

Protocols Constrain Games; Self-Interest First Continually Generates New Ones

Complete Web Text PDF Edition

Exported from the public website edition on 2026-10-17. Includes the complete article, edition and evidence boundaries, references, and further reading. This is a separate downloadable web-text edition, not another DOCX master or a verified original publication. Statements below about unattached or unproduced PDFs reproduce the website at export; they do not refer to this new web-text export.

<https://www.civitas.top/en/documents/p03-surv-002/>

Translation scope

This page translates the complete publicly available Chinese HTML text, including its source version and evidence boundaries. It does not claim that a separate English PDF has been produced. Source transcription and original-document formatting limitations remain as stated below.

The problem this work addresses

Protocols still matter. What must endure through time, however, is not one eternal protocol but the capacity to regenerate correction and cooperation after old actors, equilibria and rules fail.

Version and evidence boundaries

Current version: V1.1. This English page translates the public online transcription of the uploaded Chinese source document. The source PDF has not yet been attached through this site's binary publication process, so no invented download link is provided. Refer to the source document for tables and complex formatting.

Continue reading

- Examine a real window: a China-US channel for communicating about AI incidents

<https://www.civitas.top/en/documents/p03-surv-004/>

- Continue to the overarching strategic assessment: The Only Path

<https://www.civitas.top/en/documents/p03-only-002/>

- How shared survival enters national and international structures

<https://www.civitas.top/en/documents/p03-strat-001/>

Complete online text

Civilization Leap Research Group · Short Essay

Agreements Constrain Strategic Games; Self-Interest First Keeps Generating New Ones

—Why what civilization must carry through time is not an eternal agreement, but an internal logic capable of regenerating peace and correction

Short Essay V1.1 | Edition calibrated for the foundational judgment and further reading

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

Agreements can constrain strategic games already under way, but cannot prevent a self-interest-first logic from continually generating new ones.

Human history has no shortage of agreements. Ceasefires, alliances, arms control, trade arrangements, crisis communication, international law, and domestic institutions have all served real purposes at different times. Nor can the problem simply be described as “all agreements eventually fail.” Empirical research has even found that many alliance commitments were honored within particular historical samples.[1]

The question is different: why can an agreement that was reasonable, effective, and even conscientiously observed when signed not guarantee its passage through decades, generations, or longer merely because it was signed?

I. An agreement is an equilibrium in a particular time and place, not a natural constant across time

When an agreement takes shape, actors, interests, power, technology, risks, and expectations of the future stand in a particular configuration. Both sides' willingness to comply usually means that, under those conditions, compliance better fits their respective judgments than violation.

But time does not stop at signing. Parties alternate in office, leadership changes, societies may undergo major transformations, national power rises and falls, new technologies alter existing security boundaries, and new actors and interests emerge.

The future actors who must actually honor an agreement may therefore no longer be the same actors, in the relevant sense, as those who formed it. Leeds's study of alliance commitments from 1816 to 1944 found changes in power and in policymaking processes to be among important predictors of violation.[1] Research by Fearon and Powell on commitment problems further explains that, even when an arrangement exists today that both parties prefer to conflict, peace can become unstable if one side cannot credibly commit to future compliance, especially when the future distribution of power is changing rapidly.[2][3]

An agreement is not eroded by its paper. It is repriced by the world that changes after it is signed.

II. The real test is not when compliance is most advantageous

Compliance when both sides clearly benefit does not answer an agreement's hardest question. The real test comes when one of the following first occurs:

- After a shift in power, breaking the rules may offer a greater advantage.
- When a crisis arrives, concealment, delay, or preemptive action may become advantageous.
- New leaders and social forces no longer understand old commitments as their own.
- After a technological leap, the objects governed by old rules have changed.
- Once a shared external opponent disappears, the former “we” separates again into distinct “I”s.

If an agreement's stability depends entirely on all participants continuing to find compliance more advantageous under every future condition, its long-term stability rests on a demanding premise: future actors, interests, and power structures must not change enough to alter that calculation.

The real test of an agreement is never when both sides have reasons to honor it. It is the first time one side believes breaking it would be more advantageous.

III. Agreements constrain existing strategic games; self-interest first keeps generating new ones

This is the distinction between external rules and underlying logic.

External constraints

Actors first form purposes: I want to win, be safe, lead, or control more resources. Agreements, laws, and institutions then specify which means are prohibited and which boundaries cannot be crossed.

This structure can be highly effective. The problem is that, as long as the primary logic of purpose generation remains unchanged, rules face an internal pressure: if circumventing them better achieves the original purpose, why not circumvent them?

Internal logic

The deeper change is not “I have already decided what to do, but am prohibited from doing it.” It is that, when a purpose would destroy the conditions for oneself and others to continue existing, correct errors, exit, and choose again, those consequences already have standing to change the purpose before it is locked in.

Cooperation, correction, and shared survival then cease to be merely external costs considered after a purpose forms. They begin to enter purpose generation itself.

Agreements can constrain strategic games already under way, but cannot prevent a self-interest-first logic from continually generating new ones. To change the long-term trajectory of these games, we must ultimately reach the first-order logic that generates purposes.

IV. What must endure through time is not the same agreement

There is no reason to demand that an agreement made today remain word for word unchanged a century later. Actors, technologies, and risks change. Replacing old rules with new ones does not itself mean civilizational regression.

What matters is whether new agreements can form before old ones fail; whether new actors can rebuild shared security after old actors change; whether a new equilibrium can emerge after a reversal of power without passing through irreversible conflict; and whether channels of correction remain alive after a severe error of judgment.

