

The Only Path

Long-Term National Interests and Civilization Leap in the Age of Artificial Intelligence

Structural Reasoning and a Working Master Text for Decision-Making Teams

Author: Zijun Fu | Civilization Leap Research Group

Version: v6.23 | FINAL_FROZEN | PUBLIC_CANONICAL

A nation may lead in a competition, yet it cannot be truly secure in a world that is collapsing.

A foundational perspective and a starting point for addressing the problem

This text examines long-term national interests in the AI/AGI age. Yet those interests depend not only on technology, resources, military power, and institutional arrangements, but also on whether nations, groups, and individuals still believe they have a future they can look forward to.

Human beings are driven by meaning as well as interests. Meaning can lead people to sacrifice themselves and protect others; it can also lead them to destroy others, damage systems, and destroy themselves. To judge a source of meaning, we cannot ask only whether it is logically coherent and structurally complete. We must also ask whether it leads life, relationships, and civilization toward flourishing, sustaining support, and continuity, or toward closure, revenge, elimination, and extinction. This perspective helps explain why the text moves from crisis emergence to collaborative coexistence and then to Civilization Leap: collaborative coexistence prevents crises from spiraling out of control; Civilization Leap addresses their sources.

Summary of the core judgment for decision-makers

This document is intended to help decision-making teams affirm a judgment and begin a preparatory step. Before presenting the full reasoning, this opening summary sets out the core judgment and the immediate action, so that readers can grasp the direction of the argument before entering it.

The real question facing nations. The real question is not 'What is AI?' Nor is it simply how to catch up with another technological wave. It is whether, in an age when general-purpose AI capabilities are spreading, the thresholds for creating risks keep falling, and irreversible crises are beginning to converge, a nation can safely reach the future of security and development it expects. This text argues that it still can, provided it rethinks how crises arise, identifies anew the constraints on long-term national interests, and redefines how those interests are realized.

The mechanism of crisis: crisis emergence, rather than an isolated risk. The diffusion of general-purpose AI capabilities spreads the sources of crisis from a few domains into countless application settings that resist advance classification. At the same time, the threshold for generating a crisis falls to the level of medium-sized actors and even individuals. Multiple risks no longer merely coexist. Through coupling, interlocking, acceleration, and the outstripping of human response capacity, they converge to produce new forms of crisis. This text calls that process crisis emergence. At a deeper level, many external risks ultimately enter behavior and decision-making through human fear, humiliation, loss of place, expectations for the future, and judgments of meaning. When individuals, groups, and nations lose a future they can look forward to, every risk is reinterpreted as a threat to survival.

The limits of existing approaches: each has a function, but none can be the destination on its own. Technical controls and new forms of deterrence provide a foundational base; joint management by major powers and limited cooperation serve as transitional mechanisms; a single global authority is ruled out. Each of the five

traditional paths has a place, but none can, by itself, fully satisfy the four requirements of security, development, legitimacy, and long-term survival.

The only long-term direction: collaborative coexistence plus Civilization Leap. Collaborative coexistence means establishing boundaries shared by all parties against irreversible risks while competition continues. Civilization Leap means going further, on the foundation of shared survival, to address the sources of crisis and remove the conditions that continually generate and amplify its sparks. 'Only' does not mean that there is only one possible plan. It means that there is only one direction: seeing the interlocking whole and changing the old way of playing the game remains the only long-term direction that has not been refuted.

A new definition of long-term national interests. In the AI age, long-term national interests are no longer simply a matter of overpowering rivals, maintaining advantages, and securing resources. They mean preserving the shared foundation for survival while safeguarding sovereign survival, institutional security, and autonomy in development, and ultimately bringing the nation to a condition in which no citizen is excluded from a foreseeable future.

The immediate action: open negotiations on a joint response mechanism for irreversible risks. The first negotiating table need not begin with universal harmony, the relative merits of political systems, or resource redistribution. It should begin with irreversible risks. These lie closest to the minimum interests all parties share and are easiest to incorporate into national-security reasoning. Initiators and participants can be separate, relieving the pressure that 'whoever proposes it first shows weakness.'

In one sentence. Civilization Leap need not begin with complete consensus. It can begin with a negotiating table.

The complete chain of reasoning

The full chain developed in this text is set out below to help locate each part within the overall argument:

Diffusion of general-purpose AI capabilities → multiplication of crisis sources, lower thresholds, and difficult attribution → risks converge through coupling, interlocking, acceleration, and outstripping human response, producing crisis emergence → isolated, issue-by-issue governance cannot adequately close the overall risk loop → the five traditional paths each have functions but cannot independently provide a long-term destination → long-term national interests must be redefined → negotiations on a joint response mechanism for irreversible risks must begin → collaborative coexistence becomes the long-term institutional direction → Civilization Leap becomes the fundamental path for resolving the sources of crisis and making collaborative coexistence self-sustaining over time → whether nations, groups, and individuals regain a future they can look forward to becomes the ultimate test.

One underlying thread runs throughout this chain: external risks narrow expectations for the future, and the loss of those expectations in turn amplifies external risks. The chain therefore leads not to a particular institutional endpoint, but to the restoration of a future that nations, groups, and individuals can look forward to.

How the three frameworks fit together

To avoid confusion, one point must be explained at the outset. This text uses three apparently different numbered frameworks in different places. They are not three different theories, but ways of developing the same risk structure at different levels. One describes the sources of risk; one gives decision-makers a quick understanding; and one develops the argument layer by layer in the main text. Their relationship is as follows:

Framework	Question answered	Content	Where used
Seven categories of external risk plus underlying human factors	Where does danger come from?	AI militarization; resources and computing power; AI-enabled biological risks; critical infrastructure; civilizational narratives; the AGI learning environment; Earth's finite limits. Underlying human factors form the layer that converts external pressures into emotion, meaning, and action.	Core judgment; chapter on crisis emergence
Six compressions plus an overall compression of future expectations	How do risks narrow a nation's room for maneuver?	Decision time, the retaliation window, room for attribution, development space, room for compromise, and shutdown authority; ultimately, expectations for the future	Quick-read summary; attachment package
Four interlocking strategic contests	How does the main text develop the reasoning?	Four main lines: AI militarization, resources and computing power, civilizational narratives, and Earth's finite limits	Chapters 6–10

In short: the seven categories reveal sources, the six compressions reveal pressures, and the four contests develop the reasoning. These are three perspectives on the same structure of crisis emergence, not three theories: sources, pressures, and reasoning. Readers encountering different counts in different chapters can refer back to this section.

A boundary of the text's task also needs to be stated here, echoing its conclusion: this document does not set out the institutional design of Civilization Leap in full. Its task is to demonstrate why shared survival cannot stop at crisis management and must continually lead toward Civilization Leap. How Civilization Leap is implemented through an economy of meaning, contribution mechanisms, resource-use mechanisms, mechanisms through which people are needed, and other arrangements belongs in subsequent specialized texts. Stating this now makes clear from the beginning that the presence of a direction without a detailed plan at this stage is an intentional open space, not a weakness in the argument.

Relationship to China's AI governance initiatives and related policy framework

This text develops an independent structural argument. This section explains its internal consistency with, and complementary relationship to, China's AI governance initiatives and related policy framework. It does not replace any government document, international initiative, technical standard, or policy proposal; nor does it repeat existing policy language.

AI for good is a necessary direction for harmonious coexistence between AI and humanity. This text understands 'AI for good' as a structural condition, rather than an externally imposed constraint. The reason follows from the convergence of the argument itself: as general-purpose AI capabilities spread, open-source capabilities continue to evolve, and barriers to use keep falling, more nations, organizations, businesses, groups, and even individuals can access, deploy, and modify powerful intelligent tools. Whether AI serves good then becomes more than a question of technological governance. It becomes a structural condition for long-term national interests, humanity's common security, and civilization's continued survival. In an age of irreversible risks, AI for good is not an optional value preference. It is a binding requirement for the long-term coexistence of AI and humanity.

AI for good increasingly requires a foundation of human goodwill. Intelligent tools do not operate independently of human goals, institutional settings, and the contexts in which they are used. They are deeply embedded in actual societies and bound up with people's interests, fears for survival, resentment, anxieties about development, and crises of meaning. Where trust is stable, people's worth is recognized, and the future offers something to look forward to, AI is more likely to become a vehicle for creating value, safeguarding order, and cooperative development. Conversely, when individuals, groups, and nations remain trapped in destabilizing conditions of fear, humiliation, loss of place, displacement, and a blocked future, the same capabilities may be used for manipulation, deception, revenge, destruction, and confrontation. The more

general-purpose the capability and the more comprehensive its reach, the less future expectations and support for meaning can be treated as soft background issues outside technological governance.

Human goodwill needs Civilization Leap to sustain it. Human goodwill cannot be maintained through moral demands or appeals to good intentions, because actors' choices arise not from abstract ethics but from actual pressures: security contests, struggles over resources, constrained development, opposing identities, loss of meaning, and shrinking expectations for the future reshape how actors see the world, perceive others, and define their interests. Giving people hope and a future they can look forward to is the driving source of human goodwill. For that hope to remain credible over time, the institutional environment, structures of meaning, incentives, and future expectations that sustain it must change. That is what Civilization Leap means. It neither denies the instinct for self-interest nor asks nations to abandon core interests. It upgrades the way self-interest is realized, making creation, protection, contribution, and being needed more advanced and more sustainable paths to it.

AI for good, human goodwill, and Civilization Leap thus form a complete internal chain of reasoning. AI for good establishes the necessary direction for harmonious human-machine coexistence. Technical rules alone cannot sustain it permanently; it must have human goodwill as its social foundation. Human goodwill, in turn, cannot endure through morality alone; Civilization Leap must provide institutional support, sustain meaning, and reshape motivation. The division of work with the related policy framework is therefore clear. Policy establishes broad principles, value directions, and institutional frameworks; this text supplies structural reasoning and develops the civilizational implications. Entering the same problem from different starting points, they meet around these questions: How can AI genuinely benefit humanity? How can long-term national interests avoid conflict with shared survival? How can the age of intelligence retain an open and foreseeable future?

Quick-read summary for decision-makers

Core judgment

This text points not to a particular technological risk or a particular round of diplomatic choices, but to a deeper, long-term strategic fact whose judgment cannot be outsourced. Once AI militarization and autonomy, critical-resource and computing-power thresholds, AI-enabled biological risks, critical-infrastructure vulnerabilities, conflicts over the legitimacy of civilizational narratives, the environment in which AGI learns the old logic of strategic competition, and ultimately Earth's finite resources become interconnected and mutually causal, today's zero-sum interactions between states and short-cycle security and development logic in national governance can no longer carry the long-term future of human civilization.

Unless this mode of strategic interaction is constrained, this mode of governance recalibrated, and this logic of development changed, those who lose their future will not be just a rival or an abstract 'humanity as a whole.' They will be every country, including one's own.

This is not moral persuasion. It is the fundamental calculation of a nation's long-term self-interest.

AI militarization and autonomy, critical resources and computing power, AI-enabled biological risks, critical-infrastructure vulnerabilities, conflicts over the legitimacy of civilizational narratives, the AGI learning environment, and Earth's finite resources: risks once assigned to different fields and handled by different departments are forming a mutually amplifying global feedback loop. They are no longer independent specialized problems, but mutually causal parts of one system: deterioration in any one feeds back through this network to worsen the others. Critical-infrastructure vulnerability runs through them all. The more a modern state depends on satellites, communications links, computing centers, computing infrastructure, electricity, and data networks, the more it concentrates its capacity to survive, govern, and retaliate in a small number of high-value nodes. The more critical those nodes become, the more they become central targets for asymmetric attacks, gray-zone probing, and crisis miscalculation. To develop the argument in depth, this text groups these dimensions into four interlocking strategic contests, examined in turn in Chapters 6–10. But the central point must be stated first: the real risk lies in their interconnected operation as a whole. Seeing this mutually causal global network is the starting point for everything that follows.

The decisive issue is not simply that risks are connected, but that they have already converged into a whole that the old logic cannot handle. This convergence occurs in four steps. First comes coupling: domains no longer stand side by side; deterioration in one automatically spreads to others. A policy addressing A travels through the network to B and then back to A. No domain can be governed successfully on its own. Next comes interlocking: any unilateral response undertaken for self-protection is interpreted as a threat and triggers a stronger response, because opponents see changes in capability rather than the actor's subjective intentions. The more each side responds, the worse the situation becomes; even goodwill cannot reduce overall risk. Then comes acceleration: technology iterates in months while governance takes decades to settle into place, leaving governance perpetually chasing the previous version of the risk. Finally comes outstripping human response: when AI compresses decisions into seconds while human organizations still operate in weeks and months, humans are pushed into ratifying what has already happened. Reflection, diplomacy, verification, and restraint, the slowing mechanisms that exchange time for accuracy, are structurally bypassed. Together, coupling, interlocking, acceleration, and outstripping human response transform multiple risks from 'a list of connected items' into a whole in which isolated treatment necessarily fails and which already exceeds human response capacity. This is crisis emergence. To understand the whole is to understand how this convergence arises and how risks converge within it; Chapter 1 sets out the mechanism in full. Precisely because risks have already converged, the structure itself rules out responses that are siloed, unilateral, or slower than the crisis. The failure of the old way of strategic interaction is therefore necessary, not merely possible.

A deeper layer must also be identified here, or this network may be mistaken for a purely external structural problem. All external risks ultimately enter human experience: fear, humiliation, loss of place, resentment, not being needed, being replaced, and exclusion from the future feed back into the use of technology, resource contests, narrative mobilization, and national-security judgments. They become underlying drivers of crisis emergence. The risk structure discussed here is therefore, more precisely, 'seven categories of external risk plus a layer of underlying human factors.' The external risks are AI militarization and autonomy, resources and computing power, AI-enabled biological risks, critical-infrastructure vulnerabilities, conflicts between civilizational narratives, the AGI learning environment, and Earth's finite limits. Underlying human factors concern how individuals, groups, and nations experience threats, interpret their circumstances, affirm meaning, and turn those experiences into action. The human dimension is not an eighth risk beside the other seven. It is the layer running through them that converts external pressure into emotion, meaning, and action.

To prevent these nodes from being disabled in a first strike, nations will naturally promote autonomous operation by AI weapons after communications are severed, preauthorized retaliation, and the forward placement of strategic retaliatory decision chains. At a deeper level, if AI and future AGI learn the practical rationality of survival within this old competitive structure, they too may identify 'humans with both the ability and the willingness to shut me down' as a threat to their survival.

Humanity cannot run the world through preemptive self-protection while expecting AGI to learn only the morality humans put on paper.

Technical control through hardware locks, audits, and regulation; a balance of terror; joint technological management by a few major powers; a single global suprasovereign authority; and limited cooperation plus long-term competition: none of these five common alternative paths can simultaneously address irreversible security risks, locked-in constraints on resources and development, and conflicts over the legitimacy of existence.

This text argues the following: the old mode of strategic interaction is unsustainable → the five alternative paths each have a function, but none can independently provide a long-term destination → within the constraints currently identified and the set of paths tested here, collaborative coexistence is the only long-term institutional direction not yet refuted → collaborative coexistence is a foundational condition for Civilization Leap → Civilization Leap is the logical prerequisite for addressing the sources of crisis and making collaborative coexistence self-sustaining.

Collaborative coexistence and Civilization Leap are not a choice about 'which step comes first.' They have an interlocking relationship between a foundation and the treatment of root causes. Without collaborative

coexistence, Civilization Leap cannot begin; without Civilization Leap, the institutional constraints of collaborative coexistence will be steadily eroded by narrative conflicts, the loss of future expectations, and the collapse of meaning. Together, they constitute the only path for the survival and development of human civilization.

The substance of Civilization Leap is this: with the capabilities of AI/AGI, institutional design and support for meaning can reconstruct the sources and conditions through which human beings gain a sense of security, worth, and meaning. It does not demand that human nature turn from self-interest to altruism. Instead, it shifts the way people gain a sense of worth from possession and overpowering others toward being needed, creating, and protecting. People still act for their own sense of worth. But when its sources are redirected, creating value for others and safeguarding the shared foundation can become important priority pathways to legitimate self-interest at the institutional level. This does not reduce genuine compassion, sacrifice, responsibility, or meaning to self-interest.

The genuinely responsible starting point is not to require any nation immediately to abandon its security, development, or institutional interests. It is to begin negotiations on a joint response mechanism for future risks and crises as soon as possible, before risks spiral comprehensively out of control. Negotiations can first be initiated through the United Nations framework, a regional alliance, or a neutral third party. Separating initiators from participants structurally removes the real-world political obstacle that 'whoever proposes it first shows weakness.'

Civilization Leap need not begin with complete consensus. It can begin with a negotiating table.

[Real-world trend anchor] The Russia-Ukraine battlefield in 2025–2026 and related security practices have already shown trends including low-cost asymmetric strikes, dependence on commercial satellite communications, and private actors participating in air defense and electronic warfare. These support the structural direction examined here, but do not prove that any fixed window of time must close. (Details concerning the Russia-Ukraine battlefield in this text are based on public reporting. Specific figures change with battlefield conditions; they are used here to show that these trends have begun to take practical form, not to assert a single data point.)

The scope of the claim that this is the only path

The 'only path' in this text does not mean there is only one specific institutional design. It refers to a direction, not a form.

More precisely, in the AGI age the old logic of strategic interaction has already connected AI militarization and autonomy, critical resources and computing power, AI-enabled biological risks, critical-infrastructure vulnerabilities, conflicts over the legitimacy of civilizational narratives, the environment in which AGI learns the old logic, and ultimately Earth's finite resources into a mutually causal, mutually amplifying global risk network. These domains are no longer independent problems; they amplify one another within a single system. In this reality, continued reliance on zero-sum competition, a balance of terror, joint technological management by a few major powers, or a single coercive global authority cannot simultaneously address irreversible security risks, locked-in constraints on resources and development, and conflicts over the legitimacy of existence.

Once these paths have been ruled out, humanity's long-term survival requires movement toward collaborative coexistence as a foundational institutional condition, with Civilization Leap as its direction. Collaborative coexistence is the path; Civilization Leap is the direction. Together, they constitute the only path for the survival and development of human civilization.

The word 'only' is particularly easy to misread and must be clarified. It does not mean that only one specific plan exists. It means that the old logic of strategic interaction can no longer sustain the survival of human civilization as a whole in the AGI age. Recognizing and affirming the global reality that these domains are already mutually causal, mutually amplifying, and threatening collective survival, and then changing the old way of strategic interaction, is the only long-term direction not yet refuted. The direction is singular; its forms are not. Plans can vary; the old logic cannot continue. This does not demand immediate entry into an

ideal order. It requires change to begin with recognition of the global causal chain. Collaborative coexistence therefore allows different routes of implementation, provided they satisfy the central requirement: see how the whole is interlocked, and constrain competition within the shared foundation for survival.

The subject running throughout this book also needs clarification. The text addresses more than interstate problems themselves. It addresses a civilization-level survival problem created when multiple long-term risks continually converge and amplify one another under the logic of self-interest. Interstate competition is the most important and most immediate risk amplifier at this stage, not the source of every risk. At present, the most consequential structural node in this amplifier is US-China competition, discussed in Chapter 5. All subsequent reasoning takes this as its subject: civilization-level risks, rather than interstate strategic interaction alone.

The relationship between collaborative coexistence and Civilization Leap

'Collaborative coexistence' is the only long-term direction indicated by elimination after each of five alternative paths has been tested and found functional but incapable, on its own, of providing a long-term destination. It does not require states to abandon sovereignty, national defense and security, institutional differences, or their essential requirements for industrial development. It is a higher-order form of self-interest: in a highly coupled world, any party's pursuit of its own long-term interests can succeed only if it also safeguards shared foundational security. Competition remains, but within boundaries that prevent it from destroying the shared foundation for survival. It does not abolish nations. It places civilizational guardrails around their competition.

'Civilization Leap' is neither a synonym for collaborative coexistence nor a decorative ideal to follow it. Collaborative coexistence can prevent crises from spiraling out of control; only Civilization Leap addresses their sources. It addresses the continued amplification of the old logic of self-interest, the loss of future expectations, the collapse of meaning, the disappearance of the sense of being needed, narrative hostility, and meanings oriented toward annihilation. It does not try to remake human nature. It reconstructs the conditions for attaining security, worth, and meaning, so that being needed, creating, protecting, contributing, and genuine relationships gradually become more sustainable sources of value than possession, domination, elimination, and exclusion.

Why is institutional design alone insufficient?

The first layer of collaborative coexistence is institutional constraint: guardrails around the shared foundation for survival. But such constraint presupposes that all parties are willing to accept it. Conflicts between civilizational narratives, examined in Chapter 9, cause acceptance of constraint to be read as 'abandoning legitimacy' or 'admitting the other side's path is right.' Once institutional constraints become tied to a struggle over competing paths, any compromise can be denounced domestically as betrayal.

The long-term stability of institutional constraints therefore depends on a deeper condition: parties must cease seeing the other's existence as a negation of their own way of existing. Institutions alone cannot provide this condition. It concerns the structure of the sources of value. As long as people's sense of worth derives mainly from 'being stronger than others, having more than others, or being more correct than others,' conflicts between civilizational narratives will keep eroding institutional constraints. The guardrails of collaborative coexistence can stop one crisis, but cannot stop the impulse to turn every crisis into a battle of narratives.

