model that predicts more accurately.

2. Human value does not depend on preserving inefficiency

Strong AI can assist care, learning, creation, work, decision-making, and relationships. A mature civilization does not need to preserve inefficient labor merely to prove that human beings still matter, nor does it need to reject automation in the name of subjecthood.

Human irreplaceability is not a monopoly over functions. It is a position in existence. A person lives his or her own life, bears relationships and choices, faces time that cannot be transferred, and within those experiences generates, recognizes, questions, and changes meaning.

The Living Civilization trilogy brings this principle back to ordinary life: a future civilization must eventually enter a child's growth, two people's love, family, friendship, work, illness, aging, death, and an ordinary morning. The system must eventually learn to recede. The person must still live the life.

3. AI can extend presence without turning life into delegated completion

Spark offers another kind of human-AI relationship: not replacing the person, but amplifying the person. AI as an authorized extension of a real human being can carry care, capability, judgment, and presence farther.

Living one's own life does not mean doing every task personally. A person may use machines to see far away, use AI to express complex thought, use remote systems to care for family, and hand large amounts of repetitive labor to automation. The question is not whether an agent was used. The question is whether relationship, purpose, and consequence remain anchored in the subject.

As delegated capability becomes stronger, the danger is not simply “AI did too many things for me.” It is that a person's life may gradually be delegated away until the person is left mainly consuming experiences optimized by the system.

4. Meaning is not the last job reserved for humans

Meaning is foundational, and its sources may differ. AI may help understand, protect, convey, and extend human meaning. But a role does not settle what a being is, and understanding another subject's meaning does not create ownership over it.

Meaning should therefore not be treated as the final high-level task left to humans after automation has finished everything else. Meaning is not a job category. It is an ongoing reality formed among existence, relationship, choice, action, memory, responsibility, and experience.

Even if genuinely new non-human subjects appear in the future, this principle need not collapse. New subjects may form meanings of their own. Their meanings would not cancel already existing human meanings, and subjecthood would still not automatically generate authority over others.

5. Scarcity begins to migrate

As more cognitive and executive capability becomes abundant, the scarce objects of civilization may shift. In the past, scarcity often centered on who could calculate, who held knowledge, and who could organize resources. In the future, what becomes scarce may be whether subjects still possess a real place in the world: whether they can understand how their choices

enter a shared reality, form relationships, bear responsibility, and live a life without the system deciding it on their behalf.

This is also why system receding is not a secondary design issue. If a person's place ultimately depends on constant system rewards, reminders, rankings, and prompts, subjecthood remains propped up by the system. A mature civilization should allow stronger capability to exist while making people less dependent on being driven by it.

11 | A Mature Relationship: The Stronger the Capability, the Better the System Knows When to Recede

1. Maturity is not a system moving ever deeper into life

The Living Civilization framework states the principle clearly: a mature system does not prove itself by penetrating ever more deeply into life. It remains reliable where necessary and keeps its distance where unnecessary. It does not perpetually manufacture participation. It allows a new civilizational logic to persist through the choices, relationships, habits, and institutions of subjects after visible external driving forces have weakened.

If a behavior must forever depend on rewards, rankings, system reminders, moral exhortation, or AGI prompting, it remains primarily externally driven. The endpoint of success is not a system that becomes ever better at driving people. It is people becoming less dependent on being driven by the system.

2. Receding does not mean weaker capability; it means lower intrusion

There is no contradiction with strong AI here. A stronger system may verify authorization more accurately in the background, detect anomalies earlier, and provide help at lower cost without continuously occupying attention, choice, and relationship. Greater capability does not require greater foreground presence.

System receding depends not on lowering the integrity of rules but on moving strict validation into the background and reopening confirmation only where the relationship has materially changed: purpose changes, authority expands, the domain changes, major risk appears, a term expires, or the subject withdraws.

This also answers the earlier efficiency problem. Procedural steps and interaction density may decline. Traceability of purpose, completeness of authorization, responsibility records, anomaly feedback, and the capacity to reopen judgment may not.

3. Low intrusion is not invisible control

Background operation creates a danger of its own. A system may appear to recede while acquiring deeper invisible influence. If users cannot see which objectives are being optimized, what data are being used, or how choices are being ranked, low interaction can become high control.

A recommender or agent can interrupt a user only rarely and still shape choice through rankings, defaults, and background objectives. Low interaction is not low power.

System receding must therefore move in two directions at once: everyday disturbance to the subject goes down, while auditability of power, purpose, and anomaly does not. What recedes is prompting, not responsibility; frequency of intervention, not the boundary itself.

4. A civilizational logic becomes real when it can sustain itself

The Living Civilization framework distinguishes stability from self-maintenance. In the self-maintaining phase, rewards, rankings, strong prompts, administrative mobilization, and algorithmic driving gradually recede while the new logic continues through autonomous choice, relationship, habit, institutional inertia, and intergenerational reproduction.

