

AGI-COS 2.0

AGI Crisis Operating System → Civilizational Immune System

AGI-COS 2.0

Civilizational Immune System Protocol

Version: 2.0

Status: Open Global Protocol

License: CC BY 4.0

Year: 2026

Author: Zijun Fu

Series: Civilization Leap Protocol Framework

Canonical Statement

AGI-COS 2.0 defines a civilizational immune system in which Artificial General Intelligence functions as a continuous cognitive sensing infrastructure assisting humanity in maintaining systemic stability while preserving human sovereignty, reversibility, and future openness.

AGI analyzes.

Humanity decides.

Civilization adapts.

0. Purpose

Human civilization has entered an era characterized by:

- accelerating technological capability,
- tightly coupled global systems,
- increasing irreversible risks.

Traditional crisis governance reacts after failure emerges.

AGI-COS 2.0 shifts governance from crisis response to continuous civilizational resilience.

The objective is not to eliminate crises but to prevent systemic collapse conditions from forming.

1. Core Principles (Unchanged Constitutional Layer)

AGI-COS 2.0 preserves the three non-negotiable principles established in v1.x.

1.1 Human Sovereignty Principle

AGI shall never become a governing authority.

All strategic decisions remain human responsibilities.

AGI possesses:

- no executive power,
- no goal-setting authority,
- no accountability transfer.

1.2 Reversibility Principle

When irreversible risk exists, reversible paths must be prioritized over optimal efficiency.

Every intervention must remain:

- stoppable,
- auditable,
- recoverable,
- compensable.

1.3 Anti-Closure Principle

No system may permanently restrict humanity's future choice space.

Civilization must remain adaptive and plural.

2. Paradigm Transition

AGI-COS v1.x operated as a Crisis Operating System.

AGI-COS 2.0 establishes a:

Civilizational Immune System (CIS)

Transition:

Phase	Function
v1.x	Respond to crises
v2.0	Prevent systemic failure formation

Governance evolves from episodic intervention to continuous stabilization.

3. System Architecture

AGI-COS 2.0 introduces an immune-system-inspired architecture.

Governance Kernel



Continuous Sensing Layer



Perception Audit Layer



Human Decision Layer



Micro-Intervention Layer



Recovery & Learning Layer



Adaptive Evolution Layer

4. Continuous Sensing Layer (NEW)

Purpose

Detect systemic stress before crises emerge.

AGI operates as a civilizational sensing network.

Monitoring Domains

- economic instability signals
- technological risk accumulation
- geopolitical escalation patterns
- information ecosystem degradation
- institutional trust erosion
- infrastructure fragility

New Indicator

SSI — System Stress Index

Measures latent instability.

SSI evaluates:

- coupling tension,
- acceleration of feedback loops,
- resilience depletion.

High SSI triggers preventive observation, not emergency intervention.

5. Perception Audit Layer (Inherited from v1.1)

All AGI outputs must include:

1. Confidence Index
2. Data Lineage Map
3. Unknown Zone Declaration

Additional metric:

TCS — Trust Calibration Score

Independent auditing verifies AGI self-confidence accuracy.

6. Human Decision Layer

Human governance remains the central authority.

AGI provides:

- modeling,
- simulation,
- scenario projection,
- risk amplification analysis.

AGI never selects policy.

Decision Modes

Risk Level AGI Role

Low Analytical assistant

Medium Scenario simulator

High Risk evaluator

Extreme Red-Team stress tester only

7. Micro-Intervention Layer (NEW)

Civilizations fail when only large interventions exist.

AGI-COS 2.0 introduces micro-interventions.

Definition

Small, reversible adjustments designed to reduce systemic stress without triggering political shock.

Examples:

- communication recalibration,
- localized resource adjustment,
- algorithm deployment slowdown,
- institutional coordination improvement.

Requirements

Micro-interventions must be:

- low visibility,
- reversible,
- culturally compatible,
- continuously evaluated.

8. Recovery & Learning Layer

Failures are treated as learning events.

After anomalies:

- outcomes recorded,
- models recalibrated,
- governance updated.

All failures contribute to a global learning archive.

9. Adaptive Evolution Layer (NEW)

AGI-COS evolves through distributed experimentation.

Implementation method:

MEU 2.0 — Minimal Evolution Unit

Small-scale societal environments test adaptive governance strategies safely.

Evolution occurs through iteration, not redesign.

10. Civilizational Health Monitoring

AGI-COS 2.0 introduces continuous reporting.

Civilization Health Report (CHR)

Periodic publication including:

- SSI
- CI
- WUI
- IRI
- CSI
- PCI

Purpose:

Not policy prescription.

Only measurement.

Measurement precedes coordination.

11. Decision Support Degradation Mode

When IRI becomes extreme:

AGI shifts to Red-Team Mode.

AGI outputs only:

- worst-case projections,
- cascading failure risks,
- unintended consequence simulations.

Human authority retains full decision responsibility while avoiding cognitive overload.

12. Governance Model

AGI-COS remains:

- non-national,
- non-ideological,
- open infrastructure.

Any entity may implement local variants.

Only systems preserving core principles may claim compatibility.

13. Compatibility Certification

Designation:

AGI-COS Compatible

Requires preservation of:

- Human Sovereignty
 - Reversibility
 - Anti-Closure
-

14. Deployment Philosophy

AGI-COS 2.0 does not require global adoption.

Civilizational immunity emerges through:

- partial adoption,
 - independent experimentation,
 - gradual convergence.
-

15. Long-Term Objective

AGI-COS aims to transform civilization from a reactive system into a self-stabilizing adaptive organism capable of surviving accelerating complexity.

16. Canonical Closing Statement

The purpose of AGI-COS is not to control the future.

It is to ensure humanity never loses the ability to choose one.

当前受控阅读说明

P05-AGICOS-006 | V2.0 | 文明免疫/协调

原版本正文保留；本页只说明当前效力、边界与续读，不静默重写历史文本。

这条轨道解决什么

如何从事件响应进入连续感知、红队降级、恢复学习、分布式试验和多节点文明免疫。

当前仍有效什么

当前专业基线之一。强调持续风险感知、降级、恢复学习和分布式试验。

不能自动继承什么

MEU只能用于机制原型、接口和局部因果验证，不能冒充完整文明跃迁验证；低可见性应解释为低扰动、低侵入而非隐秘治理。

继续进入

AGI-COS 1.0 | 危机相位仍需调用危机操作轨。

AGI-COS 1.1 | 所有高后果调用继续保留感知审计。

分阶段一页入口 | 进入2.1协调、2.2技术子轨说明。

总原则：AGI-COS不存在覆盖全部任务的“最高版本”；版本可以组合调用，权限不能叠加。

版本族入口：<https://www.civitas.top/families/p05-agicos/>

分阶段入口：<https://www.civitas.top/documents/p05-agicos-011/>