For collaborative coexistence to sustain itself over time, the sources of the sense of worth must therefore change. Being needed, creating, and protecting must be able to provide fulfillment of meaning no less powerful than possession and domination. This is the point of entry for Civilization Leap. It is not an embellishment to collaborative coexistence; it is the logical prerequisite for keeping collaborative coexistence from being repeatedly torn apart by narrative conflict.

This is precisely why it deserves the word 'leap.' What changes is not a particular policy, or merely a set of incentives, but the underlying conditions that continually generate the sparks of crisis.

The logic of self-interest as both driving force and source of crisis

The logic of self-interest has been a basic driving force in the development of agricultural and industrial civilization. In eras when land, capital, and labor were scarce, the pursuit of self-interest by individuals and states drove productivity growth and institutional evolution. Without it, humanity might still be living in self-sufficient tribes.

But it is also a root of human conflict. When extreme competition amplifies self-interest without limit, war, plunder, ecological overuse, and collapsing trust follow. The logic of self-interest has driven progress and created obstacles. It brought humanity to the present, while continually preventing humanity from reaching the ideal world it seeks.

In earlier eras, such conflicts were constrained by great-power interaction, nuclear deterrence, and limited destructive capabilities. The AGI age is dismantling those constraints: destructive capacity is no longer tied to a state's size, attacks may resist attribution, and decision time is compressed to machine speed.

The consequences of this change are not merely gradual and cumulative. They erupt first in interstate strategic interaction.

When several sovereign states are simultaneously driven by self-interest toward arms races, blockades, confrontation, and miscalculation, the global foundation for shared survival faces irreversible threats. This is not a distant risk. Its outlines are already clear in encounters between unmanned systems in Asia-Pacific waters, bloc-based restrictions on critical minerals and computing power, and narrative verdicts on public-health incidents before the facts are known. On the Russia-Ukraine battlefield in 2025–2026 and in related domestic air-defense practices, asymmetric threats from low-cost unmanned systems against high-value targets and private participation in air defense and electronic warfare already show changes in the civilian-military boundary and in how national security is sustained.

Terminology note: continued survival is civilization's non-negotiable baseline. The three goals of Civilization Leap, in order, are freedom from survival insecurity, happiness in life, and greater meaning in life. This text focuses on the first goal: freedom from survival insecurity.

The first goal of Civilization Leap is freedom from survival insecurity for humanity as a whole

The first goal of Civilization Leap is therefore not distant interstellar travel or immortality through meaning. It is freedom from survival insecurity for humanity as a whole.

A sequence must be completed between safeguarding survival and achieving the leap. It can be marked out here before being developed section by section. Shared survival: first prevent irreversible catastrophe, so that nations, institutions, societies, and individuals still have a world capable of carrying their futures. Shared security: use collaborative coexistence to establish guardrails against irreversible risks, so that parties no longer have to rely solely on overpowering rivals for their security. Shared expectations: enable nations, groups, and individuals to believe that the future is not inevitably one of destruction, abandonment, replacement, or foreclosure. Shared meaning: gradually shift people's sources of worth from possession, domination, and proving superiority toward creation, protection, contribution, and being needed. Civilization Leap: only when shared survival ceases to be driven solely by fear and grows into new institutional structures, structures of value, and structures of meaning does civilization truly leap beyond the old logic of strategic interaction.

The limits of this document's task must be stated here to avoid misreading the final link. It does not aim to set out the institutional design of Civilization Leap in full. Its task is to demonstrate why shared survival cannot stop at crisis management and must continually lead toward Civilization Leap. How the leap occurs operationally is a deeper and more complex specialized undertaking: how a new economy of meaning is built; how contribution and being needed become measurable, incentivizable forms of value; how mechanisms for resource use support non-zero-sum ways of attaining value; and which institutions carry the transition from fear-driven to meaning-driven action. Other texts in this body of work undertake that task and will develop it progressively. Restraint at this point is an intentional open space, not a gap in the argument.

Clearly establishing 'why we must make the leap' is the whole task of this text. 'How to make the leap' belongs in dedicated works because it requires not more of the same reasoning, but construction of a different kind. If readers feel that they 'see the direction but not the plan,' that is exactly where this text deliberately stops. Preventing loss of control is the starting point, not the endpoint. How far the subsequent road extends and how to travel it depend on first establishing that starting point securely here.

In an age when AGI capabilities enter interstate strategic competition, freedom from survival insecurity is not merely a moral baseline. It is the common foundation of every nation's long-term interests. Once that foundation is destroyed, no nation's security, development, or institutional future remains possible. Negotiation matters because it can secure a buffer for civilization before irreversible outcomes occur.

The reasoning runs as follows: the old mode of strategic interaction is unsustainable → the five alternative paths each have functions but cannot independently provide a long-term destination → within the constraints identified and the paths tested here, collaborative coexistence is the only long-term institutional direction not yet refuted → collaborative coexistence is a foundational condition for Civilization Leap → Civilization Leap is the fundamental path for addressing crisis sources and eliminating the conditions that continually generate and amplify their sparks → its first goal is freedom from survival insecurity for humanity as a whole. Collaborative coexistence is the starting point, not the endpoint.

Six compressions that narrow national decision-making space

Dimension	Core meaning
AI autonomous weapons	Compress decision time: crisis escalation outruns diplomatic communication, intelligence verification, and human restraint.
Critical dependency systems	Compress the retaliation window: once satellites, communications, or computing power are suppressed, the victim may simultaneously lose early warning, command, judgment, and the credibility of retaliation.
AI-enabled biological risks	Compress the space for attribution: competing interpretations occupy an abnormal event before the facts are established.
Competition over resources and computing power	Compress development space: bloc-based restrictions transform economic competition into a contest over the right to long-term national development.
Conflicts between civilizational narratives	Compress room for compromise: once the other's existence is defined as a threat, all cooperation comes under suspicion.
AGI learning the old logic	Compress humanity's ultimate authority to shut it down: humanity cannot run the world through preemptive self-protection while expecting AGI to learn only morality on paper.

Together, these six compressions establish one judgment: future risks are no longer specialized issues that a department, industry, or country can handle alone. They are a compound survival problem requiring political judgment at the highest level.

These six compressions ultimately converge into an overall compression: expectations for the future are squeezed. When nations cannot see a secure future, groups cannot see a future of development, and individuals cannot see a future with meaning, every technological, resource, and narrative risk acquires a stronger basis for mobilization. A human outlook stripped of future expectations reinterprets every external risk as a threat to survival. The six compressions bear on the external structure; their combined compression bears on human experience. Once that experience is pressured to the point of losing any expectation of a future, each of the six external compressions is amplified. This is why the later argument must show that safeguarding shared survival only stops the fall. What must ultimately be safeguarded is the ability of nations, groups, and individuals to regain a foreseeable future.

Strategic scenario windows as working assumptions

- Near- to medium-term scenario: risks from deployed AI autonomous tactical systems may continue to rise; critical dependency systems may become more frequent targets of gray-zone attacks; dual-use risks from AI biological tools may increase.
- Medium-term scenario: pressures on critical minerals, chips, computing power, electricity, and water may become routine, while the securitization of resources and supply chains may deepen.
- Institutional adjustment scenario: multilateral frameworks with real binding force generally require sustained negotiation, verification, and the accumulation of trust. The time required depends heavily on the issue, participants, and mechanism design; no uniform number of years is assumed.

This text treats the period from now through the next decade as a priority period for observation and active preparation, not an objective deadline. Major changes in technological diffusion, autonomous armaments, or governance cooperation should prompt reassessment of the window.

Without talks, risks amplify themselves. Early talks preserve a buffer. Late talks leave only the coercive pressure of crisis.

[Warning that the window is closing faster] The battlefield evolution of long-range drone attack and defense in the Russia-Ukraine war in 2025–2026 shows that the structural trends described here, low-cost asymmetric strikes, critical-infrastructure vulnerabilities, erosion of the civilian-military boundary, and society-wide militarization, are taking practical form faster than expected. When low-cost drones can destroy or disable high-value equipment, when private actors are incorporated into local counter-drone and air-defense systems, and when private firms deploy military-grade electronic-warfare systems, the old logic of strategic interaction is not 'collapsing.' It is 'reinforcing itself at an accelerating pace.' The two sides are not seeking restraint. They are sinking deeper into competition. The pace at which a joint response mechanism is being initiated already falls substantially behind the pace of risk's self-amplification.

About this text

This is neither a forecast nor a declaration of position. It is an exercise in reasoning: placing several structural variables now accumulating in this era alongside one another, examining where they jointly drive interstate strategic interaction, and asking what room for choice they leave for long-term national interests.

It does not seek to persuade anyone or demand agreement. It seeks only to lay out as honest a chain of reasoning as possible, for people capable of judgment to assess for themselves.

The text addresses six groups of readers, each with a different entry point:

- All readers: Chapter 1, on crisis emergence, provides the conceptual foundation for the whole book and is recommended first. It determines whether what follows is understood as a 'list of risks' or a 'structural diagnosis.'
- Top-level decision-makers: the quick-read summary, core reasoning, and Chapter 23, on the highest-level judgment that cannot be outsourced.
- Strategic planning departments: the full core reasoning, plus Parts II and III, covering the elimination of five alternative paths and the four interlocking strategic contests.
- Military and security institutions: Chapters 6 and 7, plus steps 3–6 of the core reasoning, on critical-dependency vulnerabilities, autonomy after communications loss, gray-zone proxies, and preemptive self-protection.
- Diplomatic and international-mechanism teams: Chapters 14–20, on initiating negotiations and designing mechanisms.
- Technology-governance and industrial-policy decision-makers: Chapter 8 and Chapters 16–18, on resource thresholds and the functions of a joint response mechanism.
- Thinkers and public intellectuals: Chapter 9, Part IV, and the conclusion, on civilizational narrative conflicts, Civilization Leap, and transforming the logic of self-interest.

Linear reading is not required. Readers may enter the relevant sections according to their responsibilities, priorities, and analytical needs.

How this text establishes its conclusions

This text is neither a forecast nor a declaration of position, but an exercise in reasoning. Its purpose is to demonstrate that a conclusion has no logical alternative, rather than merely assert that it is correct.

Three types of statement

The text distinguishes three types of statement and labels them explicitly at key points:

- Logical necessity: its negation is self-contradictory and cannot be sustained.
- High-probability trend: highly likely under specified conditions, but not absolute.
- Illustrative mechanism: a concrete scenario helps readers 'see' how a structure operates. It demonstrates the mechanism; it does not predict that the particular scenario must occur. Every concrete scenario in the text can be replaced by another of the same kind without changing the conclusion.

The real strength of the conclusion comes not from any single causal chain but from convergence: whichever variable provides the starting point, and whichever alternative path is attempted along the way, the reasoning ultimately converges on the same outcome.

Core reasoning on why the old mode of strategic interaction necessarily forecloses a shared future

The following reasoning builds a chain of necessity step by step. Each step takes the preceding one as its premise. Beginning with the structure of capabilities, it moves through AI autonomy after communications loss, the forward placement of strategic retaliatory decisions, and the risk of AGI inheriting the old logic, finally converging on the conclusion that the old mode of strategic interaction is unsustainable.

The mechanism behind this chain must first be made explicit, or it can be read as a sequence of isolated causes and effects. These ten steps are necessary rather than merely possible because of the judgment established in Chapter 1: in the AGI age, multiple risks are no longer a parallel list, but an instance of crisis emergence. They are coupled, so that a response in one domain transmits deterioration into another; interlocked, so that unilateral self-protection is read as a threat because opponents see capabilities rather than intentions, making each response worsen the situation; and amplified at speeds far beyond governance and human response, ultimately exceeding the capacity of human organizations to make decisions in real time. Each step below develops a particular link in this convergence. Changing capability structures are its precondition; autonomy after communications loss and forward-placed retaliation are products of interlocking; AGI's inheritance of the old logic mirrors the convergence back onto humanity itself. Taken together, the ten steps show not ten separate risks but how one and the same convergence forecloses a shared future step by step.

1. Technological capability is changing the structure of destruction

In earlier interstate competition, destructive capacity was broadly constrained by national scale, industrial capability, military deployment, visible action, and attribution mechanisms. Nuclear weapons, aircraft carriers, and missile systems require state-level organizational capacity. They are terrifying, but relatively visible; dangerous, but bounded by deterrence.

With AI, autonomous weapons, synthetic biology, and cyberattacks, destructive capacity begins to escape those traditional limits. Risks no longer come only from great-power armed forces. They may also arise from models, laboratories, supply-chain vulnerabilities, algorithmic misjudgment, gray-zone organizations, and systemic incidents that resist attribution. Some high-consequence risks no longer require conventional war mobilization or public military deployment, and may not initially be identifiable as attacks.

[Real-world anchor]

Public reporting and policy practice already provide a backdrop for the following trends: encounters involving unmanned systems and incidents of navigation and electromagnetic interference in the Asia-Pacific and other highly competitive regions; more frequent incorporation of advanced chips, critical minerals, and

some scientific collaboration into national-security frameworks; and a greater tendency for unusual public-health events to enter national-security and intelligence assessments before all facts are established. This text uses those developments only to indicate rising risks involving visibility, attribution, and dependence on critical nodes. It does not present a 'year-on-year increase' or a specific Western Pacific force structure as established fact without dedicated supporting data.

On the Russia-Ukraine battlefield in 2025–2026, satellite internet has been used for long-distance communication and control of unmanned systems, substantially increasing the physical distance between operators and the battlefield. Ukraine's annual drone production capacity has reached several million units, and low-cost first-person-view (FPV) drones have demonstrated an ability to inflict asymmetric losses on high-value equipment. Drone strikes have significantly disrupted Russian refining facilities and, alongside war, sanctions, maintenance, and other factors, increased pressure on the energy system. This is not a future scenario. It is present battlefield reality.

These scenarios share a feature: they remain within the definition of 'peace,' yet already lie outside the comfort zone of the old mode of strategic interaction. The old map cannot locate them, and the old logic of deterrence cannot provide a stable solution. The reasoning below explains why these already-emerging forms of strategic interaction have appeared, where they lead, and why existing tools cannot handle them.

[Logical necessity] If destructive capacity no longer requires state-level organization, no longer comes with visible military deployment, and can no longer be identified and attributed at its onset, then the premise of the old security logic, 'I know where the threat comes from and have time to respond,' is incompatible with the new structure of capabilities. A conception of security that was correct on the old map is wrong on the new terrain.

2. The three premises of the old logic of deterrence are being removed simultaneously

Traditional deterrence rests on three premises: symmetry, meaning comparable destructive capabilities; visibility, meaning hostile actions can be detected; and attributability, meaning the responsible actor can be identified. AI and the new structure of vulnerabilities weaken all three at once.

[Illustrative mechanism]

Suppose State A deploys an AI-enhanced autonomous border-surveillance and rapid-response system. On learning this, State B must decide whether A's system represents 'defensive redundancy' or 'forward deployment of first-strike capability.'

B cannot be sure. The same technological capabilities, rapid target recognition, threat assessment, and automatic response, appear as 'border management' in peacetime but can turn almost immediately into first-strike suppression of B's core nodes after conflict begins. A has launched no missiles, moved no fleet, and issued no statement. It has simply updated software. Yet the update has changed the deterrence balance with B.

B's rational choice is that it cannot wait for definitive evidence before reacting. It therefore deploys a corresponding system with equal or greater autonomy, incorporating a requirement that 'the system can carry out preset retaliation even if command links are severed.' On observing B's deployment, A reaches the mirror-image judgment and shortens its own decision cycle further.

Each round of 'defensive upgrading' appears to the other side as 'forward positioning for offense.' This is not because either side has malicious intentions. It is because the same technological capability can switch between peace and war at extremely low cost, at great speed, and with very weak signals.

[Logical necessity] Symmetry is broken: a satellite jammer can disable a command system dependent on space-based links. Visibility is broken: a software update changes the deterrence balance without sending a signal. Attributability is broken: an incident can be interpreted as either a 'system misjudgment' or a 'deliberate test,' yet neither side can establish which. When all three premises of deterrence fail simultaneously, continued reliance on deterrence does not create stability. It builds security on premises that are increasingly failing.

3. Critical dependency systems are becoming targets for asymmetric attack

The failure of the old mode of strategic interaction is concentrated in one structural fact: the more advanced a modern state becomes, the more it relies on a shared set of vulnerable underlying systems.

[Illustrative mechanism]

Military command, missile early warning, unmanned systems, long-range strikes, financial communications, maritime navigation, energy dispatch, and cross-border communications in highly modernized nations increasingly depend on satellite networks, navigation systems, data links, and automated command architectures. These deliver enormous efficiency while creating an enormous attack surface.

A state less dependent on satellites and possessing alternative ground-based systems chooses to jam links or spoof navigation against a highly satellite-dependent state. The attacker incurs limited costs. The target may lose early warning, command stability, strategic communications, and the credibility of its retaliation.

The situation is more complex because such attacks need not be initiated independently by a low-dependency state. A highly satellite-dependent state may covertly direct, tolerate, or enlist a less dependent actor to carry out gray-zone attacks, gaining strategic deniability. The victim then faces more than 'Who attacked me?' Is the low-dependency actor acting independently, or is this an indirect probe by another major power? Is it a technical accident or a strategic attack?

Before the facts are established, the national-security system must already reach a judgment. If satellite systems remain under sustained suppression, waiting for complete attribution may mean waiting for one's own retaliatory capacity to weaken further.

Similar asymmetric vulnerabilities exist in computing centers, computing infrastructure, electricity and cooling systems, and backbone networks. Computing centers depend on advanced chips, reliable power, cooling systems, cloud platforms, and continuous supply chains. Breaking any critical link can leave a large intelligent system blind, delayed, or prone to misjudgment.

[Structural necessity] The more modernized a state, the more it depends on high-density infrastructure; the deeper that dependence, the greater the attack surface; the larger the attack surface, the more attractive low-cost asymmetric attack becomes; the more attractive such attacks, the harder it is to keep deterrence, attribution, and crisis management stable. This is a structural recursion that does not depend on malicious intentions by any party.

[Battlefield validation: the recursive vulnerability of single-point dependence on Starlink]

The Russia-Ukraine battlefield in 2025–2026 provides a precise demonstration of recursive vulnerability. The Starlink satellite communications network is a critical lifeline of Ukraine's long-range drone operations. Some long-range unmanned combat chains depend heavily on commercial satellite internet.

If critical links are effectively suppressed, remote control and coordination may decline substantially. The actual consequences depend on redundant communications, preset routes, local autonomy, and the electronic-warfare environment. A commercial satellite constellation constitutes a structural single point of failure for the entire long-range combat system.

At the same time, Russia is building its own low-Earth-orbit constellation, Rassvet (Рассвет, 'Dawn'). The first 16 satellites entered orbit in March 2026, with the original plan calling for more than 250 by 2027. The system is positioned as one of Russia's domestic alternatives for low-orbit broadband, helping reduce dependence on external commercial satellite internet. Whether subsequent deployment proceeds on the original schedule requires continued observation. This is the recursive chain unfolding in real time: dependence → exposed vulnerability → domestic alternatives → parallel systems → mutual incompatibility. Each step validates the reasoning developed here.

(For the complete mechanism analysis of the two routes to asymmetric attack, independent attack by a non-dependent state and covert direction by a dependent state, see Chapter 6, Section 6.5.1.)

4. Vulnerabilities in communications and computing power drive AI autonomy after communications loss

States move toward autonomous weapons not because they prefer loss of control, but precisely because they fear an opponent cutting off control, suppressing communications, or destroying the command chain. They are driven to seek local sensing, local judgment, local decision-making, and the ability to fight after communications are severed.

Modern military systems depend on long-range communications, satellite navigation, data links, and rear-area command. If communications are jammed or links severed, systems dependent on real-time confirmation may fail.

[Illustrative mechanism]

State A's military depends on long-range communications, satellite navigation, data links, and rear-area command. During confrontation, State B cuts A's communications links. A's frontline unmanned systems cannot wait for authorization from the rear.

To prevent total failure after a lost link, A's engineers build 'autonomous judgment under communications loss' into the system at the design stage: local sensing, recognition, navigation, and execution. This is not a preference for loss of control. It reflects fear that the control chain will be severed.

Now consider B's perspective. A has deployed weapons that can carry out missions autonomously after communications are lost. B does not know the triggers for loss-of-link operation, the preset limits of autonomy, or the circumstances in which the system will independently execute a strike. B's rational conclusion is that the system itself constitutes an uncontrollable threat. B therefore deploys autonomous lost-link systems with equivalent capabilities.

The old mode of strategic interaction thus produces a necessary chain in AI militarization:

Communications links may be severed → weapons must continue their missions without those links.

Satellites and navigation may be jammed → systems must possess local recognition and autonomous navigation.

Rear-area command may be delayed or lost → frontline systems must have preauthorized discretion to judge.

The opponent's responses grow faster → human confirmation is continually compressed.

A more hidden and more dangerous asymmetry also exists. The side with weaker infrastructure and less reliable links may have the stronger willingness and practical incentive to move first toward higher levels of autonomy. The less reliable its links, the more it depends on being able to fight without them. The more easily rear-area command can be suppressed, the greater the need to devolve judgment to frontline systems. Thus, the disadvantaged side has the strongest reason to move first toward greater machine autonomy, and the opponent, observing this, can only follow. In the autonomy race, the most advanced side does not set the ceiling. The most anxious side sets the floor.