A civilizational system therefore cannot treat its own continued visibility as the measure of success. The most successful Human-COS is not one that everyone uses every day. Success would mean that society becomes better at seeing consequences earlier, allowing dissent, and preserving correction even when the tool is not constantly present. The most successful AI support is not a life saturated by AI. It is strong support when needed and an intact human life when not needed.

Strong capability, low intrusion, and self-maintenance describe a mature human-AI relationship very different from the assumption that more control always means more safety.

5. Receding brings the argument back to its beginning

This essay began with the question of whether AI becoming too strong is itself the problem. By this point, the reversal should be visible. A mature civilization does not need to weaken AI as proof of safety. It needs strong

capability to operate within sound direction, explicit authorization, continuing feedback, and real correction - and ultimately not to prove its value by occupying more and more of human life.

12 | Conclusion: When Intelligence Ceases to Be Scarce, What Becomes Truly Scarce?

1. When intelligence moves from scarcity toward infrastructure

“When intelligence ceases to be scarce” is not a statistical description of the present, and it is not a prediction that everyone will receive equal access to identical intelligent capability. It is a thought experiment about direction. If increasingly powerful capacities for cognition, analysis, generation, and execution become widely available infrastructure, where do civilizational bottlenecks move?

Nor should growing capability be confused with equal access. The 2026 AI Index reports that industry produced more than 90 percent of notable AI models in 2025 and that global AI compute capacity expanded rapidly, while frontier-model production, chips, data centers, and core infrastructure remained concentrated. Capability becoming stronger and more widespread does not mean that every state, organization, or person acquires equal control, interpretability, or access to alternatives.¹²

The earlier chapters suggest the answer. Greater capability does not automatically supply purpose. Better prediction does not automatically generate decision rights. Faster execution does not automatically correct a bad direction. More information does not automatically create shared understanding. A stronger system does not automatically give subjects a more substantial place in the world.

2. The first scarcity: direction

As “how to do it” becomes easier, “why do it” becomes more important. A civilization moving in the wrong direction may reach the wrong result faster when it possesses greater capability.

Direction is not a grand value slogan. It is civilizational judgment that can enter concrete objectives, rights boundaries, authorization, and responsibility. Human-COS is meant to support direction not because it possesses the final answer to value, but because it tries to make visible, in advance, how different choices alter people and the world and whether a better path exists.

3. The second scarcity: causal understanding of the whole system

Intelligence can generate vast numbers of answers. The rarer capability is to place dispersed facts, different actors, short-term interests, long-term feedback, and irreversible thresholds into one contestable causal picture.

Civilizational failure often occurs not because nobody is intelligent, but because each actor can be locally rational while the whole system still moves toward a bad outcome.

Causal transparency is not uniform thought. It makes disagreement more precise. It lets actors see whether they disagree over facts, causal inference, values, risk tolerance, or decision rights, and it gives continued existence and shared consequence a chance to enter judgment before hostility does.

4. The third scarcity: time before irreversibility

One of the most expensive resources of the AI era may be correction time. The faster capability becomes, the shorter the interval between a bad judgment and its translation into reality. That makes it more important to preserve windows for pause, challenge, substitution, and reconstruction before irreversible consequence.

This is why reversibility cannot mean only returning to an earlier version, and Anti-Closure cannot mean only preserving options. Together they protect civilization's ability to change the result before error becomes final.

5. The fourth scarcity: a real place for subjects

When knowledge, expression, creation, agency, and execution can all be massively extended, the scarce question is not “Which jobs must still be done by humans?” It is whether people still possess a real place in the world: the ability to live, relate, choose, refuse, bear responsibility, be needed, create, and know that their actions genuinely reach other lives.

Meaning cannot be delivered uniformly by a system, and AI cannot complete it on behalf of the subject. AI may amplify human capability, and genuinely new non-human subjects may perhaps appear in the future. But abundance of capability must not reduce life to the receiving end of system optimization.

6. The fifth scarcity: shared correction capacity

The deepest safety of a civilization is not an AI that never makes mistakes, nor a population of humans who never make mistakes. It is a condition in which facts can still travel upward, dissent can still appear, responsibility can still be located, systems can still be stopped, authority can still contract, old answers can still be replaced, and new paths can still be generated.

Within the Civilizational Transition toolset, Human-COS can be understood as an augmentative tool for judgment about direction, AGI-COS as machine-side operational protection, and Nebula as a possible environment for coordination and feedback. Reality then tests the tools again. Their purpose is not to build one larger total system and certainly not to cover over the existing actions of individuals, groups, states, alliances, and communities. Their purpose is to add capabilities that make it harder for a single error to pass directly into irreversible reality.

7. AI governance returns to the civilizational question

If one governance judgment should change after reading this essay, it is not a single regulatory rule but a shift in viewpoint: present AI governance cannot govern only AI. It must also govern the purposes, authorities, and incentives humans attach to AI; the capabilities and operational risks of machines themselves; the real consequences created by their coupling; and the power boundaries of the governance institutions doing the governing.