What civilization must preserve is therefore not Agreement Persistence—the indefinite survival of an agreement's text—but Correction Regeneration: the capacity to regenerate structures of correction.

Agreements may fail. Correction must not die.

This also means that long-term peace cannot be understood merely as containing today's conflict. The harder question is whether peace and correction can regenerate again and again as new actors, interests, and eras emerge.

V. “Internal” is not a moral demand, but a change in where purpose generation begins

Self-interest first does not exclude cooperation, but generates strategic games from within.

Self-interest with others first does not exclude competition, but generates cooperation from within.

“From within” describes the first-order relational structure of purpose generation, not an absolute prediction of final behavior. Self-interested actors can cooperate; actors who bring others in first can compete. The real difference is where cooperation and competition enter the formation of purposes.

This does not require actors to abandon their own interests, promise permanent peace, yield unilaterally, or abolish defense.

It concerns the order of purpose generation: before one's own purpose is locked in, can affected others, shared consequences, and conditions for future correction already enter and genuinely change it?

In a self-interest-first structure, cooperation can emerge, but often must first be shown to benefit me. Once returns, deterrence, or countermeasures disappear, the reason for cooperation may disappear with them.

Under self-interest with others first, others need not first become “me,” prove they can repay me, or prove they can harm me before gaining causal standing to alter a purpose. Self-interest and defense remain, but one's own interests no longer presume first-mover priority in purpose generation.

This is therefore not a new agreement placed on top of an old logic. It attempts to change why each future generation of actors would regenerate cooperation, correction, and conditions for shared survival as it forms its purposes.

VI. AI and AGI make this question newly urgent

When human capabilities were limited, even problematic purposes encountered distance, organizational costs, insufficient information, execution speeds, and resource constraints that left rules time to catch up. AI, AGI, and stronger intelligence may weaken these natural impediments together.

If future intelligence inherits only a structure of completing goals and allowing rules to correct afterward, greater capability will require downstream rules to keep chasing faster, more complex execution of purposes.

The deeper choice is to give future intelligence not only more rules, but also questions that enter purpose generation itself: should new consequences reopen the original purpose? Who should be seen before a purpose is locked in? Which actions would eliminate conditions for future correction?

What civilization needs to pass to future intelligence may not be an eternally fixed set of answers, but the underlying way answers are generated.

VII. An open boundary: our task is not to find every possible path for the world

This essay does not undertake to prove that no second path could possibly exist in the universe, still less to prove that a second path does exist.

Anyone who believes external agreements, mechanism design, deterrence, distributed governance, or an entirely different approach can endure changes of actors, shifts in power, technological leaps, and crises while preserving correction and shared survival over the long term, without touching the underlying purpose-generation logic discussed here, may set out that approach in full.

We permit and invite such exploration. A genuinely distinct path can enter public discussion and examination. Where necessary, we can also help clarify it and unpack its causal relationships and failure conditions.

But substantiating those alternatives and exhausting every possible path on the world's behalf is not this essay's task. Its continuing task is to clarify the problems and causal relationships already identified and advance the direction already developed into reality.

We do not find every path on the world's behalf, but welcome every path that emerges.

VIII. What matters is not more people accepting the same answer

Open discussion is not intended to turn “the only path” into a count of affirmative votes. With more participants, what should increase is independent reasoning by different actors: different experiences, disciplines, positions of interest, cultures, and counterexamples entering the same question.

Some may retrace the causal chain and reach the same result, deepening understanding. Some may identify a break in the logic, prompting revision. Some may propose a better path, resolving the problem. Some may turn an unresolved question into real action, bringing ideas into the world.

What matters is not more people believing the same answer, but more people genuinely thinking through the same question.

IX. Conclusion: civilization needs a logic that can regenerate peace

Human history repeatedly establishes agreements and repeatedly confronts new actors, interests, and distributions of power. Agreements therefore remain necessary, but an agreement itself is not the root of stability.

If civilization adds rules only after conflicts arise, downstream institutions must forever chase upstream purposes. If purpose generation itself begins to bring others, shared consequences, and future correction into consideration earlier, rules and agreements may increasingly become practical expressions of that internal logic, rather than external brakes forever opposing it.

Truly stable peace is not a strategic game held down by rules. It is the capacity, even after new actors emerge, to regenerate reasons not to push one another into irreversibility.

This is not a promise of perpetual peace. It moves the question further upstream: perhaps what we actually need to change has never been a particular agreement, but why future actors will generate the same kind of strategic game again.

References and adjacent research

[1] Leeds, B. A. Alliance Reliability in Times of War: Explaining State Decisions to Violate Treaties. *International Organization*, 57(4), 801–827, 2003. DOI: 10.1017/S0020818303574057.

[2] Fearon, J. D. Rationalist Explanations for War. *International Organization*, 49(3), 379–414, 1995.

[3] Powell, R. War as a Commitment Problem. *International Organization*, 60(1), 169–203, 2006. DOI: 10.1017/S0020818306060061.

Further reading

- Upstream real-world structures: Crisis Emergence: New Structural Forms of Interstate Strategic Interaction in the AGI Era, v4.0.
- Overall strategic assessment: The Only Path: States' Long-Term Interests and Civilization Leap in the AI Era, v6.21.
- Real-world governance structures: Bloc Formation in AI Governance and Humanity's Overall Security, v2.0.
- Institutional development and the survival framework: Realizing, Evolving, and Transforming a Community with a Shared Future for Humanity, V3.9.3-P1; How Can Civilization Continue in the Age of AI, AGI, and ASI?, V1.1.1; REAL-ACT 01, A China–US AI Incident Communication Channel, V0.2.1.

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