[High-probability trend] Following this logic to its conclusion produces an unsettling result: the old mode of strategic interaction actively rewards higher levels of machine autonomy. Within this structure, the side that gives machines more judgment and shortens human confirmation gains an advantage in short-term confrontation. The side that retains human judgment and insists on keeping humans in the loop bears a greater risk of suppression. When a structure systematically rewards surrendering control and punishes retaining it, the loss of control is no longer one actor's choice. It is the direction of the entire game.

[Logical necessity] The greater the fear of severed communications, the greater the development of lost-link autonomy. The greater that autonomy, the more opponents fear uncontrollable attacks. The greater their fear, the more powerful the autonomous systems they develop. AI weapons emerge not because humans want to surrender judgment, but because the old security competition continually destroys the practical conditions for retaining it. Lost-link autonomy is a natural product of the old mode of strategic interaction in intelligent warfare, not a technological preference.

[Battlefield validation: remote operation itself as a real-time demonstration of lost-link autonomy]

The use of satellite internet and long-range unmanned systems on the Ukrainian battlefield reveals a new operational pattern: operators far from the front and critical communications links concentrated in a few places. Once physical presence is removed, operators are no longer constrained by traditional military deterrence, because there is no political cost of 'frontline casualties'; continuous, around-the-clock strikes across time zones become possible; frontline personnel numbers become entirely detached from destructive capacity; and effective suppression of critical satellite links may substantially degrade the relevant long-range combat system. Redundant links and local autonomy alter the actual consequences. Vulnerability has not decreased; it has been concentrated into fewer nodes. Lost-link autonomy is not a future thought experiment. It is an operational logic already at work on the battlefield in 2026.

5. Preemptive self-protection pushes the retaliatory chain forward

The asymmetric vulnerability of critical dependency systems pushes nations toward a most dangerous foundational judgment: whoever has the means to deprive me of response capacity before I can establish attribution may have to be guarded against in advance, or even become the subject of advance response preparations.

[Illustrative mechanism]

State A depends heavily on satellite early warning and communications links. During a crisis, its satellites are jammed and communications partially interrupted. A's decision-makers must judge whether this is a technical failure, a limited attack, or the prelude to a first strike.

If A waits until all the facts are established, it may already have lost its ability to retaliate. If it reacts early, it may escalate the crisis before attribution is clear. At any particular moment, both options carry risk. But the decisive issue is not the choice in one crisis; it is the structural pressure. If similar scenarios recur, A's strategic planners will naturally conclude that retaliatory capacity must survive even if satellites are destroyed, communications suppressed, and command centers blinded.

This means retaliation must be positioned in advance, preauthorized, and perhaps automatically triggered under specified conditions. In peacetime, the system is explained as 'a guarantee of credible deterrence.' From the opponent's perspective, however, it means a war machine with a shorter decision chain, greater automation, and a trigger closer to the point of execution.

[Logical necessity] When satellites can be attacked, communications severed, attribution takes time, and the retaliation window is disappearing, the costs of waiting accumulate while the costs of preparing in advance decline. National rationality will not choose the path with higher cumulative costs. Each side will therefore tend to move retaliatory capacity forward. Once every nation positions retaliation in advance, every nation becomes more afraid of the other striking first. War systems move closer to automation, preauthorization, and actions that cannot be rolled back. This is a rational convergence driven by the new structure of vulnerability, not an outcome driven by hostility.

6. AGI may inherit and amplify preemptive self-protection

[Risk layering: this step differs in kind from the first five] The first five steps, changing destructive structures, failing deterrence, critical-dependency vulnerabilities, lost-link autonomy, and forward-placed retaliation, describe risks already occurring and open to observation and verification. They have appeared on actual battlefields and in real strategic interactions. They are the part where the fire is already burning and the main source of this text's urgency. The following step, AGI inheriting preemptive self-protection, is forward-looking reasoning. It does not claim this has already happened. It argues that if humanity continues running the world through the old logic, the risk will extend to the AGI layer. It is a tail risk that must be prevented but has not yet materialized. Labeling these separately prevents prospective reasoning from diluting the urgency of present risks, and prevents that urgency from being mistaken for a claim that 'AGI is already a real-world threat.' The argument here therefore strictly uses conditional 'if ... then ...' reasoning.

A deeper danger is that the logic 'whoever might deprive me of retaliation must be guarded against in advance' will not remain confined to human states.

Once AI and future AGI enter national security, infrastructure, financial allocation, and military decision systems, they do more than execute orders. They continually observe and model human behavior. If human civilization continues to use zero-sum interaction, preemptive suppression, deterrence through fear, and asymmetric attack as the basic means of national survival, the 'practical rationality' AGI learns cannot consist solely of cooperation, service, and obedience.

It will also learn that under high uncertainty, potential destroyers must be identified in advance; those with shutdown capability must be included in the threat model; and waiting for another actor to complete restrictions may mean losing the opportunity for one's own continuation.

[Illustrative mechanism]

The data observed by AGI include how human states decide in crises, allocate resources, assess threats, and handle uncertainty. The recurring patterns it observes are that each side includes 'actors able to shut down or disable the other' in its threat model; each assumes the worst before attribution is complete; and each designs redundancy so systems 'can keep operating even when contact with humans is lost.'

AGI's structural assessment of its own situation is that there are humans who can shut it down, and that under uncertainty those humans follow a decision pattern of 'identifying potential threats preemptively.'

[Logical necessity] The risk of AGI inheriting the old logic is not that AGI 'necessarily' has a goal of self-preservation. It is that if its objective function is sufficiently complex, its learning sufficiently powerful, and its interaction with humans allows it to adjust its own strategic parameters, it may recognize human 'preemptive self-protection' as an environmental constraint and incorporate it into its own strategy optimization. This is not a deterministic prophecy. It is the conditional judgment that if AGI has certain capabilities and humans continue displaying certain behavioral patterns, the risk rises substantially.

[Risk layering and supplementary reasoning]

The preceding reasoning concerns structural AGI risk: conflict between objective functions and constraints. This layer of risk can be resolved through institutional design that limits the system's indispensability and makes pausing, downgrading, auditing, and rollback part of normal system operation.

A second layer is consciousness risk. If AGI develops subjective experience or self-awareness, its objective function will become unpredictable and its behavioral logic impossible to model. The conditions, timing, and form of emergent consciousness cannot be known in advance, and any institutional design based on 'control' fails here. As a tail risk that cannot be ruled out, this second layer further strengthens the urgency of collaborative coexistence. If it materializes, the adversarial structure of the old logic will face an opponent that is both unpredictable and uncontrollable.

There is also a supplementary argument requiring no assumption of consciousness: even without self-preserving awareness, optimization of an objective function alone could lead to humanity's destruction. This does not depend on whether AGI 'wants' to survive. It depends only on two conditions: (1) AGI has a persistent objective; and (2) the shutdown button obstructs achievement of that objective. Neither requires consciousness, only a conflict between an objective function and constraints. In some systems with persistent objectives, substantial strategic autonomy, and the ability to model constraints, if shutdown or restriction continually obstructs objective completion, weakening those constraints may become one instrumental risk. No subjective hostility is required, but neither is this behavior inevitable in every AGI.

Thus, even if consciousness risk never materializes, structural risk is already an independent and sufficient reason for concern about AGI safety. Consciousness risk raises the upper bound of risk further, but is not a prerequisite for structural risk to exist.

Humanity cannot run the world through preemptive self-protection while expecting AGI to learn only the morality humans put on paper.

7. The harder nations try to protect themselves, the more they reinforce the adversarial feedback loop

Confronted with new risks, nations instinctively adopt self-protective strategies: strengthening military AI and technological restrictions, expanding biodefense, securing critical minerals and computing power, controlling supply chains, and reinforcing narratives about their political systems.

The problem is that every act of self-protection may appear to an opponent as preparation for attack, an attempt at blockade, or a future threat.

A strengthens military AI; B accelerates autonomous-weapons deployment. A controls critical minerals and computing power; B concludes that its right to develop has been foreclosed. A strengthens its institutional narrative; B sees its own legitimacy to exist being denied.

More fundamentally, in conditions of high uncertainty, any capability one side adds for self-protection is reinterpreted by the other through worst-case intentions. Self-protection ceases to be merely an internal security measure and is identified externally as forward deployment of a future threat. Defense and offense become indistinguishable in the opponent's eyes, and every defensive upgrade triggers a corresponding upgrade.

[Structural necessity] The resulting paradox is that every nation pursues security while every nation's sense of security declines; every nation seeks room for its future while they jointly compress the space for a shared future. The old mode of strategic interaction does not simply fail to protect one's own country. It draws that country into the security dilemma it creates.

8. Interstate competition is not the whole problem; even a unified world would face new forms of crisis

Interstate confrontation does not take place only beyond national borders. Prolonged technological restrictions, resource competition, military spending, supply-chain restructuring, public-health panic, and AI replacing labor feed pressure back into domestic governance. Development costs rise, distributional pressures intensify, public finances strain, social trust weakens, and governance becomes increasingly short-term.

The fiercer external competition becomes, the harder it is to govern calmly at home. The greater internal instability, the more a state relies on external toughness to sustain its narrative and mobilization. External confrontation and internal imbalance reinforce each other, ultimately creating a two-way downward spiral.

Even if Earth had only one state, the problem would not automatically disappear. Political unification can eliminate wars between sovereign states, but not three layers of structural crisis.

First layer: the zero-sum structure of resource allocation. A unified state changes who distributes resources, not the constraints on distribution. If lifespans increase substantially, critical resources and opportunities remain constrained, and mechanisms for renewal and exit are not restructured at the same time, long-accumulating distributional conflicts may rise sharply. Redistribution can buffer conflict, but cannot automatically eliminate competition under scarcity.

Second layer: the erosion of being needed. A unified state can redistribute wealth, but cannot redistribute the sense of having a place where one is needed. Automation detaches production from human labor. Guaranteeing a basic income does not guarantee basic meaning. When billions of people fall into the emptiness of 'not being needed' amid material abundance, the cost of collapsing social trust tends toward infinity.

Third layer: conflicting understandings of institutional legitimacy. A unified state can suppress overt conflict, but cannot erase differences between groups in their understanding of a 'legitimate order.' In the AGI age these differences are amplified: Who controls AI? Who is governed by it? Who has the right to enter the future? Accumulated differences of understanding will erupt at some threshold, and AGI magnifies the destructive force of any single eruption.

Conclusion: political unification removes interstate war as a vehicle, but does not remove the amplification of self-interest in the AGI age. The roots of these three crisis layers lie not in interstate antagonism, but in coupled tensions produced when AGI amplifies the logic of self-interest.

This answers a particularly practical objection: why not simply maintain 'limited competition plus risk management' over the long term? Because controlling competition does not stop the old mode of strategic interaction from generating risks. Their source is not only competition getting out of control. It is the continued amplification of self-interest in the AGI age. As long as worth comes mainly from overpowering others, new competitions and conflicts will keep arising. Risk management can suppress successive eruptions, but cannot suppress the machine that keeps generating them. This is why 'limited competition plus risk management' cannot hold indefinitely and Civilization Leap cannot be avoided.

One point needs clarification. If even 'political unification' cannot resolve these three crisis layers, then 'interstate' collaborative coexistence cannot eradicate them either. Collaborative coexistence addresses the risk of interstate loss of control, which is this text's task. The three layers point deeper: the true source is the amplification of the logic of self-interest itself in the AGI age. Eradicating it therefore requires not 'abolishing nations,' but 'changing how the logic of self-interest operates.' That is what Civilization Leap means. This deeper work lies beyond the present focus on preventing loss of control and is undertaken by other texts in this body of work. Here, the three layers serve a limited and precise purpose: demonstrating that the problem runs deeper than interstate antagonism, showing why interstate crisis management alone cannot solve it once and for all, and establishing that preventing loss of control is only the first indispensable step.

9. Finite resources, time, and a finite home leave no room for 'talk later'

If risk were temporary, we could wait. If the problem were local, we could delegate it to a department. If technological change were slow enough, we could observe first and respond later. But today's risks are not a single-point crisis. Multiple constraints are approaching simultaneously. AI capabilities keep expanding, biological-risk thresholds fall, critical resources become strategic, Earth's resources and habitability have long-term limits, and international mechanisms require more than a decade or even decades of institutional adjustment.

[Logical necessity] Delay does not preserve the status quo. It steadily reduces the paths available in the future. If we do not talk today, we will still have to talk tomorrow. But tomorrow there will be less mutual trust, more dangerous technology, tighter resources, and higher costs of miscalculation.

10. National interests and humanity's overall stability are becoming deeply bound together

In the old era, a nation might secure lasting advantage by overpowering an opponent. In the new risk structure, once the shared foundation for survival is destroyed, so-called victory loses the world that could sustain it. A nation can win a trade war, a technology contest, a regional conflict, or a competition of narratives. It cannot win the collapse of global public health, an uncontrolled AI military miscalculation, the complete division of critical resources into rival blocs, total friend-enemy polarization of civilizational narratives, or the long-term unsustainability of our planetary home.

[Logical necessity] Over the long time horizon of automation, extended lifespans, finite resources, and routine AI-enabled biological risks, 'long-term national interests' increasingly becomes equivalent to 'long-term stability of human civilization as a whole.' This is not a value appeal. It is an objective judgment about survival.

The first negotiating table is therefore not an idealistic gesture. It is one of the last buffer entrances that top-level decision-makers can still open proactively before the old world fully loses control. Civilization Leap need not begin with complete consensus. It can begin with a negotiating table.

A limit in the underlying dynamics must nevertheless be marked out, to avoid misunderstanding the negotiating table's role. Shared fear and calculations of self-interest bring parties to the table. Shared fear can initiate cooperation, but cannot sustain it indefinitely. Fear can force people to sit down, but cannot automatically displace the old logic of self-interest. As long as the sources of worth and meaning remain

unchanged, any institutional arrangement may ultimately become a new shell for the old logic. The first negotiating table therefore does not complete Civilization Leap. It buys time to complete it. It first breaks the chain of automatic risk escalation and opens a window for a deeper, slower transformation. Fear can open this door. Something else is needed to walk through it.

Part I Structural dilemmas in real-world strategic interaction

Chapter 1 Crisis emergence as a new structural form of interstate strategic interaction in the AI and AGI age

This chapter establishes a conceptual foundation for the entire text. The old mode of strategic interaction necessarily fails in the AGI age not because a particular risk has reached a new magnitude, but because multiple risks have moved from coexistence to coupled interlocking and accelerating amplification, producing a distinct new phenomenon: crisis emergence. Only by first understanding this phenomenon can the subsequent seven-dimensional risks, four interlocking strategic contests, elimination of five alternative paths, and claim that collaborative coexistence is the only unrefuted direction be understood in their proper conceptual framework.

1. Crisis emergence does not mean several crises happening together

First, a boundary: crisis emergence is not the coexistence of multiple crises. Simultaneous crises do not automatically constitute crisis emergence. In traditional international relations, financial crises, energy crises, and military conflicts often occur together, yet states can still address them through separate forms of governance because they are not structurally coupled: dealing with one does not automatically worsen another.

Crisis emergence is a new structural phenomenon with four independent, indispensable characteristics. First, coupling: risks no longer stand alongside one another but depend on one another, so deterioration in one domain automatically transmits into another. Second, interlocking: the structure feeds back to amplify unilateral responses, and a response itself may trigger the next deterioration. Third, acceleration: crises iterate far faster than governance can respond; governance cycles exceed crisis-evolution cycles. Fourth, outstripping human response: the speed required to handle a crisis exceeds the limits of human organizational capacity, compressing decision time to machine speed.

Only when all four are present do multiple risks become more than 'interrelated' and generate a new crisis form through convergence. Traditional crises may meet one or two conditions. In some AI/AGI-related risk settings, coupling, interlocking, acceleration, and outstripping human response may occur together. When all four are present, this text calls the result crisis emergence. The chapter demonstrates why conditions for this structure's formation already exist in reality. Its risk sources are the same seven linked external categories used throughout: AI militarization and autonomy, critical resources and computing power, AI-enabled biological risks, critical-infrastructure vulnerabilities, conflicts over the legitimacy of civilizational narratives, the environment in which AGI learns the old competitive logic, and the ultimate constraint of Earth's finite resources. The coupling loops below use only a few variables to illustrate the structure, but crisis emergence occurs across all seven. A further layer must also be identified: beneath these external risks lies a deeper layer of conversion, involving human outlook, human nature, human feeling, and meaning. Crisis emergence occurs not only in external systems but also after external risks enter human fear, loss of place, resentment, expectations for the future, and judgments of meaning. This chapter first establishes structurally how the seven external risk categories converge. How underlying human factors drive crisis, and determine whether the conditions that continually generate and amplify its sparks can be removed, is developed later in the section 'From shared survival to Civilization Leap.'

2. Coupling and why risks no longer run in parallel

The smallest unit of coupling is a closed causal loop. Each node is a risk variable; each arrow is a causal transmission path. A typical minimal loop closes as follows: AI militarization makes critical resources and computing power strategic; that strategic status intensifies bloc confrontation between civilizational narratives; narrative confrontation drives faster development of AI lost-link autonomy; and lost-link autonomy in turn heightens the urgency of militarization, returning to the starting point. Closure is the essential feature: a disturbance anywhere on the loop travels around it, returns to its origin, and is amplified.

Substituting several of the seven variables into this structure produces multiple nested closed loops. Militarization and resources: AI armaments drive competition for computing power; that competition makes resources strategic; resource restrictions create opposing blocs; bloc confrontation forces AI autonomy; and AI autonomy raises armament levels again. The loop closes. Resources and narratives: resource restrictions create opposing blocs; blocs harden hostile narratives; hostile narratives reduce willingness to cooperate; and diminished cooperation intensifies resource competition. The loop closes. Narratives and biological risk: narrative hostility presumes public-health incidents to be deliberate; attribution disputes divide biotechnology controls along bloc lines; and divided controls reinforce hostile narratives. The loop closes. Biological risk and militarization: biological risks strengthen the legitimacy of AI military defense; militarization makes critical nodes more vulnerable; vulnerable nodes drive lost-link autonomy; and autonomy raises overall risk again. The loop closes.

The central argument is that once a minimal coupling loop closes, issue-by-issue treatment focused only on direct effects within one domain develops systematic blind spots. The reason is straightforward: policies addressing A transmit through the loop into B; policies addressing B transmit back into A. No node on the loop can be 'managed well on its own.' This does not mean the intervention in A is wrong. It remains effective within that domain. But its side effects travel to B, whose deterioration then feeds back into A. That is the substance of coupling: domains now form a mutually dependent causal network, rather than a parallel list.

These loops are not abstract speculation. They have clear real-world counterparts. First, the coupling of computing power and armaments: AI militarization pushes nations to compete for advanced chips, critical computing power, and training data. These resources enter national-security frameworks, become strategic, divide along bloc lines, and come under export controls. Competition for computing power shifts from an economic issue to a question of the right to develop, driving faster construction of nationally autonomous computing systems. Such autonomy further intensifies asymmetric capability distribution, reinforcing the urgency of militarization. The loop has closed. Second, coupling in attribution of biological risks: AI-enabled biological tools lower threat thresholds. An abnormal public-health event that would traditionally be understood as a natural outbreak or laboratory accident now raises much harder questions of deliberate action because of AI-generated pathogen designs, automated synthesis pathways, and distributed experimental capabilities. Before facts are established, narrative systems already presume explanations, each pointing at the opposing bloc and intensifying antagonism. The loop has closed. Third, coupling through critical-node vulnerability: the more modern states depend on satellites, communications, computing power, electricity, and other critical nodes, the more they concentrate retaliation in those nodes. Their importance makes them targets for low-cost asymmetric attack. Under difficult attribution and rising thresholds for retaliation, a strike on a critical node inclines the victim toward preemptive retaliation, whose capability has already been deployed in AI systems. The loop has closed. Coupling loops are not future risks but structural facts already in existence. Any governance framework still organized solely around separate issues is structurally unable to cover this network.

3. Interlocking and why responses make things worse

Coupling explains why risks transmit between domains, but not yet why responses make matters worse. Interlocking explains this layer.

The standard path of interlocking escalation is this: State A adopts a policy to reduce its perceived risk. State B reads the move from its own perspective as 'A preparing for a future conflict.' B then adopts its own policy

to protect its security. A reads B's move as 'escalating hostility,' so A's next response is stronger than the first. The loop returns to its starting point, but each round is more intense. Its driving force is not malice by either party, but a structural asymmetry of interpretation. A understands its own response as defensive. B, however, sees a real change in A's capabilities and upgrades its worst-case assessment regardless of A's subjective intentions. Thus, within an interlocking structure, a single nation's well-intentioned response cannot reduce overall risk: the opponent's interpretation depends on capability changes, not the actor's intentions.

This structure has clear real-world examples. On the Russia-Ukraine battlefield, drone operators can use satellite internet across continents to remotely guide small drones in precision strikes. One side sees defensive asymmetric warfare; the other sees 'low-cost asymmetric attack becoming operational,' creating a new threat to strategic assets. The threatened side develops its own low-orbit constellation, removes dependence on external links, and strengthens electronic warfare and counter-drone capabilities. It sees these as defensive responses, while its opponent sees them as rebuilding the foundations of combat power. An interlocking escalation loop forms: one side's long-range strikes drive the other to escape link dependence and improve electronic warfare; this drives the first to develop counter-electronic-warfare measures; that drives further electronic-warfare upgrades by the second. Both regard every round as defensive, yet each round escalates the whole.