This does not abolish conventional AI safety, model evaluation, privacy, accountability, regulation, or international rules. It places them inside a longer causal chain.

8. What civilization itself must learn to do better

The future worth pursuing is neither one that keeps AI permanently weak as proof of safety nor one that leaves human beings standing forever beside a nominal final button. The more important task is to make increasingly powerful intelligence a shared civilizational capability while civilization itself develops matching capacities for direction, causal understanding, responsibility, and correction.

As intelligence becomes less scarce, what becomes scarce is not simply more intelligence. It is sound direction; the ability to see whole causal

systems; time that still exists before irreversibility; a real place for subjects in the world; and the power of a civilization to choose again after discovering that it was wrong.

The measure of whether civilization uses strong AI well is therefore not how weak AI has been forced to remain. It is whether civilization becomes better able to continue, whether lives can be more fully lived and meaning can continue to emerge, and whether - after acquiring enormous capability - civilization still retains the capacity to judge, correct, and become itself again.

Notes on Evidence, Mechanisms, and Open Questions

External literature in the Chinese baseline supports specific facts, methods, and institutional experiences. It does not prove the Civilizational Transition framework as a whole.

The three civilizational aims, meaning as foundational, the account of human irreplaceability as a position in existence, the civilizational extension of correction capacity, the positive definition of Anti-Closure, and the five scarcities in the conclusion remain normative claims and theoretical syntheses of this framework.

Whether future AI develops subjective experience, non-transferable continuity of subjecthood, and corresponding standing remains open. Current institutional responsibility assigned to humans must not be converted into a permanent conclusion about what kind of being future AI may become.

Chapter 9 gives only an account of possible directions of changing payoff structures, information structures, and conditions of shared risk. Existing efforts by individuals, groups, organizations, states, alliances, communities, international institutions, and transnational networks are the real actors from which mechanisms emerge; they are not variables waiting for a new system to activate them.

The China-US calculable-cooperation study remains only an early bilateral prototype. Alliance, bloc, community, multilateral, and cross-domain mechanisms require separate research.

If Human-COS is completed and enters meaningful trial operation, it may become one tool for multi-actor causal simulation, counterfactual mechanism comparison, threshold identification, and feedback. Other AI tools, existing institutions, professional methods, and the judgments and actions of real actors remain independent parts of mechanism research and practice. This edition does not pre-claim Human-COS effects that have not been demonstrated in trial use, and it does not treat Human-COS as the only tool or direction.

Further Reading and Research Entry

If this essay has shifted the question from how powerful AI becomes to where civilization intends to take that power, the following English-language entries provide the next reading paths.

English Global Entry - The main English entry into the Civilizational Transition framework. [Open](#)

Civilizational Transition: A Global Introduction - A concise introduction to the framework's overall problem and direction. [Open](#)

Why Civilization Must Transition in the Age of AGI - The AGI-specific civilizational argument. [Open](#)

A Living Civilization: What Could Life Actually Be Like? - The framework translated into ordinary life, relationships, work, care, and human experience. [Open](#)

AGI-COS: Keeping Civilization Open Under Advanced Intelligence - The machine-side governance and Anti-Closure entry. [Open](#)

1 Stanford Institute for Human-Centered AI, The 2026 AI Index Report, especially “Research and Development” and the report overview.
<https://hai.stanford.edu/ai-index/2026-ai-index-report>

2 UNESCO, Recommendation on the Ethics of Artificial Intelligence, adopted 2021.
<https://www.unesco.org/en/artificial-intelligence/recommendation-ethics>.
This supports current institutional responsibility, not a permanent ontological conclusion about future AI subjecthood.

3 Liming Zhu, Qinghua Lu, Ming Ding, Sung Une Lee, Chen Wang et al., “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 (Springer, 2019), 23-51, <https://doi.org/10.1007/978->

3-030-05252-2_2.

5 Marjolijn Haasnoot, Andrew Warren, and Jan H. Kwakkel, “Dynamic Adaptive Policy Pathways (DAPP),” in *Decision Making under Deep Uncertainty* (Springer, 2019), 71-92, https://doi.org/10.1007/978-3-030-05252-2_4; see also Marjolijn Haasnoot et al., *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, Article 3(3), on precaution where there are threats of serious or irreversible damage and on cost-effective measures.

<https://unfccc.int/resource/docs/convkp/conveng.pdf>. Used here as a methodological analogy, not as validation of Human-COS.

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>. MANAGE 4.1 includes post-deployment monitoring, appeal and override, decommissioning, incident response, recovery, and change management.

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,” OECD.AI. <https://www.oecd.org/en/topics/sub-issues/ai-risks-and-incidents.html>

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, *The 2026 AI Index Report, “Responsible AI.”* <https://hai.stanford.edu/ai-index/2026-ai-index-report/responsible-ai>

12 Stanford Institute for Human-Centered AI, *The 2026 AI Index Report, “Research and Development.”* <https://hai.stanford.edu/ai-index/2026-ai-index-report/research-and-development>