

AGI-COS v1.1

Stability Upgrade Edition

Zijun Fu

Produced by Civilization Leap Research Group

AGI-COS v1.1

AGI Crisis Operating System

Stability Upgrade Edition

1. Version Statement (New)

Version: v1.1

Upgrade Type: Stability & Reliability Upgrade

Compatibility: Fully compatible with v1.0

AGI-COS v1.1 introduces three structural safeguards:

1. Perception Audit Layer
2. Pre-Deployment Protocol
3. Compatibility Certification Standard

These upgrades strengthen operational reliability without changing core philosophy.

2. Core Architecture Update

The AGI-COS architecture is upgraded from five layers to six:

Governance Kernel

↓ Perception Audit Layer ← (New) ↓ AGI Perception Layer

↓ Human Decision Layer

↓ Intervention Layer

↓ Recovery & Evolution Layer

3. New Module 1

Perception Audit Layer

3.1 Purpose

AGI outputs are not treated as facts, but as:

Cognitive advice accompanied by a confidence level.

This module is intended to prevent:

- AGI hallucinations
- Data contamination
- The effects of information warfare
- Out-of-distribution misjudgments

3.2 Mandatory Outputs

All AGI analyses must include:

1. Confidence Index

Levels:

- High Confidence
- Medium Confidence
- Low Confidence

Low-confidence results must not be the sole basis for a decision.

2. Data Lineage Map

The following must be shown:

- Categories of data sources
- Data freshness
- Relative weights
- Potential sources of bias

3. Unknown Zone Declaration

AGI must explicitly identify:

Areas in which the current model cannot make reliable judgments.

3.3 Updated Rules for Human Decisions

Human governing actors must record:

- Whether low-confidence advice was adopted
- The reasons for adopting it

4. New Module 2

COS Pre-Deployment Protocol

4.1 Core Principle

AGI-COS is not an improvised crisis tool.

Preparations must be completed before a crisis occurs.

4.2 Pre-Deployment Checklist (Mandatory Preparation)

Organizations must establish the following in advance:

1. Local Governance Dataset

- Legal texts
- Policy frameworks
- Management procedures

2. Crisis Semantic Dictionary

- Risk terminology
- Local social context
- Administrative keywords

3. Communication Interface

- AGI outputs → Mapping into decision-making language
- Probability statements → Action levels

4.3 Drill Requirements

The following are recommended annually:

- Simulated COS activation drills
- Indicator-generation tests
- Human-AI collaboration exercises

5. New Module 3

AGI-COS Compatibility Certification

5.1 Background

An open protocol permits forks, but must prevent:

- Misrepresented versions that remove core principles
- Authoritarian misuse
- Conceptual confusion

5.2 Compatibility Standard

Only systems that meet all the following conditions may use the designation:

AGI-COS Compatible

The following must be retained:

1. Human Sovereignty Principle
2. Reversibility Principle
3. Anti-Closure Principle

5.3 What Is Permitted

- Localized modifications
- Language adaptations
- Differences in technical implementation

5.4 What Is Not Permitted

Systems must not:

- Give AGI decision authority
- Remove rollback mechanisms
- Close off auditing and alternative paths

6. Updated Constraints on Path Design (New)

New requirement:

Socio-Cultural Compatibility Requirement

Intervention paths proposed by AGI must include:

1. Technical Optimal Path
2. Social Acceptance Estimate
3. Cultural Conflict Warning

The decision-making layer must examine all three results together.

7. Indicator System Updates

New indicator:

PCI — Perception Confidence Index

Measures:

- Data trustworthiness
- Model stability
- Information consistency

When PCI is too low:

Automatically reduce the weight given to AGI advice.

8. Updated Rules for Human-AI Collaboration

Logic for adjusting decision weights:

IRI Level	AGI Weight	Human Weight
Low	50%	50%
Medium	35%	65%
High	20%	80%
Extreme	≤10%	100%

9. Backward Compatibility Statement

AGI-COS v1.1 :

- Does not change the principles of v1.0
- Does not change the governance philosophy
- Only adds safeguards for stability

Implementers of v1.0 can upgrade without rebuilding their systems.

10. Core Proposition (Updated)

The goal of AGI-COS is not to rely on AGI to solve crises,
but, in an era when complexity exceeds human intuition,
to use auditable, reversible, and verifiable human-AI collaboration
to ensure that civilization always retains the ability to keep choosing its future.

Current Controlled Reading Note

P05-AGICOS-005 | V1.1 | Perception and Cognitive Safety

The original version text is retained. This page explains only its current applicability, boundaries, and further reading; it does not silently rewrite the historical text.

What this track addresses

In high-consequence judgments, how can model outputs carry sources, confidence levels, data lineage, and declarations of unknowns, rather than being treated as facts?

What remains applicable

This track remains active across stages. Perception audits, data provenance, the Unknown Zone, reasons for adopting low-confidence advice, and pre-deployment preparation should be included in high-consequence uses.

What is not automatically inherited

Fixed human-AI decision weights no longer apply. Risk indicators cannot automatically allocate governance authority. The cognitive-load and adversarial-resilience provisions in 1.2 may be combined as a controlled sub-track.

Continue reading

[AGI-COS 2.0 | Move from individual judgments to continuous civilizational immunity.](#)

[AGI-COS 1.0 | Use the crisis lifecycle when entering the crisis phase.](#)

[One-Page Guide by Stage | See when other tracks should be combined.](#)

General principle: AGI-COS has no highest version that covers every task. Versions may be used together; their authorities must not accumulate.

[Version family](#)

[One-Page Guide by Stage](#)