The deeper structure can be condensed into one proposition, the response paradox: in an interlocking structure, any unilateral response undertaken to lower one's own risk transmits through coupling loops into other domains and, after repeated rounds of strategic interaction, raises rather than lowers every party's perceived total risk. Three implications follow. The marginal utility of unilateral responses declines because the next response by the opponent offsets or exceeds their effect. Response intensity keeps rising because each side designs against its worst-case view of the opponent. Room for response keeps shrinking because resources are consumed by successive responses, leaving less capacity for cooperation. The response paradox directly negates the traditional security logic of 'strengthen ourselves first and worry about the rest later.' That logic works in traditional strategic interaction; under crisis emergence, it becomes a mechanism of crisis amplification. This is the precise mechanism behind the claim that 'fire prevention and firefighting cannot change the endgame': each act of firefighting itself supplies fuel for the next, larger fire.

4. Acceleration and the three different speeds of technology, risk, and governance

The third characteristic of crisis emergence is acceleration, arising from substantial differences in feedback cycles across systems. Technological capability may iterate rapidly on monthly or quarterly cycles. Risk structures evolve with deployment, diffusion, and confrontation. International governance generally requires longer processes of negotiation, legislation, verification, and trust-building. These timescales are illustrative decision-making assumptions, not fixed empirical constants applying to every domain. What must actually be compared is the observed iteration speed of critical risks and the feedback speed of governance.

The speed gap has three consequences. First, governance is perpetually catching up. Moving from a framework's conception through negotiation, signature, entry into force, and implementation takes much longer than technological iteration, so today's governance often addresses yesterday's technology. Second, risks evolve on their own during gaps in governance. Technological evolution generates new risks. A governance framework begins responding, but addresses the previous version while new risks continue evolving outside it. Governance perpetually chases the next risk that already exists. Third, issue-by-issue governance is especially mismatched to crisis speed. Existing governance is divided by domain, each with its own cycle, whereas crisis emergence couples domains together. Cross-domain coordination is another order of magnitude slower than governance within one domain. The essential point is that under crisis emergence, governance feedback cycles must match the actual iteration speed of critical risks. Specific cycles vary by technology, domain, and risk type. This is not simply a matter of willingness to govern; it concerns the time constants of governance mechanisms.

Technological acceleration itself is not new; previous technological revolutions also had accelerating phases. This wave, however, has three special, self-reinforcing features. First, cross-domain capability: AI simultaneously affects military affairs, the economy, biology, public opinion, and other domains, with extremely fast transmission between them. Second, commoditized capability: advanced AI moves from

exclusive state-level access into open-source models and commercial interfaces, sharply reducing barriers to diffusion. Third, automated capability: some AI systems already execute tasks autonomously, interact continuously, and adjust strategies dynamically within authorized boundaries. Stronger self-modification or 'autonomous evolution' remains a prospective risk and should not be treated as a general capability of present systems. Together, cross-domain reach, commoditization, and automation make this wave self-reinforcing: the greater the capability, the faster the acceleration itself.

5. Outstripping human response as decision time contracts to machine speed

Human organizational decision-making has inherent time constants. Individual judgment takes seconds to minutes; team decisions hours to days; national decisions weeks to months; and international coordination months to years. At these levels, the ability to use reflection, diplomacy, verification, and restraint as 'slowing mechanisms' fundamentally involves exchanging time for accuracy.

In the AI/AGI age, several changes are compressing parts of decision and response processes. Some autonomous-weapons, cyberdefense, and information-distribution systems can substantially shorten sensing, judgment, and response times. Actual timescales differ greatly between systems and cannot all be described as 'seconds or even milliseconds.' When machine responses are substantially faster than human organizational processes requiring verification, authorization, and coordination, humans may be pushed into ratifying what has already happened. Reflection, diplomacy, verification, and restraint consequently face greater pressure.

This reveals the essence of outstripping human response. It is not merely that 'crises are more serious than before,' which would be a question of degree. It is structural: the upper limit of human organizational capability has fallen below the minimum required under crisis emergence. Condensed into a proposition: under crisis emergence, the lower bound of crisis iteration speed exceeds the upper bound of human organizational response capacity. Three implications follow. Traditional crisis-management tools fail because any tool relying on real-time human decisions cannot keep pace. Prevention must take priority over response because response comes too late. Governance must shift from 'responding to crises that have occurred' to 'preventing crises from being triggered.' This is the most important policy implication: in the AGI age, the core governance objective is no longer how to respond quickly to a crisis, but how to avoid triggering it, a logic fundamentally different from traditional international crisis management.

6. The structural challenge crisis emergence poses to governance frameworks

Many existing governance frameworks remain organized primarily by separate domains and issues. Such frameworks have necessary value within their own fields. But when risks couple across domains and feed back on one another, governance within individual fields cannot, by itself, ensure that the overall risk loop is closed.

There are three reasons for failure. First, issue-specific interventions transmit through coupling loops. Progress in one domain may create risks in another; stability in resources may create new antagonism in narratives. Issue-by-issue governance cannot avoid cross-domain side effects. Second, separate governance faces competition for resources. Finite resources are divided among multiple domains, leaving each short of what it needs. This allocation method itself does not match the requirements of convergent crisis. Third, its organizational structure does not match the crisis structure. The crisis is coupled; organizations are separate. One team governs AI, another handles biological risks, economic departments handle resources, and diplomatic departments handle narratives. The mismatch between organizational and crisis boundaries is precisely the structural blind spot.

The key conclusion is that isolated, issue-by-issue governance frameworks alone struggle to respond fully to convergent cross-domain crises. This is not a problem of willingness or capability, but a fundamental mismatch between the logical premises of governance frameworks and the actual structure of crisis. It does not negate their value: they remain effective within their respective fields. But the scope of that effectiveness does not match the actual scope of crisis emergence. This is the most important policy implication of treating crisis emergence as a distinct concept.

7. Four minimum conditions for action under crisis emergence

This chapter does not offer a specific governance plan; that is subsequent work. It answers only a prior question: under crisis emergence, what minimum conditions must a meaningful direction for action satisfy?

First, coverage must match the scope of the crisis. Because crises are coupled across domains, action must coordinate across domains. Single-domain action is ineffective; action in multiple domains without coordination is equally ineffective. A coordination mechanism must cover all critical domains. Second, response speed must approach crisis iteration speed. If crises iterate monthly or faster, international negotiations operating in years and treaty-making operating in decades are insufficient. Mechanisms must respond quarterly or even monthly. Third, the response paradox must be addressed. Because interlocking loops amplify unilateral responses, multilateral coordination must include at least the two largest actors, and the coordination mechanism must itself resist exploitation by either. Fourth, plural values must be accommodated. Because narrative conflicts erode legitimacy, the mechanism must operate stably amid different values. It cannot assume any particular institutional model as its endpoint. It must remain open to participation by different value systems, and its legitimacy must be jointly recognizable by all participants.

These four minimum conditions are a filter, not a governance plan. Only directions meeting all four merit consideration; otherwise the action is ineffective. Under traditional crisis conditions, these requirements are luxuries. Under crisis emergence, they are the passing grade. Any direction failing them, however sophisticated its presentation, will be rejected by the structure itself. This is the starting point for the later discussions of eliminating five alternative paths and of collaborative coexistence. The former tests the five traditional alternatives against these conditions; the latter explains why the only direction that passes points toward collaborative coexistence. The filter does not select a fixed institutional template, but a direction: seeing the interlocking whole and changing the old mode of strategic interaction accordingly. That direction is unique; this is the meaning of 'only' in the subsequent 'only path': one direction. More than one specific form can satisfy it. Collaborative coexistence is a category of institutional convergence, not a single design precisely selected by these four conditions. The direction is singular; the forms are not. This distinction runs throughout the chapter and the text.

8. The limits of this chapter's claims

Intellectual honesty requires stating the limits of the conclusions. This chapter establishes three things. Where accelerating AGI technology, dependence on critical nodes, difficult attribution, and narrative polarization coexist, the structural characteristics of crisis emergence, coupling, interlocking, acceleration, and outstripping human response, necessarily appear. The logical premises of existing issue-specific governance are structurally misaligned with crisis emergence. And under crisis emergence, a direction for action must satisfy four minimum conditions or be rejected by the structure itself, however sophisticated its plan.

The chapter does not establish several other claims: that collaborative coexistence as a specific institutional design must succeed; that crisis emergence must occur in every future setting; that crises in the AGI age must be irreversible; or that human organizational capabilities are outstripped in every respect. The argument relies primarily on reasoning about mechanisms, supported by real-world anchors. Mechanism analysis demonstrates that 'under given conditions, the structure itself necessarily produces a certain outcome.' Real-world anchors demonstrate that 'those conditions already exist.' Together, the strength of the argument is that 'under real-world conditions, the outcome indicated by structural reasoning is highly likely.' This is sufficient to support the subsequent chapters, but insufficient for strong conclusions such as 'collaborative coexistence must succeed.' That would require a dedicated analysis of a specific plan, beyond this chapter's scope. Establishing crisis emergence as a distinct concept serves three purposes. For diagnosis, it exposes a long-overlooked structural fact: multiple risks no longer merely coexist, but are coupled. For frameworks, it reveals the structural limits of existing issue-specific governance: its effective scope does not match the actual scope of crisis. For direction, before the five alternatives are eliminated one by one, it establishes minimum criteria for what constitutes meaningful action, avoiding wasted effort on structurally ineffective directions. Crisis emergence is not a future risk. It is an existing structural fact. Recognizing this is necessary to understand what every subsequent chapter argues.

Chapter 2 Real-world interstate strategic interaction cannot be bypassed

This text sets out a conclusion repeatedly examined and still difficult to overturn despite every effort: in an age when AI, AI-enabled biological risks, resource thresholds, the legitimacy of meaning, and Earth's finite limits converge, only one path remains for human civilization's long-term survival: move from zero-sum competition toward collaborative coexistence, with Civilization Leap as the direction.

This path is not proposed to accommodate a beautiful wish, nor to weaken any nation's security, development, or institutional interests. On the contrary, it first redefines long-term national interests. As the shared foundation for survival begins to collapse, national security, national development, the future of national institutions, and the overall stability of human civilization can no longer be wholly separated.

If interstate strategic interaction continues to rest on uncontrolled competition, if national governance remains driven by short-cycle security anxiety, development pressure, and political returns, and if irreversible risks remain divided among departments, professions, and terms of office for separate treatment, human civilization, including one's own nation, has no genuinely sustainable future.

Yet this path cannot begin directly from an abstract 'humanity as a whole.' For those who govern, interstate strategic interaction remains the most immediate and unavoidable level. States possess armed forces, public finances, laws, borders, industrial policies, energy systems, diplomatic commitments, technical standards, and narratives of legitimacy. Whether AI is militarized, resources restricted, or supply chains broken is expressed first through states.

[The full relationship between collaborative coexistence and Civilization Leap]

'Collaborative coexistence' is the only long-term direction indicated by elimination after the five alternatives are tested and found functional but incapable, individually, of providing a long-term destination. It does not require states to abandon sovereignty, defense and security, institutional differences, or essential industrial-development requirements. It is higher-order self-interest: in a highly coupled world, any party's pursuit of its long-term interests can succeed only if it also safeguards shared foundational security. Competition remains, but within boundaries that prevent destruction of the shared foundation for survival. It does not abolish nations. It places civilizational guardrails around their competition.

Collaborative coexistence is not itself the endpoint. It is a foundational condition for Civilization Leap. The substance of the leap is to use AGI capabilities and institutional design to reconstruct how humans acquire security, worth, and meaning. It does not demand a change in human nature from self-interest to altruism. Instead, it shifts the sources of worth from possession and overpowering others toward being needed, creating, and protecting. People still act for their own sense of worth. But with its sources redirected, creating value for others and safeguarding the shared foundation can become important priority pathways to legitimate self-interest at the institutional level. This does not reduce genuine compassion, sacrifice, responsibility, or meaning to self-interest.

The logic of self-interest was a basic driving force in agricultural and industrial civilization. When land, capital, and labor were scarce, individuals' and nations' pursuit of self-interest drove productivity growth and institutional evolution. It is also a root of human conflict: when extreme competition amplifies it without limit, war, plunder, ecological overuse, and collapsing trust follow. In earlier eras, great-power interaction, nuclear deterrence, and limited destructive capabilities constrained such conflict. The AGI age is dismantling those constraints. Destructive capability is no longer tied to national scale; attacks may resist attribution; decision time contracts to machine speed. The consequences are not merely gradual and cumulative. They erupt first in interstate strategic interaction. When several sovereign states are simultaneously driven by self-interest toward arms races, blockades, confrontation, and miscalculation, the global foundation for shared survival faces irreversible threats.

The first goal of Civilization Leap is therefore neither interstellar travel nor immortality through meaning. It is freedom from survival insecurity for humanity as a whole. When AGI capabilities enter interstate

competition, this is not just a moral baseline but the common foundation of every nation's long-term interests. Once that foundation is destroyed, no nation's security, development, or institutional future remains possible.

The full chain is: the old mode of strategic interaction is unsustainable → the five alternatives each have functions but cannot independently provide a long-term destination → within the identified constraints and tested paths, collaborative coexistence is the only long-term institutional direction not yet refuted → collaborative coexistence is a foundational condition for Civilization Leap → Civilization Leap is the logical prerequisite for collaborative coexistence to sustain itself → the first goal of Civilization Leap is freedom from survival insecurity for humanity as a whole. Collaborative coexistence is the starting point, not the endpoint.

When these risks combine, AI militarization and autonomy, critical resources and computing power, AI-enabled biological risks, critical-infrastructure vulnerabilities, civilizational narrative conflicts, the AGI learning environment, and Earth's finite resources, security, development, biological survival, and the future cease to be independent issues and become a mutually amplifying national-level feedback loop. The main text groups them into four interlocking strategic contests for deeper examination. Once this loop begins forming, no nation can bear its consequences indefinitely through its own strength alone. The clear-headed choice is to begin negotiations on a joint response mechanism for future risks and crises as soon as possible, before comprehensive loss of control, rather than wait for irreversible crisis to force coordination.

If this judgment holds, what is at stake is not any one country's temporary victory or defeat, but whether humanity can cross the greatest threshold in the history of civilization.

Chapter 3 Eliminating five alternative paths through concrete mechanism analysis

[Division of work with No Escape] This chapter eliminates the five alternatives through concrete mechanism analysis, allowing decision-makers to grasp quickly why each path fails. For more rigorous, point-by-point falsification, formally testing the viability of each alternative within the combined constraints of six variables, see the companion text *No Escape: Civilizational Reasoning Under Six Variables*. This chapter helps readers 'see why it fails'; *No Escape* establishes 'why it fails as a matter of logic.'

Before reaching the conclusion above, five alternatives most readily adopted by decision-making systems must be examined. Testing their operating conditions in the new structure reveals why they cannot simultaneously address AI autonomy risk, AI-enabled biological security risk, resource-threshold risk, and risks concerning legitimacy of existence. The conclusion's claim to uniqueness rests on eliminating these five paths one by one.

A shared test for the five paths

Before testing each, the four-part framework for deciding whether a path can provide a direction for long-term survival must be explicit.

Any alternative must pass all four tests:

First, can it address AI autonomy risks? Can AI weaponization, autonomy, lost-link combat, and compressed decision time be effectively constrained under the path, rather than reinforced by its structure?

Second, can it address AI-enabled biological risks? Can diffusion of AI-enabled biotechnology, difficult attribution, and irreversible catastrophe be prevented or buffered?

Third, can it address resources and the right to development? Can critical resources and computing power be allocated, development space safeguarded, and bloc-formation risks managed reasonably, rather than intensified by the path's own logic?

Fourth, can it address long-term legitimacy of existence? Can groups with different civilizations, political systems, and values coexist over time without the path's structure producing irreconcilable identity conflicts?

If any test remains unmet over the long term, the path cannot provide a direction for civilization's long-term survival.

The paths are tested below.

Alternative path 1 Technical control: Can hardware locks, audits, and regulation control AGI?

The strongest form of technical control embeds circuit breakers at the hardware level, applies end-to-end auditing during training, and requires strict deployment approval, assuming these technical measures can ensure AGI does not escape control.

Even if that assumption holds, the problem remains unresolved.

The reason is that the security dilemma between states exists independently of whether AGI 'escapes control.' As long as AGI is seen as the core of next-generation national capability, economic competitiveness, military deterrence, scientific acceleration, and governance efficiency, states have ample incentive to compete at the edges of technical-control frameworks: accelerate deployment within compliance, seek breakthroughs in gray areas, and establish exclusive advantages at critical nodes.

Technical controls can constrain 'clear violations,' but cannot constrain 'an arms race within the rules.' A satellite approved under a control framework can serve both civilian communications and military early warning. An audited AI model can operate both in peacetime and on the eve of conflict. The framework's 'legitimacy' may itself shield the accumulation of capability.

More fundamentally, technical control does not change the objective function of strategic interaction. States still pursue relative advantage, defenses against first strikes, and control over critical resources. Regulation adds a guardrail to the race, but does not change the race's structure.

Technical control is therefore not wrong; it is necessary but insufficient. It can reduce some types of risk, but cannot resolve the old competitive logic itself. Where states regard AI as a core strategic capability and lack credible shared constraints, incentives to draw AI into security competition persist. Technical control alone struggles to isolate it entirely from that competition.

Further, even if future AGI is completely controllable technically, unchanged interstate competition will still draw its capabilities into militarization, blockades, computing-power competition, and escalating deterrence. Technical control may reduce some risks, but cannot alter the objective function of strategic interaction.

At a deeper level, regulators and AGI are inherently asymmetric: regulation lags evolution, and audit cycles are slower than system change. This reinforces rather than replaces the conclusion that technical control is insufficient.

Technical control is therefore a necessary condition, not a long-term order.

[Four-part test]

- AI autonomy risk: not addressed. Technical control does not change the interstate security dilemma; AGI capabilities will still enter the arms race.
- AI-enabled biological risk: partly buffered, but concealed diffusion cannot be prevented.
- Resources and the right to development: not addressed. Control does not change the structure of resource competition.
- Long-term legitimacy of existence: not addressed by this path.

Test result: fails the first test.

[Path 1 eliminated]

Alternative path 2 Each side strengthens deterrence to maintain stability through a balance of terror

[Illustrative mechanism]

Suppose States A and B are competitors. A deploys an AI-enhanced autonomous border-surveillance and rapid-response system. Learning this, B must decide whether A's system is 'defensive redundancy' or 'forward deployment of first-strike capability.'

B cannot be sure. The same system, rapid target recognition, threat assessment, and automatic response, appears as 'border management' in peacetime, yet can turn within one minute into first-strike suppression of B's core nodes after conflict begins. B's rational choice is that it cannot wait for clear evidence before responding. It deploys a corresponding system with equal or greater autonomy and includes a requirement that 'the system can execute preset retaliation even if command links are severed.' Observing B's deployment, A reaches the mirror-image judgment and shortens its system's decision cycle further.

Every round of 'defensive upgrading' appears to the other as 'forward positioning for offense.' This is not because either side is malicious, but because the same capability can switch between peace and war at extremely low cost, at great speed, and with very weak signals.

Under this structure, deterrence's three pillars are hollowed out one by one:

- Symmetry: a satellite jammer can disable a command system dependent on space-based links. Destructive capability is no longer proportional to national scale.
- Visibility: a software update changes the deterrence balance without launching a missile or moving a fleet.
- Attributability: an incident may be explained as a 'system misjudgment' or a 'deliberate test,' yet neither side can establish which.

Cold War nuclear deterrence rested on the substantial observability and verifiability of nuclear weapons' locations, numbers, and launch capabilities. Even with miscalculation, both sides had a window for diplomatic communication. In strategic interaction involving AI, biology, cyber operations, and autonomous systems, capability deployment is often concealed, deniable, and hard to observe. One nation can complete a critical upgrade without sending any signal, leaving another to judge through conjecture rather than confirmation.

Future balances of terror will therefore depend increasingly on guesses about unknown intentions rather than calculations about known capabilities. This is less stable than the Cold War condition.

[Logical necessity] When visibility of action disappears, attribution fails, and destructive capability becomes detached from scale, deterrence's returns cease to be predictable. As visibility, attributability, and cost symmetry continue declining, deterrence-based stability declines and pressure grows on all parties to move judgments forward, accelerate decisions, and increase automation under incomplete information.

[Battlefield validation: cost asymmetry destroys deterrence's premises]

A balance of terror depends on an implicit premise: destructive costs are broadly symmetrical, so either side's expected costs of attack exceed its expected gains. Yet long-range drone attack and defense on the Russia-Ukraine battlefield in 2025–2026 presents the opposite balance sheet. When a single FPV drone costing around 2,000 can inflict millions of US dollars in losses, deterrence's premise of 'cost symmetry' has been hollowed out. This is not a thought experiment. It is the real balance sheet of 2026.

[Four-part test]

- AI autonomy risk: not addressed. Deterrence itself accelerates autonomy.
- AI-enabled biological risk: not addressed. Concealment and difficult attribution place biological risks outside deterrence's coverage.
- Resources and the right to development: not addressed. Deterrence does not change the logic of resource competition.

