

AGI-COS v1.2

AGI Crisis Operating System

Ecological Resilience Edition (Ecological Resilience Edition)

1. Version Statement

Version: v1.2

Upgrade Type: Ecological Adaptability & Cognitive Resilience Upgrade

Compatibility: Fully compatible with v1.1 and v1.0

AGI-COS v1.2 introduces three evolutionary mechanisms to address human cognitive limits, adversarial environments, and social feedback loops:

Cognitive Load Optimizer (Cognitive Load Optimizer)

Adversarial Defense Module (Adversarial Defense Module)

Societal Resonance Loop (Societal Resonance Loop)

These upgrades transition the system from "stable operation" to "ecological co-evolution".

2. System Architecture Update

Building on the six-layer structure of v1.1, the AGI-COS architecture strengthens the logic of interaction between layers:

Governance Kernel

↓

Perception Audit Layer

↓ (New interaction interface)

Cognitive Interface Optimizer (Cognitive Interface Optimizer)

↓

AGI Perception Layer

↓

Human Decision Layer

↓

Intervention Layer

↓

Recovery & Evolution Layer

Note: the Cognitive Interface Optimizer is not an additional physical layer, but a "protocol converter" embedded between the Perception Audit Layer and the decision layer to ensure that information density matches human cognitive bandwidth.

3. New Module 1

Cognitive Load Optimizer (Cognitive Load Optimizer)

3.1 Problem Background

During high-pressure crises, the detailed audit information required by v1.1 (provenance maps and confidence details) may cause "cognitive overload" and decision paralysis among human decision-makers.

3.2 Adaptive Display Protocol

The system automatically adjusts output information density based on WUI (Window Urgency Index):

Green Mode (Normal WUI): Display full technical details and data provenance.

Yellow Mode (Elevated WUI): Collapse secondary data sources, displaying only key deviations and core recommendations.

Red Mode (Critical WUI): Activate "decision dashboard mode", displaying only:

Core risk alerts

Reversibility status

Recommended action options

3.3 Action Rationale

Every AGI recommendation must generate an Action Rationale of no more than 50 characters, ensuring that decision-makers understand the core logic within 10 seconds.

4. New Module 2

Adversarial Defense Module (Adversarial Defense Module)

4.1 Defense Targets

The defense target expands from "random errors/hallucinations" in v1.1 to "malicious manipulation":

- Data poisoning
- Adversarial example attacks
- Manipulation through information warfare

4.2 Red-Teaming Protocol

At critical decision points, the system must trigger Shadow Red-Teaming:

- Simulate an attacker's perspective to find logical flaws in the current AGI recommendation.
- If clear signs of adversarial manipulation are found in the decision path, automatically trigger a Security Alert and freeze that recommendation.

4.3 Data Immunity Mechanism

Conduct a "motivation audit" on input data:

- Identify anomalous data patterns (such as unnatural abrupt changes in time series).
- Automatically down-weight or quarantine data from a single source or an untrustworthy source.

5. New Module 3

Societal Resonance Loop (Societal Resonance Loop)

5.1 Core Concept

The effectiveness of governance depends on social cooperation. A technically optimal solution without societal resonance will lead to implementation failure or a secondary public-opinion crisis.

5.2 New Indicator: SRI — Societal Resonance Index (Societal Resonance Index)

Measures:

- Public understanding of intervention measures
- Stability of social sentiment
- Trust in official information vs the rate of rumor propagation

5.3 Dynamic Feedback Mechanism

When SRI falls below the set threshold:

- Execution-layer adjustment: suspend highly coercive measures and shift toward communication and reassurance strategies.
- Output-layer adjustment: require AGI to convert "administrative terminology" into "plain language for society" for release.

6. Path Design Constraint Upgrade

Building on "sociocultural compatibility" in v1.1, add:

Implementation Resistance Assessment

Intervention paths proposed by AGI must include:

Technical Optimal Path (Technical Optimal Path)

Social Acceptance Estimate (Social Acceptance Estimate)

Implementation Resistance Factor (Implementation Resistance Factor) ←(New)

The decision layer must confirm: if the implementation resistance factor is too high, priority must be given to adjusting the plan or increasing communication resources rather than forcing it forward.

7. Indicator System Update

In addition to the existing indicators (CI/WUI/IRI/CSI/PCI), add:

CLI — Cognitive Load Index (Cognitive Load Index)

- Measure whether the current volume of decision information exceeds human processing limits.
- When CLI is too high, the system automatically enforces simplification of the output interface.

SRI — Societal Resonance Index (Societal Resonance Index)

- Quantify societal acceptance and cooperation in real time.

8. Human-Machine Collaboration Rule Update

Introduce "Dual-Factor Veto" logic:

Rule A: Confidence Circuit Breaker

If $PCI < 20\%$ (extremely low perception confidence):

- The system prohibits recommending "irreversible actions".
- Only "observation, defensive isolation, and tentative action" are permitted.

Rule B: Societal Resonance Down-Weighting

If $SRI < 20\%$ (extremely low societal resonance):

- Even if confidence in an AGI recommendation is high, the system automatically lowers the AGI weight by one level.
- Mandatorily introduce a "human consensus meeting" procedure to prevent a disconnect between technology and public opinion.

9. Backward Compatibility Statement

AGI-COS v1.2 :

- Fully inherits the principles of v1.0 and the audit mechanisms of v1.1.
- v1.1 implementers need only update the "interface-layer protocols" and "indicator monitoring set" to upgrade.
- No restructuring of the core governance architecture is required.

10. Core Proposition (Further Evolution Edition)

The goal of AGI-COS is not only to preserve the ability to choose the future, but also to build a mechanism

through which humans, assisted by machine intelligence, can still exercise civilizational sovereignty calmly, lucidly, and empathetically.

End of Release