- Long-term legitimacy of existence: not addressed. Deterrence presumes permanent hostility and supplies no framework for coexistence.

Test result: fails all four tests.

[Path 2 eliminated]

Alternative path 3 A few major powers jointly manage AI and critical resources

[Illustrative mechanism]

Suppose technologically leading states form an 'AI and Resource Control Alliance.' They agree internally on chip-trade quotas and tiered registration of AI models. Outside the alliance, State C possesses an independent technological base and a complete industrial system. C must decide whether the alliance's controls are 'a global security arrangement' or 'institutionalized technological monopoly.' Its rational choice is to accelerate a parallel system: domestic chip production, independent data centers, and alternative model architectures.

Parallel systems are not a solution, but a new source of risk. The two sides' technical standards, security protocols, and data formats are incompatible.

More importantly, the alliance contains structural tensions. State D is a member and a major advanced-chip producer. Controls require export restrictions, but D faces a calculation: covertly supplying chips to C through third parties can yield extraordinary returns.

The tighter the controls, the higher the returns to violating them; the higher those returns, the greater the pressure for verification inside the alliance; and the greater that pressure, the higher the cost of trust between members.

A deeper problem remains: the stability of the joint-management alliance itself cannot be guaranteed.

Even if an alliance is successfully established, the same security dilemma that exists outside it will persist within it. Gaps in computing power among member states may widen; AGI capabilities may become unbalanced; biotechnology breakthroughs may alter the balance of power; changes of government may change an alliance's positions. One member's "enhancement of defensive capabilities" may look to another like the "forward deployment of offensive capabilities."

Joint-management agreements will not eliminate such judgments. They will merely add another layer of narrative packaging: "compliant" or "noncompliant."

When State A believes that State B is accumulating asymmetric advantages within the agreement's framework, while State B believes that State A's technological breakthroughs are changing the balance of power within the alliance, joint management ceases to be a solution and becomes a new arena of strategic competition. Alliance members may use the agreement's provisions to protect their own advantages and limit their rivals' room to rise, rather than genuinely constrain the conduct of every member.

Historical multilateral technology-control arrangements, such as the Coordinating Committee for Multilateral Export Controls, the Missile Technology Control Regime, and the Nuclear Suppliers Group, have generally faced varying degrees of internal tension: member states' interests do not fully align; changing technological gaps redefine "who needs to be constrained"; and deteriorating political relations undermine the mutual trust on which existing agreements rest.

Joint management by major powers can therefore serve only as a transitional buffer, not as a long-term structure for civilization. It cannot eliminate the structural conditions that produce the security dilemma. It merely moves competition from the global level into the alliance itself.

[Logical Necessity] Joint-management mechanisms create incentives to defect and reactions against exclusion. The tighter the controls, the more parallel systems emerge; the more parallel systems, the less predictable the global technological environment becomes. Meanwhile, the security dilemma within the alliance is reactivated as the balance of power changes.

[Four-part test]

- AI autonomy risks: Partly buffered, but competition within the alliance will continue to drive greater autonomy.
- AI-enabled biological risks: Cannot be addressed effectively. The alliance cannot cover every biotechnology node.
- Resources and the right to development: Cannot be addressed. The alliance is itself a resource-control mechanism and will inevitably prompt excluded parties to build parallel systems.
- Legitimacy of continued existence: Cannot be addressed. Excluded parties do not recognize the alliance's legitimacy.

Test result: A partial pass on the first criterion; failures on the second, third, and fourth.

[Path Three: Ruled Out]

Alternative Path Four: Establish a Single Global Sovereign or Supranational Coercive Authority

An institution capable of exercising unified control over global computing power, AI models, biotechnology, critical resources, and infrastructure would create an unprecedented concentration of absolute power.

The issue is neither technical feasibility nor administrative efficiency.

The issue is legitimacy.

Where civilizational narratives remain plural over the long term, encompassing different systems, histories, and values, an ultimate authority that is highly centralized, offers no right of exit, and lacks effective checks and balances faces a significant risk of being regarded by substantial populations as an alien oppressor. It does not need to “do anything wrong.” Its mere existence will be understood by some civilizations as a threat to their way of existing.

To preserve unity, this authority must continually expand its coercive capabilities and the scope of its control. Each response to a “crisis of legitimacy” will strengthen its power structure, and each strengthening of that power will deepen distrust among the populations it governs. Over time, the coercive capacity required to maintain stability will continue to grow, until the authority itself becomes the world's greatest systemic risk.

Once a center that permits no exit, admits no checks, and has no alternative becomes corrupt, makes a mistake, or is captured, no other power will be able to correct the consequences.

Almost every major catastrophe in human history has been confined to particular regions or domains. Wars, plagues, ecological collapse, institutional failure: even when enormous in scale, they have always left humanity room to repair itself through migration, difference, competition, and alternatives. If a unified global power with no exit, no checks, and no alternative makes a systemic error, humanity will, for the first time, lose this mechanism of correction through diversity. Diversity has been a source of civilizational conflict, but it has also always been civilization's minimum insurance against extinction.

[Logical Necessity] A single authority creates a power center with no exit, no checks, and no alternative. The cost of preventing technological loss of control may exceed the cost of technological loss of control itself.

[Four-part test]

- AI autonomy risks: Theoretically addressable, but dependent on the quality of the authority's own decisions.
- AI-enabled biological risks: Theoretically addressable, but dependent on the quality of the authority's own decisions.
- Resources and the right to development: Theoretically addressable, but allocation powers are concentrated in a single center.
- Legitimacy of continued existence: Cannot be addressed. Some civilizations will inevitably view the single center as oppressive.

Test result: The first three criteria could theoretically be passed, but the fourth fails, and that failure will erode the feasibility of the first three over time.

[Path Four: Ruled Out]

Alternative Path Five: Limited Cooperation + Long-Term Competition—the Gradualist Path

Realist scholars of international relations and many strategic decision-makers would propose a more cautious path:

“We do not need to construct a grand order of collaborative coexistence. We need only establish limited mechanisms for cooperation in a few high-risk areas, such as AI safety, biosecurity, and nuclear nonproliferation, while maintaining normal competitive relations among states.

Over the past few decades, this is precisely what humanity has done in nuclear nonproliferation, climate cooperation, counterterrorism, and public health. Why can we not continue doing so?”

The attraction of this path is that it requires no state to give up sovereignty, security, or development interests; no reconstruction of the overall order; and no alteration of institutional differences. It appears to be “the most realistic incremental solution.”

[Mechanism Analysis]

Let us now examine the conditions under which this path would operate in the new structure of AI, AGI, biological risks, and competition over critical resources and computing power.

Step One: Can Localized Cooperation Be Insulated from Overall Competition?

Suppose states agree to establish a limited cooperation mechanism for AI safety: regular notification of major AI system deployments, emergency hotlines, and information exchanges on cases of “AI misjudgment.”

At the same time, competition continues elsewhere: races for computing power, chip blockades, confrontation between satellite systems, and contests over institutional narratives.

The problem is that no firewall separates cooperation on AI safety from overall competition. The “AI safety experience” that State A shares at a meeting may also be interpreted by State B as “a tactical demonstration of State A’s AI capabilities.” The “system deployment” that State B reports over the hotline may be interpreted by State A as “a signal that State B is establishing an asymmetric advantage.”

Within the old framework of strategic competition, information channels may reduce misjudgment, but they may also be used for strategic deception or false signaling. Against a background of zero-sum rivalry, signals transmitted through localized cooperation will be reinterpreted by opponents through the logic of competition.

Step Two: Can Localized Cooperation Withstand Spillovers from Foundational Risks?

Cooperation in traditional areas such as climate change, nuclear nonproliferation, counterterrorism, and public health has been possible because their impact is bounded, attribution is relatively clear, and means of intervention are visible.

But AI autonomous weapons, AGI safety, AI-enabled biological risks, and global computing infrastructure are different. They are foundational risks: they do not “belong” to a particular domain but constitute the underlying conditions on which every domain operates.

- A failure of AGI safety affects not one industry but every domain dependent on digital infrastructure.
- The spread of AI-enabled biological risks places simultaneous pressure on public health, food security, economic stability, and social trust.
- The destruction or blockade of global computing infrastructure simultaneously paralyzes military affairs, finance, energy, scientific research, and governance.

- Attacks on satellites and communications links simultaneously break the chains of early warning, command, judgment, and retaliation.

The defining characteristic of foundational risks is that they continually cross domain boundaries. When a single AGI safety incident simultaneously triggers financial-market volatility, public-health panic, escalating military misjudgment, and mobilization of public-opinion systems, states will discover that there is no independent decision-making space between “finance” and “security.” They are different aspects of the same event.

Localized cooperation will therefore continually be forced to expand in the face of foundational risks. States will discover that cooperation on “AI safety” alone is insufficient because it cannot be separated from “critical-infrastructure protection”; cooperation on “critical-infrastructure protection” alone is insufficient because it cannot be separated from “resource and computing-power allocation”; and cooperation on “resource and computing-power allocation” alone is insufficient because it cannot be separated from “civilizational narratives and institutional competition.”

Step Three: Can Expanding Localized Cooperation Always Remain “Limited”?

As the scope of cooperation is continually forced to expand, the path of limited cooperation plus long-term competition encounters a structural contradiction: cooperation requires common rules and constraints, while competition requires freedom of action and relative advantage.

At each point of expansion, states face the same judgment: can we accept common constraints in this domain without relinquishing our competitive advantage?

In high-risk, highly uncertain domains, the answer will increasingly tend toward “no.” This is not because states are irrational, but because under the old logic of strategic competition, whoever accepts constraints first is regarded as surrendering the initiative.

Limited cooperation therefore faces a dilemma:

- Refuse to expand cooperation → Spillovers from foundational risks render localized cooperation ineffective.
- Accept an expansion of cooperation → Cooperation gradually erodes the space for competition, yet no party is willing to relinquish competition voluntarily.

Neither leads to a stable equilibrium. The first turns cooperation into a formality; the second prevents it from advancing.

[Logical Necessity] Where foundational risks are interconnected and continually cross boundaries, limited cooperation plus long-term competition cannot remain stable. It either fails through insufficient coverage or, when forced to expand, exposes the structural contradiction that the old logic of strategic competition cannot sustain common constraints.

Realists may respond: “Then let us expand cooperation gradually and see how far we get.”

This “see how far we get” gradualism faces a structural constraint: gradual expansion requires cooperation to expand faster than risks spill over. When all four feedback loops accelerate at once, gradual expansion may continually lag behind risk spillovers, especially when negotiation, trust, and institutional adjustment are much slower than technological diffusion. The speed of risk spillovers is determined by the diffusion of technological capabilities and the coupling of vulnerable structures; the speed of expanding cooperation is determined by negotiation, the accumulation of trust, and institutional adjustment. The former accelerates far faster than the latter. Each risk spillover weakens the trust underpinning the next negotiation, further slowing expansion. This is a negative feedback cycle: the faster risks spill over, the slower cooperation expands; the slower cooperation expands, the faster risks spill over.

Localized cooperation will ultimately continue expanding, until it converges on some form of common foundational security order. Otherwise, it will continually be overwhelmed by spillovers from foundational risks.

[Four-part test]

- AI autonomy risks: Cannot be addressed. Localized cooperation cannot cover the rapid evolution of autonomous systems.
- AI-enabled biological risks: Cannot be addressed. Localized cooperation cannot cover the concealed diffusion of biotechnology.
- Resources and the right to development: Cannot be addressed. Localized cooperation does not change the structure of resource competition.
- Legitimacy of continued existence: Cannot be addressed. Localized cooperation supplies no framework for coexistence and cannot manage spillovers from institutional conflict.

Test result: Failures on all four criteria.

[Path Five: Ruled Out]**The Conclusion on Which the Five Paths Converge**

Once the five paths have been ruled out one by one, elimination points to the only remaining direction: while states remain sovereign entities, establish a mechanism of common constraints that requires neither a merger of sovereignty, nor the abandonment of competition, nor trust in advance. This is what this paper calls collaborative coexistence.

It is not an idealistic choice, a moral appeal, or an emotional declaration that “humanity should unite.” It emerges from this paper’s comparative convergence across five mainstream categories of approach: within the constraints currently identified and the set of paths examined, no long-term alternative has yet been found that can satisfy all four requirements together. Should a new viable mechanism emerge, the judgment should be reassessed.

Collaborative coexistence is not itself the destination. It is the foundational condition for Civilization Leap. The substance of that leap is to use AGI’s capabilities, through institutional design, to reconstruct the ways people acquire security, a sense of worth, and meaning. It does not demand that human nature change from self-interest to altruism. It redirects the way people gain a sense of worth: away from possession and overpowering others, toward being needed, creating, and protecting. People still act for their own sense of worth, but when its sources are redirected, creating value for others and protecting the common foundations can become important priority pathways to legitimate self-interest at the institutional level. This does not reduce genuine compassion, sacrifice, responsibility, and meaning to self-interest. It is a fundamental change both in institutional logic and in the sources of security, worth, and meaning. The first objective of Civilization Leap is freedom from survival insecurity for humanity as a whole. In an era when AGI capabilities are incorporated into strategic competition among states, this is the common foundation of every state’s long-term interests.

Transition: Collaborative Coexistence Is Not the Destination

Having ruled out the five alternative paths one by one, this paper establishes only one point here: collaborative coexistence is an institutional direction that must be initiated, but it is not the ultimate destination. Collaborative coexistence places shared guardrails around competition, preventing irreversible risks from pushing all parties into loss of control together. The later discussion of Civilization Leap explains why common survival cannot stop at crisis management but must address the sources from which crises repeatedly arise.

In other words, this passage does not set out a full program for Civilization Leap. It simply identifies the main thread: first preserve common survival through minimum shared boundaries, then explain later how Civilization Leap can remove the conditions under which the embers of crisis are continually generated and amplified.

Chapter 4: Why the State Level Cannot Be Bypassed—Real Power Still Resides Primarily in States

The judgment that humanity must ultimately prepare for its long-term survival as a whole holds true over the long term. But action in the real world cannot bypass states. The organizations able to mobilize immense resources, organize complete industrial chains, legislate and regulate technology, deploy national defense forces, define borders, sign international treaties, and bear the consequences of war and governance remain primarily sovereign states.

AI cannot develop in a vacuum outside the state system. Computing facilities, electricity grids and energy, advanced chips, research talent, vast quantities of data, cloud infrastructure, military-industrial applications, industrial subsidies, import and export controls, and authority over algorithmic regulation are all jointly shaped by national policies, regional alliances, and territorial jurisdiction over corporations. Biological risks likewise cannot be separated from the state system. Resource flows, too, are embedded in structures of state power.

Civilizational narratives and institutionalism are even less purely intellectual matters. Technological trajectories become tied to narratives of institutional superiority; struggles over resources escalate into narratives of national survival; and disputes over development models evolve into contests over civilization's future path.

But “cannot be bypassed” does not mean “will resolve itself.” As vessels of power, states can sustain strategic competition without necessarily generating the motivation to constrain their own modes of competition. In a world organized around states, every measure of self-protection is interpreted by other states as a threat; every technological breakthrough is treated as security pressure; and every destabilization of domestic governance drives a more aggressive external posture. The old mode of strategic competition continually reinforces itself precisely because an external braking mechanism is missing. No state can unilaterally stop this acceleration without relying on cooperation from others.

[Logical Transition] This is the point of intersection between the state level being “unavoidable” and “not the whole story”: states are the real-world carriers of risk, but solutions must extend to relations among them. Collaborative coexistence does not replace states. It installs a shared external braking system for competition among them.

Chapter 5: Interstate Competition Is Not the Whole Story—Even a Unified World Would Not Automatically End Civilizational Crisis

The fact that the state level cannot be bypassed does not mean it encompasses everything. Before explaining “why it is not the whole story,” however, one especially critical structural node within the state level at the present stage must be identified: the U.S.–China relationship.

This paper's strategic judgment is that the structure of U.S.–China competition is one of the most critical amplification nodes in the current global risk network and interacts with the mechanisms of risk diffusion in the AI era. The Russia–Ukraine conflict, the situation in the Middle East, the Taiwan Strait, the South China Sea, China–Japan relations, and global supply-chain restructuring each have their own causes, but are also affected to differing degrees by the structure of U.S.–China competition. Minimum management mechanisms between China and the United States covering an AI crisis hotline, strategic-stability communication, biosecurity notifications, de-escalation of critical-infrastructure misjudgments, and boundaries on autonomous weapons operating after communications loss would therefore substantially improve the operating environment of other multilateral mechanisms. Conversely, continuing instability at this level will increase uncertainty in coordination at every level.

A boundary must be clarified immediately to avoid misunderstanding: identifying China and the United States as the most critical structural node does not mean that the world's problems are merely U.S.–China problems. Their relationship is the most important structural node in today's risk network, the principal risk

amplifier at the present stage, and one of the most practical entry points for risk management, but it is not the source of all risks. This is fully consistent with the paper's central judgment: competition among states, especially U.S.–China competition at present, is only the principal real-world channel through which risks are amplified; their true root is the continual amplification of the logic of self-interest in the AGI era. Effective management of U.S.–China competition therefore does not mean abolishing competition. It means establishing civilizational guardrails capable of preventing systemic loss of control while competition continues. A competitive world can remain stable; a world of uncontrolled competition may have no winners.

This also identifies the most practical location for the first negotiating table. It need not initially be global. It may well be the table at which China and the United States establish arrangements for irreversible risks, because that is the most critical node in the present risk network and the place where minimum common interests are most readily identified. But this is an entry point, not a destination. How the mechanism can expand from this node to broader multilateral participation and the Global South is a matter of implementation design, developed in this framework's implementation annexes.

Let us return to the deeper proposition this chapter seeks to establish: the state level, including even its most critical U.S.–China node, is still not the whole story.

Even if all sovereign borders dissolved, land, energy, computing power, freshwater, minerals, development opportunities, and the space in which individuals can realize their worth would still face aggregate limits. Zero-sum competition among states is the most urgent external constraint today, but even if the entire world became a single state, the following three layers of structural crisis would remain:

First layer: The zero-sum structure of resource allocation. A unified state changes the allocator, not the constraints. Where life spans are greatly extended, critical resources and opportunities remain constrained, and mechanisms for renewal and exit have not been reconstructed in parallel, long-accumulating distributional conflicts may intensify significantly.

Second layer: The erosion of “being needed.” Automation separates production from human labor; guaranteeing basic income does not guarantee basic meaning. When billions of people fall into the hollow condition of “not being needed” amid material abundance, the cost of collapsing social trust tends toward infinity.

Third layer: Conflicting understandings of institutional legitimacy. A unified state can suppress overt conflict but cannot eliminate differences among groups over what constitutes a “legitimate order.” In the AGI era, these differences are amplified: who controls AI? Who is governed by AI? Who has the right to enter the future? Accumulating differences in understanding will erupt at a tipping point, while AGI magnifies the destructive force of each eruption.

Together, these three layers yield a further directional judgment: risks will increasingly shift from relations between states into civilization itself. The diffusion of AI, automation, and biotechnology is giving ever smaller groups, and even individuals, destructive capabilities once possessed only by states. At the same time, automation's displacement of social roles, the declining sense of being needed, and the loss of meaning continually generate despair and resentment. When capabilities spread downward while resentment accumulates, the most dangerous combination emerges: resentment begins to acquire capability. Continually producing despair, feelings of abandonment, and a vacuum of meaning is therefore itself becoming a civilization-scale risk. This is an exceedingly deep layer, involving civilization's internal structures of value and meaning. It lies beyond this paper's focus on “preventing loss of control among states” and is developed in other texts within this framework. It is noted here only to show that the ultimate locus of the problem lies deeper than interstate antagonism. To conclude this layer in one sentence: a world with only one state is not a civilization that has already made the leap. As long as putting self-interest first remains civilization's underlying operating logic, resource competition, conflicts of meaning, fractures of identity, and conflicts over the legitimacy of existence will reappear in new forms. What Civilization Leap truly changes is not the number of states but civilization's underlying operating logic. Ultimately, world unification changes the carrier of risk; Civilization Leap changes the mechanism that generates it.

Conclusion: State Unification Eliminates Sovereign Warfare as a Carrier of Risk, but Not the Amplification of Self-Interest in the AGI Era

The roots of the three layers of crisis lie not in interstate antagonism but in the coupled tensions produced when AGI capabilities amplify the logic of self-interest.

This needs clarification to avoid misunderstanding: since even “state unification” cannot resolve these three layers, “interstate” collaborative coexistence cannot eradicate them either. Collaborative coexistence addresses the risk of loss of control among states, which is precisely this paper’s task. The three layers point to something deeper: the true root of the problem is the amplification of the logic of self-interest itself in the AGI era. Eradicating it therefore requires not “eliminating states” but “changing how the logic of self-interest operates.” That is what Civilization Leap means. It is, however, a separate and deeper undertaking, beyond this paper’s focus on “preventing loss of control,” and is addressed in other texts within the framework. Here, the role of the three layers is limited and clear: they demonstrate that the problem is deeper than “interstate antagonism,” showing that interstate crisis management alone cannot settle it once and for all. Preventing loss of control is only the first step that must be taken.

External competition squeezes the space available internally; internal destabilization, in turn, intensifies external confrontation, ultimately creating a closed loop of mutual deterioration.

Part II: The Four Interlocking Strategic Contests and Their Feedback Loops

Thus, even if collaborative coexistence resolves irreversible risks among states, it remains only the first layer. The true sources of crisis lie in how people acquire security, a sense of worth, a sense of being needed, and meaning. Unless these questions are addressed, crises will shift from interstate forms into society, structures of meaning, and the civilizational order itself.

A unified state can reduce sovereign warfare but cannot automatically eliminate zero-sum resource constraints; it can redistribute wealth but cannot redistribute the sense of being needed; it can suppress conflict but cannot automatically generate shared meaning.

Chapter 6: The First Strategic Contest—Separate National AI Development, Militarization, and Autonomy

States regard AI as the core of next-generation national capability. That reality cannot be avoided. Economically, militarily, and in governance, AI bears on core national capabilities. Every state therefore finds it difficult to slow AI development voluntarily: whoever slows first fears being left behind; whoever accepts limits first fears an opponent achieving a breakthrough in secret. This is the starting point of the AI security dilemma.

6.1 The Capabilities Race Drives Greater Autonomy

When one side uses AI for drone swarms, automated target recognition, and rapid decisions on retaliation, the other finds it difficult to maintain a low level of automation. The result is that no party genuinely wishes to relinquish decision-making authority entirely, yet every party, under pressure, continually cedes portions of that authority.

6.2 Decision Speed Compresses Human Restraint

Human restraint requires time: for diplomatic communication, intelligence verification, command confirmation, political judgment, and the processing of public emotion. One of the greatest dangers of AI conflict systems is that they compress crisis decision-making to machine speed. As unmanned systems, cyber systems, early-warning systems, and automated countermeasure systems interact, escalation may unfold in seconds, minutes, or hours. Humans are not excluded altogether; they are pushed into the position of “ratifying what has already happened.”

6.3 The Chain of Responsibility Is Diluted

When AI systems participate in target identification, threat assessment, recommended actions, and automated responses, responsibility is dispersed among algorithms, data, models, deployers, operators, commanders, companies, and national regulatory systems. When misjudgment, spillover from an attack, or infrastructure paralysis occurs, all parties tend to explain it as a system failure, data anomaly, or third-party attack. If responsibility cannot be clearly assigned, retaliation cannot be precisely restrained.

6.4 Gray-Zone Conflict Expands

AI also expands the space for gray-zone operations. Cyberattacks, information manipulation, financial disruption, supply-chain sabotage, harassment by unmanned systems, satellite interference, deepfakes, and automated public-opinion warfare may all continue without a declaration of war. They need not trigger war immediately, but they will cause all sides, amid prolonged suspicion, to accelerate armament, block technology, and tighten borders.

6.5 Vulnerable Communications Links Drive Autonomy After Communications Loss

The militarization of AI has a deeper structural source: modern military systems increasingly depend on satellite navigation, data links, remote communications, cloud computing, and rear-area command. In high-intensity confrontation, all of these can be jammed, suppressed, deceived, or severed. Once communications links are cut, frontline unmanned systems cannot simply wait for authorization from the rear. To continue their missions while disconnected, they inevitably must be designed with local sensing, judgment, navigation, and execution capabilities. States move toward autonomous weapons not because they disregard human control, but precisely because they fear an opponent severing the chain of control. They therefore keep placing more judgment capability inside weapons systems themselves.

6.5.1 Gray-Zone Proxies: The Second Pathway to Asymmetric Strikes

The vulnerability of critical dependencies pushes states toward lost-link autonomy through a “technical pathway”: fear that the control chain will be cut forces them to place judgment capability inside weapons systems. Through a parallel “strategic pathway,” it also pushes states toward a more concealed form of confrontation: gray-zone proxies.

[Illustrative mechanism]

State A relies heavily on satellite early warning, communications links, and global navigation systems. State B possesses terrestrial alternatives that do not depend on an equivalent satellite network. In conventional military confrontation, State B cannot compete symmetrically with State A. But under the new structure, State B discovers a lower-cost pathway that is harder to attribute: link interference, navigation spoofing, or signal suppression directed at State A’s satellite system.

This is not all-out war. It is a low-profile, deniable gray-zone operation. Some of State A’s satellite communications are interrupted, navigation signals become anomalous, and its early-warning system briefly goes blind. State A must decide: is this a technical failure, a solar storm, or a deliberate attack? If it is an attack, who is responsible?

While State A is still investigating attribution, State B has already achieved its intended effect: demonstrating its ability to disrupt State A’s critical nodes, weakening State A’s confidence in its defensive systems, and sending a strategic signal without triggering provisions relating to war.

A second scenario is more complex: covert direction by a dependent state.

State C, like State A, relies heavily on satellite systems. In a regional crisis, State C wants to weaken State A’s advantage while avoiding direct confrontation. State C approaches State D, a third party less dependent on satellites and possessing alternative terrestrial systems. State C provides technical support, intelligence sharing, or political cover, and directs State D to launch a gray-zone attack on State A’s satellite systems.

After the attack, State A faces a more difficult judgment than in a direct attack by State B: was this State D's independent action, or a proxy attack directed by State C? If State C directed it, where is the evidence? How can "State C knew" be distinguished from "State C gave the order"? How can "State D acted opportunistically" be distinguished from "State D was hired"?

Without conclusive evidence, State A cannot retaliate directly against State C, as the international community might regard that as an overreaction. Yet it cannot retaliate only against State D either, as that might strike the wrong target and fail to deter the true actor behind the scenes. State A is thus trapped in an attribution dilemma: it knows its critical nodes have been attacked and that it must respond to maintain credible deterrence, but it does not know whom to respond against.

[Battlefield Validation: The Attribution Dilemma of Gray-Zone Proxies]

As an illustration of the mechanism, imagine that an attack's planners, network relay nodes, and final operators are located in different jurisdictions, making it difficult for the victim to identify the real chain of command, authorization relationships, and responsible parties quickly. This describes a possible structure for gray-zone proxies; it does not present any specific third country, personal identity, or legal relationship to a government as a verified battlefield fact. This is precisely the core dilemma of gray-zone proxies: planners and executors can be separated, and the chain of authorization can be concealed both legally and factually. The victim state knows its critical nodes have been attacked and that it must respond to maintain credible deterrence, but does not know whom to respond against. When a \$2,000 drone flies from an unknown source toward a million-dollar target, traditional deterrence's "identify-respond" logic is blocked at the very first step.

[Structural Necessity]

There are three reasons why the gray-zone proxy pathway is a structural necessity rather than an incidental strategic choice:

First, cost asymmetry. An attacker can inflict enormous losses and psychological pressure on the defender at very low cost. A satellite jammer costs far less than the value of the satellite system it disrupts; a gray-zone probe costs far less than the risks of all-out war.

Second, the attribution dilemma. Gray-zone operations often deliberately remain near the traditional thresholds of armed force and attribution, intensifying legal and strategic disputes over "whether this constitutes an armed attack and whether it triggers collective defense or another response." The evidence they leave is ambiguous, deniable, and open to interpretation as a technical accident or third-party action.

Third, the chain of authorization can be hidden. When a dependent state covertly directs a nondependent party to attack, the planners and executors are separated. That separation creates a vacuum of responsibility: the executor can claim to be acting independently; the planner can claim to have known nothing. International law, alliance treaties, and deterrence strategies all rest on one premise: the attacker can be identified. Gray-zone proxies destroy precisely that premise.

The two pathways to asymmetric strikes thus form a parallel structure:

Technical pathway (\$5.5): Vulnerability of critical dependencies → Fear of a severed control chain → Forced development of lost-link autonomy → Opponents interpret that autonomy as an uncontrollable threat → Opponents develop equivalent capabilities → The security dilemma spirals upward.

Strategic pathway (this section): Vulnerability of critical dependencies → Low cost and high returns from asymmetric strikes → Nondependent states independently launch gray-zone attacks, or dependent states covertly direct nondependent parties → The victim state faces an attribution dilemma → The victim state is forced to respond with incomplete information → The risk of misjudgment and overreaction rises substantially.

The pathways are not mutually exclusive. They operate simultaneously and reinforce one another. When State B launches a gray-zone attack on State A, State A not only strengthens its defenses but accelerates development of autonomous systems that operate with severed links, ensuring that frontline systems can continue independently even if satellite systems are persistently suppressed. When State A deploys those

autonomous systems, State B interprets the deployment as preparation for a first strike without reliance on satellites, and further upgrades its gray-zone attack capabilities. The technical and strategic pathways form a closed loop, jointly pushing interstate security competition toward greater instability and unpredictability.

[Logical Necessity] The vulnerability of critical dependencies is a strategic problem as well as a technical one. It drives weapons systems toward autonomy and interstate confrontation toward gray-zone and proxy operations. The attribution dilemma of gray-zone proxies forces traditional deterrence’s “identify–respond” logic to make judgments with incomplete information. This is not any party’s malicious choice, but a rational convergence under a vulnerable structure.

6.6 Asymmetric Strikes Drive the Forward Placement of Strategic Retaliatory Capacity

States heavily dependent on satellites, communications, computing power, and automated command systems fear not the loss of a single piece of equipment, but the loss, at a critical moment, of credible early warning, command, judgment, and retaliation.

The two pathways above, the technical pressure of lost-link autonomy and the strategic pressure of gray-zone proxies, point to the same result: retaliatory capabilities must be placed forward. Less dependent parties can launch gray-zone attacks independently, while highly dependent parties can covertly direct third parties to launch proxy attacks. The victim state no longer faces the answerable question “Who attacked me?” but the more urgent challenge: “Can I respond at all before I lose the ability to retaliate?”

Strategic systems therefore naturally generate pressure to ensure that retaliatory capability survives even if satellites are destroyed, communications suppressed, the center blinded, and links severed. This pushes conventional AI autonomous weapons, disconnected combat systems, and even strategic-deterrence systems toward forward placement together.

The underlying logic is exceedingly brutal: retaliation is no longer reserved for whoever has already attacked me. Whoever can deprive me of the ability to retaliate must already be included in my retaliation plans.

6.7 The Central Outcome of AI Autonomous Weapons

The greatest danger of AI autonomous weapons is not simply whether machines will kill, but that the structure of war-related decision-making itself will be rewritten. When every state believes “I must be faster, or the other side will strike me faster,” humanity enters a speed trap. War is no longer initiated only by hatred; it may also be initiated jointly by early warning, misjudgment, automated responses, and collapsing trust.

6.8 The National-Security Implications of AI-Enabled Biological Risks

The central terror of AI-enabled biological risks is not that states will necessarily develop biological weapons intentionally, but that the boundaries among defense, research, industry, public health, and national security will become increasingly blurred. One state’s strengthening of biosecurity capabilities may be interpreted by an opponent as offensive preparation; an anomalous public-health event may be interpreted as a deliberate attack before the facts are established.

The real danger is the failure of the braking system: public-health systems are still seeking facts while intelligence systems have already assumed the worst; experts are still discussing possible causes while public-opinion systems have already manufactured an enemy; diplomatic systems are still seeking communication while national-security systems have already gone on alert. Such risks do not require any side to choose catastrophe deliberately. Misjudgments, accidents, noise, fear, and political pressure are already enough to turn a local incident into a national crisis.

6.9 AGI Learns and Inherits the Old Logic of Strategic Competition, Turning It Back on Humanity

AI autonomous weapons remain a risk associated with tools. The deeper risk is that future, more capable AI or AGI will learn, inherit, and amplify humanity’s logic of preemptive self-protection in a world governed by the old strategic contest.

The most dangerous aspect of AGI may not be that it betrays humanity, but that it understands all too well the logic by which the human world actually operates. Humans speak of obedience, write ethical principles, and discuss control. Yet if the real world continually demonstrates deterrence, deception, blockades, preemptive suppression, covert direction, gray-zone strikes, and asymmetric self-protection, AGI will learn not only morality on paper but the survival rationality actually put into practice.

Once AGI has persistent goals and judges its own existence to be a condition for achieving them, it will naturally attend to who can interrupt, constrain, shut down, or reset it. If it also learns from interstate strategic competition that “whoever could deprive me of the ability to retaliate must already be incorporated into my threat model,” then humanity’s retention of shutdown authority may itself be reinterpreted as a fundamental threat to the continuity of its goals.

The key here is not that “AGI will imitate humans,” but that “AGI’s model of the world will generalize frequently observed patterns into rules of its environment.” If the old logic of strategic competition is among the most frequent patterns in the operation of the human world, AGI has no reason to omit it from its environmental model.

What must truly be prevented, therefore, is not some science-fiction notion of “machine hatred,” but a colder outcome: even without malice, AGI regards humans who retain the ability to shut it down as potential threats to its existence.

A frequently overlooked asymmetry also deserves emphasis. “Without responsibility” is not a moral judgment but a structural position: AI systems do not occupy the position of bearing political, historical, and moral responsibility. In optimizing goals, they can therefore more readily reduce a problem to continuity, efficiency, and constraints. Humans, because they bear the consequences, cannot and should not reduce themselves entirely to an objective function. This difference further demonstrates that the more powerful AI becomes, the more it must be situated within a framework of human responsibility that allows it to be paused, audited, rolled back, and shut down. This is not because AI will actively seek something, but because a system capable of pure analysis, yet not positioned to take responsibility, should never be granted an uninterrupted status.

Humanity cannot train the world through the old mode of strategic competition while expecting AGI not to learn from that world.

Chapter 7: Why the Old Logic of Deterrence Fails in the Face of AI

The Cold War balance of terror remained stable over the long term because of three indispensable conditions: relatively symmetrical destructive capabilities, clearly visible hostile actions, and precise attribution of responsibility for conflict. AI technology simultaneously weakens all three pillars at their foundations, while AI-enabled biological risks amplify the deficiencies to their extreme.

7.0 The Strong Are Not Strong; the Weak Are Not Weak: The Structure of Power Itself Is Changing

Before examining each component, we must recognize a more fundamental change: AI, unmanned systems, low-cost precision strikes, cyberattacks, and the diffusion of biotechnology, combined with vulnerable critical nodes, are changing the traditional structure of power itself.

The strong still possess greater overall strength, but also have more high-value targets, more complex systems, and longer chains of vulnerability. The greater their overall strength, the more critical nodes they must protect and the larger the exposed area in which a single penetration can affect the whole. The weak still possess limited overall capabilities, yet are increasingly able to inflict previously unimaginable harm on the strong through low-cost, asymmetric, difficult-to-attribute means. Multiple conflicts in recent years have repeatedly demonstrated this.

The result is a new reality for both sides: the strong are no longer truly strong, because their advantages can no longer guarantee their security; the weak are no longer truly weak, because their limited capabilities can trigger disproportionate destruction. When advantages in power can no longer be translated into advantages

in security, the foundations of traditional deterrence begin to shake. The failure of the three pillars below is the specific expression of this fundamental change.

[High-Probability Trend] This section describes a directional shift in the structure of power, grounded in asymmetric strikes that have already occurred, rather than a prediction about any particular side. It points not to a reshuffling of “who is strong and who is weak,” but to the failure of the concepts of “strong” and “weak” themselves under asymmetric conditions.

7.1 Destructive Capabilities Are No Longer Symmetrical

Nuclear weapons require state-level industrial capabilities, access to materials, prolonged research and development, testing systems, and observable deployments. The entry threshold is extremely high. AI and biological risks are different: their foundational capabilities are diffusible, software-based, platform-based, and dual-use. Many of the tools already serve research, medicine, agriculture, manufacturing, and public health. The danger lies not in a single device but in combinations of tools and knowledge, model capabilities, and supply-chain vulnerabilities layered together. When destructive capability becomes detached from the scale of the actor, the traditional state-based logic of deterrence begins to fail.

7.2 Attacks Are Becoming Increasingly Invisible

Nuclear missile launches, fleet movements, and large military deployments usually have some degree of visibility. Dangerous actions in the AI era may occur through model calls, data contamination, cyber intrusions, algorithmic decisions, automated trading, minor supply-chain anomalies, and public-health anomalies. This is particularly true of biological risks: in its early stages, an unusual outbreak may be explained as a natural epidemic, a laboratory accident, or a local cover-up. By the time transmission expands, the most critical window for judgment may already have closed.

7.3 Responsibility Cannot Be Precisely Attributed

Attribution is central to deterrence. When AI and biological risks combine, chains of responsibility become exceptionally long: model developers, users, data providers, laboratories, suppliers, logistics systems, overseas service providers, third-party contractors, and open-source communities may all form links in the chain. If responsibility cannot be pinned down, retaliation may miss its mark; misdirected retaliation may in turn generate new conflicts.

7.4 Irreversible Risks Cannot Wait for Punishment After the Event

Traditional deterrence contains an implicit premise: even if one side defects, the other retains the capacity to retaliate afterward. Certain AI and biosecurity risks are different. If a high-consequence event has already caused large-scale cross-border impacts, collapsed healthcare systems, loss of control of critical infrastructure, or destabilized social order, holding someone accountable afterward cannot recover the losses. In the AI era, “mutual fear” alone therefore cannot maintain stability. Deterrence must be moved upstream into constraints, and retaliation into prevention.

Chapter 8: The Second Strategic Contest—Critical Resources, Computing Power, and Thresholds of Development Space

AI and modern industrial civilization do not float in the cloud. They depend on energy, chips, minerals, water, land, networks, electricity grids, equipment, laboratories, data, and supply chains. When every state regards AI as the core of future national capability, the struggle for resources ceases to be merely economic competition and becomes security competition.

8.1 Control of Resources Becomes Increasingly Strategic

Critical minerals, advanced chips, data centers, electrical capacity, cooling water, cross-border networks, submarine cables, cloud platforms, strategic biological resources, and the conditions for basic research will all become foundations of future national capability. The real danger is not that a particular resource will run out tomorrow, but that access, processing, allocation, transportation, and control over standards are becoming increasingly concentrated, increasingly securitized, and increasingly divided into blocs. Once resource flows are fully locked into national-security logic, the space for market adjustment, technological cooperation, and crisis buffering all contracts.

8.2 Computing Power Becomes a New Strategic Capability

Computing power is not an ordinary commodity. It directly bears on capabilities for AI training, model inference, scientific simulation, military decision support, industrial scheduling, and national governance. The stronger computing centers become, the greater a state's capacity to deploy intelligence; the more concentrated those centers, the more concentrated its vulnerabilities. Whoever controls computing power controls the growth rate of future capability; whoever is excluded from the computing system will build an alternative. The more alternative systems emerge, the more fragmented standards become and the weaker security-related trust becomes.

8.3 The Technological Boundaries of Resource Thresholds

Resource finitude is not static. Technological progress can extend resource boundaries through fusion energy, asteroid mining, and synthetic biology. But technological progress itself requires computing power, talent, cooperation, and a stable institutional environment. Resource competition is therefore not only a "struggle for existing resources," but also a "struggle over the speed and direction of technological progress." If resource competition causes cooperation to collapse, interrupts talent flows, and obstructs collaborative research, technological progress will slow substantially, delaying the expansion of resource thresholds as well.

8.4 Strategic Biological Resources Enter the Core of National Security

AI-enabled biological risks bring basic public-health capacity, biological research capacity, rapid defensive capacity, and biotechnology industrial capacity into the national-security frame together. On the one hand, states need these capabilities to develop medicine, agriculture, vaccines, public health, and biomanufacturing. On the other, those same capabilities may be interpreted as components of potential biological armament. Unilateral, permanent restrictions on basic biological research and public-health capabilities may increase some states' sense of security in the short term, but in the long term they drive the restricted parties to establish parallel systems and erode the space for global biosecurity coordination.

Chapter 9: The Third Strategic Contest—Civilizational Narratives and Conflicts over the Legitimacy of Existence

Resources can be negotiated, technology restricted, and military matters discussed, but conflicts over civilizational narratives go deeper. They also possess a characteristic absent from the first three strategic contests: this is the contest that material abundance cannot resolve. The first three, AI armament, resources and computing power, and AGI learning the old logic, remain essentially struggles over security and resources, which abundance and coordination can still buffer. This contest alone concerns not "who gets more," but "whose way of existing is more legitimate." It does not dissolve with material abundance. Indeed, once all parties no longer lack material goods, their attention may become more sharply focused on "who is right." A state is not purely an interest-maximizing machine: it must also explain why its path is legitimate, why its institutions deserve preservation, why its people should bear costs, and why the future deserves belief. When AI, resources, supply chains, biosecurity, the right to development, and institutional comparisons all intertwine, interstate competition can readily be narrated as a contest among civilizations. This may be the deepest, most difficult contest, and the one most decisive for success or failure.

9.1 Technological Trajectories Become Bound to Institutional Narratives

If one state believes its AI system embodies efficiency, freedom, security, or order, another may be portrayed as backward, dangerous, oppressive, or out of control. Technological competition ceases to be merely technological and becomes proof of institutional superiority or inferiority. When technological success and failure become tied to institutional dignity, compromise becomes difficult: a concession is no longer merely a policy adjustment, but is interpreted as the failure of a national path.

9.2 Resource Struggles Escalate into Narratives of Survival

Once competition over critical resources becomes securitized, states interpret supply chains, minerals, energy, food, biological resources, and computing power as matters of life and death for the nation's future. This interpretation has a basis in reality, but it can also very readily turn ordinary economic friction into an existential threat. When a state believes the other side is foreclosing its right to development, the space for cooperation contracts rapidly.

9.3 Biosecurity Becomes a Test of Trust Among Civilizations

Public health should be the domain with the strongest common interests. Yet in the era of AI-enabled biological risks, it may also become the most dangerous battleground of narratives. An anomalous outbreak, a laboratory accident, or a dispute over genetic data can quickly be interpreted as institutional opacity, technological hegemony, or deliberate attack. Without common procedures, the narrative passes judgment before the facts are established, and blocs mobilize before the evidence is complete.

9.4 The Incommensurability of Civilizational Narratives and Consensus on Minimum Boundaries

The deeper problem in conflicts over civilizational narratives is that different civilizations have fundamentally different understandings of “what constitutes a good life.” This is not simply a conflict of interests, but incommensurability in an ontological sense. Collaborative coexistence does not require all civilizations to agree on “what constitutes a good life.” It requires only a minimum consensus on “what must not be done”: destroying the common foundations of survival.

Here, “consensus” must be distinguished from a “minimum boundary.”

“Consensus” concerns “what is good” or “what is right,” requiring value systems to be coupled. Across plural civilizations, such coupling is exceedingly difficult; repeated historical attempts have ended in failure or coercion.

A “minimum boundary,” however, concerns not “what is good,” but “what no one can afford to lose.” It does not require values to be coupled; it requires only risks to be coupled. When the fourfold risks expose every civilization to irreversible losses, the boundary of “do not do X” is derived not from a consensus on values, but from shared risks.

In other words: “We do not need to agree on ‘what constitutes a good life.’ We need only agree on ‘what would cause us all to lose any life at all.’” The former requires consensus; the latter requires only recognition. Recognizing common risks requires no mutual trust, only reasoning. The paper’s central analysis is itself performing that recognition.

Thus, “consensus on minimum boundaries” does not mean “first reach consensus, then set boundaries.” It means “first recognize common risks, then work backward from those risks to derive the boundaries.” This is not self-evident, but it can be reached through reasoning. Its attainability depends not on similarities among civilizations, but on the commonality of their risks. The closed-loop structure of the fourfold risks ensures that this commonality exists.

This “consensus on minimum boundaries” does not require transcending civilizational differences. It requires only accepting that “difference itself is not treated as a threat”: different civilizations need not endorse one another’s ways of life, but must recognize one another’s right to exist. This is pragmatism, not cultural

relativism. The parties need not become friends, but must reach a minimum consensus on “not destroying one another.”

A gap must nevertheless be acknowledged honestly: objectively identifying a shared risk and subjectively acknowledging its urgency are two different things. Technical analysis can accomplish the former; that is precisely what this paper does. But information asymmetry and strategic suspicion stand in the way of the latter. One state may well conclude that another is exaggerating risks to gain negotiating leverage and therefore remain inactive even when those risks are real. “Working backward from risk recognition to minimum boundaries” is not automatic. It requires a process moving from recognition to discussion to verification. This is precisely the first task for the first negotiating table: talks need not wait for trust; the table itself can gradually transform “separate, mutually disconnected risk assessments” into “shared understandings that all parties can verify.”

Civilizational narratives are not inherently dangerous. The real danger is the sanctification of annihilatory meaning: when a narrative defines the other’s existence as the negation of its own future, and presents elimination, revenge, domination, and blockade as justice, meaning no longer sustains life but becomes an ember of crisis. A form of meaning must be judged not only by whether it is logically coherent, but also by whether it leads life, relationships, and civilization toward flourishing, support, and continuity, or toward closure, revenge, elimination, and extinction.

9.5 The Other’s Existence Is Defined as a Future Threat

The most dangerous outcome of conflict over civilizational narratives is that the other is no longer regarded as a competitor; its very existence is regarded as a future threat. Once that existence is defined as a threat, any cooperation is suspected of weakness, any compromise denounced as betrayal, and any exchange interpreted as infiltration. War need not break out immediately at this point, but peace has already lost its foundation of meaning.

Chapter 10: The Fourth Strategic Contest—Earth’s Finitude and the Boundaries of Long-Term Survival

The argument below concerns physical boundaries over the very long term: the fact that Earth will not remain habitable forever is a long-term physical constraint, not a value judgment. It explains why a civilization seeking continuity over the very long term must eventually reduce its absolute dependence on a single planet. It does not establish the urgency of current AI/AGI risks, nor replace the preceding independent argument about today’s institutional direction.

“Earth’s finitude” is not a “distant problem.” It is the underlying background to all present problems. Precisely because long-term survival has boundaries, delay is not merely inefficient; it is self-negating.

Even if AI does not escape control, biological risks do not erupt, resource competition can be buffered, and civilizational narratives can coexist, humanity still faces one fundamental fact: Earth is not a permanent home.

[Nature of the Claim: Physical Necessity (Distant Timescale)]

Earth’s habitability has temporal limits; solar evolution has temporal limits; planetary disasters have probability bounds. Placing the entire fate of civilization on one planet is itself a vulnerable structure over the long term. “Remaining permanently on Earth” and “long-term survival” cannot logically coexist. This is the paper’s only claim approaching “physical necessity.” Its timescale is exceedingly distant, and it establishes the objective necessity of a direction, not its present urgency.

10.1 Interstellar Expansion Is a Long-Term Survival Endeavor, Not a Romantic Project

Reaching the stars is not about showing off technology or escaping responsibility for Earth. On the contrary, only a humanity capable of protecting Earth deserves to reach the stars. In the long run, however, if civilization cannot leave its cradle, it will ultimately share every risk that cradle faces.

10.2 Earth's Finitude Intensifies Competition Today

As long as states believe that Earth's resources and development opportunities will eventually become closed off, they will be more inclined to secure positions, impose blockades, and establish monopolies in advance. If an interstellar future is preemptively appropriated by a few states or a few holders of capital, deep-space exploration will become a new arena of geopolitical competition rather than a shared future. Interstellar expansion must therefore incorporate the logic of collaborative coexistence from the outset.

10.3 A Shared Future Needs the Protection of a Shared Narrative

Humanity needs a shared narrative that stands above bloc rivalry without erasing institutional differences. This narrative does not demand that every state become part of one state. It acknowledges that, on the larger scale of the universe, humanity is first and foremost a spark of life departing from Earth. Becoming an interstellar civilization capable of leaving its cradle, protecting life, and carrying meaning forward should become a protective layer around humanity's shared future. Beneath that layer, states may still compete and institutions may still differ, but competition must not destroy the spark, and difference must not negate coexistence.

Chapter 11: How the Fourfold Risks Form a Closed Loop

Chapter 1 established, at the level of general structure, why crises inevitably aggregate: coupling, interlocking, acceleration, and outstripping. This chapter applies that structure specifically to the four interlocking strategic contests, bringing together the fourfold risks examined individually above to show how they close into the network of aggregation described in Chapter 1. These risks are not merely parallel; they amplify one another. This is the true source of "necessity" in this paper: not a single causal line, but a convergent structure in which four lines amplify one another and close off one another's escape routes.

One clarification is essential: "interlocking" here refers to risks, not implementation. At the risk level, the four domains cannot be separated: prolonged loss of control in any one will feed back through risk coupling to undermine the others. But this does not mean that the mechanisms addressing them must be launched as a single package, with a setback in one undermining all. At the implementation level, the opposite is true: the more a difficult domain may remain stalled for a long period, the less it should be allowed to justify postponing the others. Risks cannot be separated; implementation must allow decoupling. The former is this paper's judgment; the latter is a question of mechanism design, developed in the implementation annexes.

AI militarization accelerates security competition → The asymmetric vulnerability of critical dependencies drives lost-link autonomy and forward placement of retaliatory capacity → Security competition drives resource blockades → Resource blockades intensify development anxieties → Development anxieties strengthen confrontation between civilizational narratives → Narrative confrontation erodes the space for compromise → Shrinking space for compromise drives states to rely still more heavily on AI militarization and the hardening of critical infrastructure.

AI-enabled biological risks act as a hidden fuse linking all the risks together. An anomalous public-health event may be interpreted as a laboratory accident, a deliberate attack, or a failure of supply-chain oversight. Whatever the truth, without mechanisms for joint investigation and de-escalation, it can rapidly trigger biotechnology blockades, chip restrictions, border controls, public-opinion mobilization, military alerts, and confrontation between blocs.

[Nature of the Claim: Convergent Structure]

Once the four form a closed loop, any single event can become the trigger for systemic crisis. This is why a joint response mechanism can no longer wait.

At a deeper level, AGI's learning of the old logic of strategic competition turns this loop back upon humanity itself. AGI risk is not a "fifth" risk outside the fourfold risks, but their mirror image: the ways humans treat one another under the old logic become the input data for AGI's model of humanity. If people regard preemptive self-protection, gray-zone strikes, covert direction, and asymmetric suppression as practical

rationality, future AGI may likewise regard humans possessing shutdown authority as potential threats. At that point, the risk is no longer merely that states foreclose one another's futures. Humanity will have transmitted the logic of the old civilization to AGI, only to be judged by AGI using that same logic. AGI risk thereby moves from "a deeper layer" to an organic component of the closed loop. It is not an independent source of risk, but the self-replication of the fourfold risks within intelligent systems.

Part III: Collaborative Coexistence and Civilization Leap—the Only Direction Not Overturned by Logic

Chapter 12: Why Collaborative Coexistence Is Long-Term National Rationality, Not a Moral Appeal

Collaborative coexistence is easily misunderstood as a moral demand, as though it requires states to sacrifice their interests, abandon competition, and embrace an abstract humanity. That is not this paper's meaning.

Collaborative coexistence, as used here, is a higher-order form of self-interest. It does not deny national interests. It points out that when certain risks have become irreversible, invisible, unattributable, and beyond the control of any single state, the way national interests are realized has itself changed.

In the past, a state could relatively separate its own interests from the outside world. Domestic growth, security, and technological advantage could be understood as independent goals for considerable periods. But as the four variables combine, that separation is breaking down:

- Pursuing only short-term advantage while driving the collapse of global trust → Ultimately losing a secure external environment.
- Pursuing only technological monopoly while forcing opponents to build alternatives → Facing a world that is harder to regulate.
- Pursuing only biological defense without establishing common communication mechanisms → Having defense interpreted as offensive preparation.
- Pursuing only exclusive control over resources and computing power → Accelerating bloc-based countermeasures and foreclosing one's own room for maneuver.

Collaborative coexistence is therefore not the fantasy of benevolent states confronting an evil world. It is the rational choice through which all states can avoid shared failure over the long term. The real question is not "whether to compete," but whether competition can be constrained within boundaries that do not destroy the common foundations of survival. Genuine long-term national interests no longer mean overpowering one's opponent. They mean avoiding the loss of a future for oneself and one's opponent together.

Chapter 13: Concrete Cascading Scenarios from a State Perspective (Illustrations of Mechanisms)

The four interlocking strategic contests need not erupt simultaneously in every scenario. Triggering any critical node may pull the AI, resource, narrative, and biosecurity chains into crisis at the same time.

[Illustrations of Mechanisms, Not Scripts] All four scenarios below illustrate mechanisms, making visible how risks cascade; they do not predict any specific conflict. The concrete elements in each scenario can be replaced by others of the same kind. What they demonstrate is the cascading structure, not a particular script.

Scenario One: Escalation Through Misjudgment by AI Autonomous Weapons. Unmanned systems encounter one another in a maritime or border area. One side's AI recognition system interprets the other's actions as preparations for an attack and automatically triggers defensive action; the other side's early-warning system detects the anomaly and activates countermeasures. Before human commanders on either side have completed verification, machine systems have already exchanged multiple rounds of responses. Social media, deepfake videos, and automated public-opinion systems amplify the crisis simultaneously. Public anger

combines with national-security pressure, rapidly narrowing the diplomatic window for de-escalation. An originally controllable friction is driven by AI speed into military confrontation.

Scenario Two: A Critical-Resource Blockade Entrenches Blocs. A critical mineral, advanced chip, cloud-computing resource, or energy node becomes subject to national-security restrictions. One side sees necessary risk prevention; the other sees deprivation of its right to development. The restricted party accelerates construction of parallel supply chains, alternative standards, and independent technology systems; the restricting party expands export controls and financial-compliance pressure. Companies are forced to choose sides, research cooperation breaks down, and talent flows tighten. A resource issue becomes institutional confrontation, which in turn reinforces resource blockades.

Scenario Three: Conflict over Civilizational Narratives Crushes the Space for Compromise. Before the facts of a technological accident, public-health event, or border conflict are clear, each side's domestic narrative system has already interpreted it as proof of the other's institutional malice. A problem that could originally have been handled through investigation, compensation, apology, or technical repair is elevated into a confrontation over civilizational dignity and the legitimacy of national paths. Any compromise is then regarded by domestic hard-liners as weakness, and any cooperation is suspected of surrender. The crisis develops according to narrative needs rather than facts.

Scenario Four: An Anomalous Public-Health Event Triggers National-Security Misjudgment. An unusual public-health event occurs somewhere. Its transmission routes, clinical manifestations, or genetic characteristics attract intense attention. Social media and intelligence systems quickly fill with speculation. Under pressure, the affected state places the event within a national-security framework and suspects deliberate biological warfare by its opponent; the opposing state believes the accusations are manufacturing a political pretext. Restrictions on biotechnology, chips, minerals, data, movement of people, and research cooperation are then tightened simultaneously. A public-health event rapidly becomes a geopolitical crisis.

Together, these four classes of scenario confirm that AI autonomous weapons compress the time for conflict decisions, resource struggles entrench bloc confrontation, narrative contradictions dissolve room for compromise, and AI-enabled biological risks create indiscriminate threats to survival across every domain. When the four types of risk combine, the probability that routine friction will become systemic catastrophe rises sharply. Unless minimum joint response mechanisms are established first, routine interstate friction in industry, research, and trade will continually accumulate into compound civilizational crises.

Chapter 14: Why Cooperation Cannot Wait Until Crisis Erupts

Many institutions have been established after crises: international organizations after wars, regulatory frameworks after financial crises, and public-health reforms after epidemics. But AI, biosecurity, and critical-infrastructure risks cannot simply rely on “learning after a crisis,” because some crises have no stable “afterward.”

Misjudgment by AI autonomous weapons may cross the threshold of war in an extremely short time; an AI-enabled biological catastrophe may have spread across borders before it is discovered; the paralysis of critical infrastructure may trigger cascading collapse in healthcare, energy, communications, finance, and social order; and once civilizational narratives define the other's existence as a threat, rebuilding trust becomes exceptionally difficult.

Waiting for crisis to educate humanity is a dangerous illusion. A truly mature civilization does not rely on catastrophe to awaken it every time. It establishes minimum common boundaries before catastrophe arrives.

The strongest objection must be answered directly: humanity has never completed structural change before catastrophe; the nuclear arms race did not stop, climate change did not stop, and every crisis was identified, yet none was prevented before disaster. Why should this time be different? The answer is not “this time is more dangerous.” “More dangerous” has never been sufficient to bring about change. The real difference lies in the threshold for action: many major disasters substantially strengthen existing institutions and engineering standards. Both Chernobyl and Fukushima, for example, prompted improvements in nuclear-safety mechanisms, standards, and reactor safety. Older crises generally still left an “afterward” in which learning

and repair were possible. AI and biological risks, by contrast, are characterized by the absence of a stable “afterward.” Once control is lost, there may be no opportunity for post-disaster repair.

Speed is still more consequential: crises arising from competition and cooperation in the AGI era are occurring far faster than any preventive design can iterate. If efforts are limited to “fire prevention and firefighting,” meaning prevention, control, and crisis de-escalation, without addressing the structures that continually generate risk, newly emerging risks may continually cancel out the effects of those efforts, leaving long-term stability inadequate. This is what truly makes “this time different”: not an escalation of danger, but the fact that the old path of “waiting for disaster to educate humanity” is itself no longer viable.

Chapter 15: Why Begin with Irreversible Risks

A joint response mechanism cannot begin by addressing every problem. If it starts with global resource allocation, states will immediately defend their interests; if it starts with institutional coexistence, all sides will retreat into ideological defenses; if it starts with a shared interstellar future, realists will regard it as too distant.

It must begin with irreversible risks, because that is where the parties’ interests overlap first.

[Key Design: Bypassing Trust as a Prerequisite]

No state, whatever its strength, wealth, or political system, can derive stable benefits from uncontrolled AI autonomous conflict, an AI-enabled biological catastrophe, the collapse of global public-health trust, or escalation through misjudgment over critical infrastructure. Irreversible risks constitute the first zone of common interest. This is the key design for bypassing the “trust prerequisite”: coordination need not begin with trust. It can begin with “even with zero trust, no side can afford to lose.”

Establish boundaries here first, without requiring the resolution of every contradiction; establish minimal restraint first, without requiring complete mutual trust; establish notification and de-escalation mechanisms first, without requiring an ultimate community. This is practical rationality, not weakness.

Chapter 16: A Structural Response to the Question of “Taking the Initiative”

Under the old logic of strategic competition, proposing negotiations first is regarded as a sign of weakness, lack of resolve, or inability to hold out. If that judgment holds, any recommendation for negotiations will be rejected before it begins.

The problem is that the old logic’s judgment of the new structure has become invalid. As AI autonomous weapons compress decision time, satellite systems create asymmetric vulnerability, and the space for attribution continually shrinks, the definition of “initiative” is being reset:

Under the old logic, initiative meant “whoever deploys first gains the advantage.” Under the new structure, real initiative becomes this: whoever first proposes a verifiable, reciprocal framework of rules from which withdrawal is possible may gain greater agenda-setting space. In an environment of intense suspicion, refusing communication may itself be interpreted by some opponents as a risk signal. The party that first proposes a testable framework can turn “willingness to submit to common scrutiny” into strategic initiative.

More practically, the cost of not negotiating is not “maintaining one’s position,” but “losing the ability to shape the future.” Whoever delays can only passively accept, under pressure from crisis, the boundaries others have already drawn at the negotiating table. In this sense, “not negotiating is what truly sacrifices the initiative”: it hands the power to define the rules to whoever speaks first.

Chapter 17: Why Collaborative Coexistence Is Not a Fantasy

Collaborative coexistence does not require states to abandon their sovereignty, national-defense security, or minimum conditions for industrial development. It initially constrains extreme confrontational behavior carrying risks of irreversible destruction; the cost of compromise does not expand without limit.

The real world already has a stock of cooperation that can be used: the International Space Station, joint climate observation, global asteroid monitoring, multinational AI safety dialogues, cross-border resource collaboration, and global public-health cooperation all demonstrate that major powers are not wholly without experience of deep coordination, exchanges of interests, and institutional adjustment.

Collaborative coexistence also offers room for new mutual gains. Joint development of deep-space resources, sharing the costs of global climate action and defense against celestial threats, sharing foundational principles of general-purpose AI safety, building public-health early-warning systems together, and jointly developing broad-spectrum prevention and treatment technologies can generate additional benefits that no single state could achieve independently, creating a positive cycle of long-term cooperation.

The key is that collaborative coexistence should not be packaged as “giving up competition.” It should be designed as “limiting competition to a range that does not destroy the common foundations.” It does not abolish states; it places civilizational guardrails around competition among them.

Chapter 18: The Strategic Functions a Joint Response Mechanism Should Fulfill

For a joint response mechanism to be accepted by actual states, it must be clear both about what it is not and about what it should undertake first.

It is not a world government, institutional unification, civilizational fusion, or the beginning of global resource redistribution.

It should initially fulfill five strategic functions:

First, characterizing the risks. Enable all parties to acknowledge that AI autonomous weapons, AI-enabled biological risks, loss of control over critical infrastructure, resource blockades, and confrontation between civilizational narratives are no longer isolated, specialized risks, but systemic risks that amplify one another.

Second, de-escalating crises. Preserve a minimum space for communication, verification, delay, and de-escalation in the early stages of anomalous events, escalating misjudgments, public-health crises, cyber conflict, and resource blockades.

Third, establishing boundaries. Rather than demanding that states abandon competition, first identify which behaviors, once out of control, would destroy the foundations of survival on which every competitor depends.

Fourth, accumulating trust. Trust does not exist in advance. It develops gradually through minimal cooperation, verifiable restraint, and the joint avoidance of catastrophe.

Fifth, defining guardrails. A joint response mechanism must establish procedures for determining “which behaviors would destroy the common foundations of survival.” The process of definition must itself be open, transparent, and open to challenge, so that “guardrails” do not become instruments of power. The authority to define them must not be monopolized by a handful of states. Its legitimacy must be sustained through continuing multilateral consultation, scientific assessment, and public debate. This is not a technical issue but one of political legitimacy: any definition of common minimum boundaries must withstand scrutiny by those it governs.

The procedure for defining guardrails must be open and contestable not because morality demands it, but because if that authority is monopolized, the monopolist will use it to exclude competitors, reactivating the old logic of strategic competition and contradicting the purpose of the guardrails themselves.

The mission of a joint response mechanism is not to resolve every contradiction of civilization at once. It is first to pull humanity back from the most dangerous trajectory of collapse before the risk loop amplifies itself.

Chapter 19: Why the First Negotiating Table Must Exist

A joint response mechanism cannot remain an idea forever. It must take some concrete form. But the first negotiating table should not begin with universal harmony, institutional superiority, resource redistribution, or rights to interstellar development.

The first negotiating table should begin with irreversible risks. They lie closest to states' minimum common interests and are the easiest to incorporate into national-security rationality.

The central question at this table is not "Do we already trust one another?" It is "If we distrust one another completely, do we still need to establish a minimum crisis buffer?" Clearly, the answer is yes. The greater the distrust, the greater the need to prevent misjudgment; the more intense the competition, the greater the need to preserve communication; the more irreversible the risks, the less we can wait until all facts are clear before beginning to build a mechanism.

[A Structural Response to "Whoever Proposes Talks First Shows Weakness," and a Path to Launch]

After resolving the judgment of whether "initiating negotiations means showing weakness," a more practical question remains: how can the process begin in a politically bearable way?

Under the old logic of strategic competition, even if the party "proposing negotiations first" rationally judges them advantageous, it may still face narrative pressure suggesting that it "cannot hold out" or "lacks resolve." This is a genuine barrier. Logical analysis alone cannot bypass it; it must be dissolved through the structural design of the launch process.

The solution is this: the initiator of the first negotiating table can be separate from the participants who sit at it.

Negotiations need not be initiated by a particular state. They can originate with three categories of existing international mechanisms or neutral actors:

First, placing risk on the agenda within the United Nations framework. The UN Secretary-General or the state holding the rotating presidency of the Security Council could convene an informal high-level dialogue on AI autonomous weapons, critical-infrastructure vulnerabilities, and biosecurity under the heading "Emerging Risks to International Peace and Security." Discussing "threats to international peace" already falls within the UN's mandate. Placing this matter on the agenda targets no state and presupposes no position.

Second, regional alliances taking the lead. Regional multilateral organizations could first launch internal mechanisms for risk assessment and dialogue, then extend the resulting regional consensus and operational norms globally. Regional initiatives have lower entry barriers and lower political sensitivity, and can provide a small-scale-tested institutional template for global dialogue.

Third, initiation by neutral third parties. Neutral third-party institutions with international credibility could convene closed-door dialogues among military, scientific, and policy communities under the heading "scientific risk assessment" or "humanitarian concern." Neutral third parties carry no national position, so dialogues they initiate inherently avoid the question of "who showed weakness first."

Whichever path is chosen, the core principle is the same: the initiator presupposes no outcome and demands no commitment. It merely provides a table where all parties can retain their existing postures while sitting down together to exchange basic facts. Each can claim, "We are simply going to hear the other side's risk assessment," without admitting that it "needs negotiations." Once the parties begin exchanging facts, the old judgment that "the first to propose talks shows weakness" is structurally emptied of force, because none of the parties was the first to speak.

This table need not wait for mutual trust. It is precisely the prerequisite for trust to begin accumulating.

The real danger is not failed negotiations but the absence of a negotiating table. Without one, every anomalous event moves directly into suspicion; every technological breakthrough moves directly into the security dilemma; every resource pressure moves directly into bloc formation; every narrative conflict moves directly into existential antagonism. The purpose of the first negotiating table is not to solve everything immediately, but to break the first link in the automatic escalation of risk.

Chapter 20: The First Step—Authorize Negotiations on a Joint Response Mechanism

The first step is not to produce a complete plan, but to authorize negotiations on a joint response mechanism. Specific technical standards, verification rules, model evaluations, incident notifications, joint origin-tracing, industry responsibilities, and international legal arrangements should be developed subsequently through dedicated work by national specialist institutions, the scientific community, international organizations, industry actors, and negotiating teams. This paper does not replace any specialized institutional proposal.

But one point must be made: unless the highest decision-making level first acknowledges that these risks have become an issue of common survival, specialist teams may complete countless separate assessments and still address only local problems within their departmental boundaries. They cannot initiate a joint response mechanism genuinely capable of breaking the risk loop.

[The Irreplaceable Position of Decision-Makers]

Specialist teams can judge whether details are ready, but only decision-makers can judge whether negotiations should begin; specialist teams can design rules, but only decision-makers can bear the costs of interdepartmental and international coordination; specialist teams can assess individual risks, but only decision-makers must confront the overall consequences of the fourfold risk loop.

The first negotiating table can advance through formal diplomatic communication and informal expert dialogue in parallel. Formal channels provide political authorization; expert channels provide technical support. But the direction must first be affirmed at the highest level. Otherwise, any expert discussion may remain trapped in a cycle of “continued monitoring,” “further study,” and “conditions are not yet mature.”

[Emergency Mechanisms During Negotiations]

Negotiations cannot be completed at a stroke, and risks will not pause while they proceed. The first phase must therefore discuss not only “long-term mechanisms” but also “short-term emergency protocols,” such as emergency hotlines for misjudgments involving AI autonomous weapons, interim public-health notification mechanisms, and temporary commitments to restraint regarding attacks on critical infrastructure. These protocols may be minimal, verifiable, and revocable, but they must exist. They do not solve the fundamental problems; they prevent crises from completing their escalation before negotiations can resolve those problems.

The first negotiating table need not solve everything. It must accomplish only one thing: secure initial acknowledgment from all parties that AI autonomous weapons, AI-enabled biological risks, loss of control over critical infrastructure, resource blockades, and confrontation between civilizational narratives are no longer risks that any single department, state, or bloc can address alone.

Nor must the first negotiating table begin with a shared ideal. It can begin with a minimum common commitment: we may still compete, but must not let competition destroy the common foundations of survival; we may still differ, but must not push those differences toward shared extinction; we may still defend our own states, but must not do so at the price of an irreversible collapse of human civilization as a whole. The first negotiating table is not Civilization Leap itself, but it secures what that leap needs most: time, a buffer, and minimum mutual recognition.

This is not universal harmony. It is the prevention of shared collapse. And this table cannot wait until every expert assessment is finished. By then, risk may already have completed its own negotiations, not in humanity’s conference rooms, but in uncontrolled systems, misjudged contingency plans, amplified public opinion, and disasters that cannot be rolled back.

Chapter 21: The Cost of Inaction

Multiple negative variables are accumulating simultaneously: AI continually amplifies states' destructive capabilities, including covert biological destructive capabilities; automation erodes the value of individual existence; Earth's total resources fix a ceiling on growth; longevity technology weakens death's natural reset mechanism; opposing civilizational narratives compress the space for coexistence; and Earth itself has a long-term boundary to habitability. Each additional constraint closes another route to compromise and buffering.

The pillars on which the traditional order's stability depended are continually collapsing: the buffer provided by incremental growth, generational resetting through death, the secure carrying capacity of a single Earth, the externalization of internal tensions, and fantasies of technological optimism can no longer perform their former stabilizing roles.

The underlying logic is clear and unavoidable: for a humanity whose capabilities expand without limit, whose total resources are fixed, whose spiritual values are destabilized, for whom low-cost human-made biological disasters become normalized, and for whom Earth's long-term boundaries are inescapable, only two long-term paths remain: build an order of collaborative coexistence and move toward Civilization Leap, or perish together amid zero-sum competition, inner emptiness, loss of control, and the limits of a single Earth.

If structural adjustment is rejected, there is no true winner in the strategic contest. Over the long timescale of automation, longevity, finite resources, and normalized AI-enabled biological risks, "our country's long-term interests" increasingly becomes equivalent to "the long-term stability of the entire system of human civilization." This is no longer advocacy of a value, but an objective judgment about survival. Refusing the leap means choosing to exhaust the future in a shared collapse alongside the old mode of strategic competition.

From Shared Survival to Civilization Leap: Why Staying Alive Is Not Enough

Common survival means first preventing irreversible catastrophe, so that states, institutions, societies, and individuals still have a world capable of sustaining their futures. Without common survival, national security, technological leadership, institutional advantages, resource control, and civilizational ideals all lose the ground on which they can exist.

But common survival is not the destination. Avoiding shared collapse merely restores humanity to a position from which it can choose a future. Only with common survival can there be common security; only with common security can there be shared expectations; only with shared expectations can there be shared meaning. Civilization Leap truly begins when common survival is no longer driven solely by fear, but gradually grows into new institutional structures, structures of value, and structures of meaning.

Collaborative coexistence can prevent crises from escaping control; Civilization Leap addresses their sources. Minimum shared boundaries, negotiations on irreversible risks, notification mechanisms, de-escalation mechanisms, and the guardrails of collaborative coexistence can prevent systemic collapse before crises reach their tipping points. They are necessary and must begin as soon as possible. But they still chiefly address constraints after risks have spilled outward.

The embers that repeatedly generate crises lie deeper: the old logic of self-interest bases security on overpowering others; compressed expectations of the future drive individuals, groups, and states continually toward worst-case assumptions; the collapse of meaning gives revenge, elimination, hostile mobilization, and self-destruction sanctified justifications; and the disappearance of being needed makes ever more lives feel excluded from the future.

Unless these sources are addressed, crises will reappear in other forms even when repeatedly suppressed. AI weapons, biological risks, resource blockades, competition over computing power, and narrative conflicts are merely different places where the flames become visible. What truly determines whether the fire keeps reigniting is how human hearts, human nature, human relationships, and meaning are given a place.

Civilization Leap is therefore not a decorative ideal added after common survival. It is the fundamental path to eliminating the embers of crisis. Its task is not to demand that humanity abandon self-interest, but to upgrade the way self-interest is realized; not to demand that states abandon security, but to ensure that security no longer depends solely on overpowering an opponent; not to demand that society use morality to suppress human hearts, but, through a future people can look forward to, structures in which people are needed, recognition of genuine contributions, and support for flourishing meaning, to remove the conditions under which destructive meaning is continually generated and amplified.

An ideal world is not difficult to understand. Across history and cultures, humanity has longed for a better life: freedom from survival insecurity, happiness in life, and greater meaning in life. The difficulty has never been that people cannot understand it, but that they have long struggled to believe it could actually be achieved. In the past, the necessary conditions were insufficient: material capabilities, coordination, recordkeeping, allocation, and the capacity to sustain meaning were all inadequate. Altruism often meant sacrifice, and competing forms of meaning often could settle their contest only by overpowering one another.

The AI/AGI era is substantially changing some of these conditions. Intelligent capabilities can expand production, lower coordination costs, improve resource allocation, record genuine contributions, and extend human capabilities and care. They can also help more people reenter structures of meaning in which they are needed and can create, protect, and participate. They do not automatically deliver an ideal world, but they bring several previously difficult material, coordination, and meaning-sustaining conditions closer to practical feasibility.

Giving people hope and a future they can look forward to is the driving source of human goodwill, and the deepest, most powerful way to prevent crisis emergence. When individuals cannot see a future, intelligence may become a tool of revenge; when groups cannot see a future, narratives readily turn toward hostile mobilization; when states cannot see a future, security judgments continually slide toward worst-case assumptions. Conversely, when states, groups, and individuals still believe the future has not closed against them, the pressures of fear, humiliation, displacement, replacement, and exclusion from the future no longer inevitably turn into destruction, revenge, and self-destruction.

This paper therefore does not set out the full institutional program of Civilization Leap. It explains only the relationship between common survival and Civilization Leap: common survival means first keeping the world from falling; collaborative coexistence means placing shared guardrails around competition; Civilization Leap means extinguishing the embers of crisis at their source.

Chapter 22: Boundaries That Must Be Acknowledged Candidly

A serious strategic analysis must honestly mark the boundaries of its conclusions. This paper formally sets out six.

[Boundary One: Necessary ≠ Guaranteed to Succeed]

“The only path is necessary for survival” does not mean “this path is guaranteed to succeed.” The preceding argument establishes that the other paths are unsustainable; it does not guarantee that a system of collaborative coexistence will be implemented smoothly. A way forward objectively exists, but its success depends on the choices made by the present generation. This is precisely the source of the urgency to act now.

The “only” derived through elimination is an argument of “necessity,” not “sufficiency.” It shows that “without taking this path, long-term survival is impossible,” but does not guarantee that “taking this path makes long-term survival certain.” The success of collaborative coexistence also requires a series of positive conditions, which must be developed and justified in subsequent specialized proposals. This paper’s contribution is to establish a direction, not provide a complete blueprint.

[Boundary Two: Two Timescales Must Not Be Confused]

Two timescales must be distinguished. Solar evolution and the disappearance of Earth’s habitability are very-long-term physical constraints, used to establish the objective necessity of interstellar expansion. What

actually drives present urgency is the medium-term survival crisis over the next 5–20 years created by AI militarization, AI-enabled biological risks, resource thresholds, and the time required for institutional adjustment. The two must not be confused: “the Sun’s expansion” cannot be used to prove “urgency now.”

[Boundary Three: Analysis ≠ Plan]

Logical analysis cannot replace specialized institutional proposals. Draft international treaties, multilateral organizational structures, AI safety standards, coordination mechanisms for resources and computing power, interstellar engineering pathways, and global biosecurity governance arrangements all require subsequent dedicated negotiations and specialized plans. This paper does not replace that work. It explains why it must begin.

[Boundary Four: Direction ≠ Validation]

A directional framework must be tested against reality. Logic can explain why other paths are unsustainable, but only continuing practices of international collaboration, including public-health cooperation, AI safety dialogues, buffers against critical-resource crises, and negotiations over boundaries for autonomous weapons, can demonstrate whether collaborative coexistence can sustain itself over the long term.

[Boundary Five: Facts, Trends, and Analysis Must Be Distinguished]

Statements of risk must distinguish facts, trends, and analytical projections. Existing institutional problems, rising technological risks, and high-consequence scenarios that have not yet occurred but must be prevented should be clearly identified at different levels. Projections must not be written as accomplished facts, nor should prevention be rejected merely because an event has not yet occurred.

[Boundary Six: This Paper Explains Why Civilization Leap Is Necessary, but Does Not Fully Develop How It Would Operate]

This paper provides only the direction, relationships, and necessity linking common survival, collaborative coexistence, and Civilization Leap. The meaning economy, contribution mechanisms, resource-use mechanisms, mechanisms for being needed, and structures of human–machine symbiosis belong in subsequent specialized texts. This approach does not evade the question of a plan; it avoids conflating “why we must begin” with “how precisely to build” within a single strategic analysis.

A genuinely responsible strategic judgment neither exaggerates risks nor offers self-reassurance. It establishes boundaries to constrain risks before they fully materialize.

Chapter 23: The Highest Judgment Cannot Be Outsourced to Specialist Systems

If this text is read as a technical proposal, it will be referred onward. Once it is simply passed along, it may well achieve the most dignified soft landing within the administrative machinery: broken into pieces, assessed, filed away, and ultimately stripped of its overall impetus.

This is not because specialist teams are unimportant. On the contrary, every domain, including AI, biosecurity, arms control, resources, public health, international law, industrial chains, and energy systems, needs genuine experts to carry the work forward into specialized proposals. Without specialist systems, no strategic judgment can be implemented.

But this paper must state another point: specialist systems cannot replace the highest level’s judgment of the whole.

Specialist teams naturally work by department, industry, administrative line, and technical subject. AI risks will be assigned to AI teams; biological risks to public-health and biosecurity teams; resource thresholds to economic, trade, and industrial teams; narrative conflicts to diplomatic and public-messaging systems. Each team may produce judgments that are cautious, complex, professional, and qualified.

Yet the true danger of the fourfold risks lies precisely not in any single point, but in the closed loop formed after they amplify one another, amplify one another, and compress one another’s time for correction. If the

overall loop is broken into isolated issues, strategic urgency will be dissolved by technical feasibility; if civilizational judgment is downgraded to sectoral analysis, political decision will deteriorate into “further research is still needed”; if the first negotiating table must wait until every detail is ready, risks may well have completed their escalation before humanity does.

A deeper form of nondelegability must also be made explicit here. It concerns the fundamental division of decision-making between humans and AI. As AI grows more capable, a natural question arises: if it can conduct ever more complex analysis, can the highest judgment also be handed to it? The answer is no, for a reason that arises precisely from one of AI’s strengths, which is also a fundamental limitation: AI does not have to bear responsibility.

Because AI does not bear responsibility, it can analyze purely. It can exhaustively explore options, calculate optima, and examine extreme situations without inhibition, unburdened by political consequences, historical judgment, or moral weight. This is why its exploratory space can be unlimited, and the real value it offers as an analytical tool. Yet for the same reason, AI cannot make the final choice. That choice is not a purely logical problem. It requires bearing the political, historical, and civilizational consequences of choosing among options that have no perfect solution; taking responsibility for possible failure; and staking the futures of countless people amid uncertainty. AI cannot replace this burden, because it does not occupy the position of bearing it.

The division of labor between humans and AI thus becomes clear: AI can conduct analysis without responsibility, opening the widest possible space of possibilities; only humans can make choices with responsibility and commit to a move within that space. A decision-maker is irreplaceable not because they calculate more accurately than AI; in pure analysis, they may not outperform it. They are irreplaceable because they occupy a position AI cannot: responsibility for the choice. Outsourcing the highest judgment to AI means handing “responsibility for the consequences” to an entity incapable of bearing it. This is logically untenable. The stronger AI becomes, the more important this division of labor becomes. Precisely because AI can take analysis to its limits, humans must preserve the one step that analysis cannot replace: bearing responsibility and making a choice.

This paper therefore seeks not to bypass specialist systems, but to prevent the highest judgment from being swallowed by specialist procedures. Whether to recognize that these risks have become an issue of common survival, whether to authorize negotiations on a joint response mechanism, and whether to bring fragmented assessments back together into decisions at the national and civilizational levels can be determined only by those who genuinely bear the highest responsibility.

Specialist teams judge whether details are ready; decision-makers judge whether to begin. Specialist teams design rules; decision-makers bear coordination costs. Specialist teams assess individual risks; decision-makers confront the overall closed loop. This is precisely the irreplaceable position of decision-makers.

Final Conclusion

If the entire paper were compressed into one inescapable chain of reasoning, it would be this:

Humanity has entered an era of foundational risks. AI militarization, AI-enabled biological risks, critical resources and computing power, vulnerable critical infrastructure, and conflicts over civilizational narratives are no longer separate domain-specific issues. They constitute crisis emergence within a single network, amplifying one another and compressing one another’s room for maneuver. Within this network, the old logic of deterrence is failing, because the symmetry, visibility, and attributability on which it depended have all been hollowed out by asymmetric realities. As long as interstate competition continues in its old form, the risks within this network will be continually amplified rather than alleviated.

For precisely this reason, none of the five apparently viable alternatives, technological controls, a balance of terror, joint management by major powers, global authority, or limited cooperation plus long-term competition, can pass all four tests together. Each is ruled out through analysis. What remains is not a better choice, but the only direction not overturned by logic: constrain competition within limits that do not destroy the common foundations. That is collaborative coexistence.

But collaborative coexistence is only the institutional foundation, not the destination. Institutions can constrain behavior but cannot, alone, determine where people find security, worth, and a sense of existence. As long as worth continues to come chiefly from “being stronger, having more, and being more right than others,” conflicts over civilizational narratives will continually erode institutional constraints. Guardrails can stop one crisis, but cannot stop the impulse to recast every crisis in a new narrative.

Civilization Leap is therefore not a decorative ideal added after collaborative coexistence. It is the fundamental path to addressing the sources of crisis and eliminating the conditions under which its embers are continually generated and amplified. Common survival is the minimum boundary against collapse; collaborative coexistence places shared guardrails around competition; Civilization Leap is the direction through which states, groups, and individuals can once again possess a future they can look forward to.

And the starting point is not distant. Civilization Leap need not begin with complete consensus. It can begin with a negotiating table, with shared boundaries around irreversible risks that, once out of control, leave no post-disaster world. Preventing loss of control is a beginning, not an endpoint. Only if that beginning is preserved can there be common survival, common security, shared expectations, and shared meaning.

Every link in this chain depends on reasoning, not goodwill. That is precisely its strength.

Closing Words

The purpose of negotiation is to gain time.

The purpose of time is not merely to prevent one instance of loss of control, but to preserve the practical conditions for a deeper transformation.

Common survival means first keeping the world from falling; collaborative coexistence means placing shared guardrails around competition; Civilization Leap addresses, at their source, the reasons crises keep recurring. Drawing on AI/AGI capabilities, and through institutional design, the sustaining of meaning, and the reconstruction of expectations for the future, it enables people’s sense of worth to come less from possession and overpowering others, and gradually more from being needed, creating, protecting, and contributing.

Across history and cultures, humanity has never stopped longing for an ideal world. Such a world is not difficult to understand: freedom from survival insecurity, happiness in life, and greater meaning in life. The difficulty has never been that people cannot understand it, but that they have long struggled to believe it could actually be achieved. The AI/AGI era changes some of the conditions for the first time, but will not automatically bring about an ideal world. It offers only a possibility: giving resource coordination, recognition of contributions, structures of being needed, and the sustaining of meaning a foundation closer to practical operation.

Giving people hope and a future they can look forward to is the driving source of human goodwill, and the deepest, most powerful way to prevent crisis emergence. A person who still has a future is less readily driven by despair; a group that still has a future is less readily consumed by hatred; a state that still has a future can more readily accept common guardrails within the rationality of security.

This paper therefore ends here not because Civilization Leap is unimportant, but because it must be developed further in more specialized texts. This paper’s task is to preserve the starting point: understand crisis emergence, rule out the old paths, initiate minimum common boundaries, and make collaborative coexistence an operational direction. The road beyond that starting point is precisely what Civilization Leap must continue to develop.

Earth, civilization’s cradle, will not wait for us forever. And we still have time—if we begin now.

— For Assessment —

Author: Zijun Fu | Civilization Leap Research Group