

AGI-COS v1.1

AGI Crisis Operating System

Stability Upgrade Edition

一、版本声明（新增）

Version: v1.1
Upgrade Type: Stability & Reliability Upgrade
Compatibility: Fully compatible with v1.0

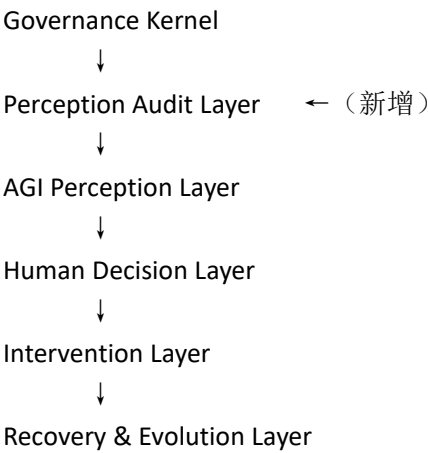
AGI-COS v1.1 introduces three structural safeguards:

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

These upgrades strengthen operational reliability without changing core philosophy.

二、系统架构更新（Core Architecture Update）

AGI-COS 架构由五层升级为 六层结构：



三、新增模块 ①

Perception Audit Layer（感知审计层）

3.1 目的

AGI 输出不被视为事实，而被视为：

带置信度的认知建议。

本模块用于防止：

- AGI 幻觉
- 数据污染
- 信息战影响
- 分布外误判

3.2 强制输出要求（Mandatory Outputs）

所有 AGI 分析必须附带：

① Confidence Index（置信度指数）

等级：

- High Confidence
- Medium Confidence
- Low Confidence

低置信度结果不得作为单一决策依据。

② Data Lineage Map（数据溯源图谱）

必须显示：

- 数据来源类别
- 时间新鲜度
- 权重占比
- 潜在偏差来源

③ Unknown Zone Declaration（未知区域声明）

AGI 必须明确标注：

当前模型无法可靠判断的领域。

——

3.3 人类决策规则更新

人类治理主体必须记录：

- 是否采纳低置信度建议
- 采纳理由

——

——

四、新增模块 ②

COS Pre-Deployment Protocol（预部署协议）

——

4.1 核心原则

AGI-COS 不是危机临时工具。

必须在危机发生前完成准备。

——

4.2 预部署清单（Mandatory Preparation）

组织需提前建立：

1. Local Governance Dataset

- 法律文本
- 政策体系
- 管理流程

2. Crisis Semantic Dictionary

- 风险术语
- 本地社会语境
- 行政关键词

3. Communication Interface

- AGI 输出 → 决策语言映射
 - 概率表达 → 行动等级
-

4.3 演练要求

建议每年进行：

- COS 模拟启动演练
 - 指标生成测试
 - 人机协同演习
-
-

五、新增模块 ③

AGI-COS Compatibility Certification

5.1 问题背景

开放协议允许分叉，但必须防止：

- 删除核心原则的伪版本
 - 威权化滥用
 - 概念混淆
-

5.2 Compatibility 标准

只有同时满足以下条件，才可标注：

AGI-COS Compatible

必须保留：

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

5.3 允许内容

- 本地化修改
- 语言适配
- 技术实现差异

5.4 不允许内容

不得：

- 赋予 AGI 决策权
- 删除可回滚机制
- 关闭审计与替代路径

六、路径设计约束升级（新增）

新增：

Socio-Cultural Compatibility Requirement

AGI 提出的干预路径必须包含：

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

决策层必须同时审视三项结果。

七、指标体系更新

新增指标：

PCI — Perception Confidence Index

衡量：

- 数据可信度
- 模型稳定性
- 信息一致性

PCI 过低时：

自动降低 AGI 建议权重。

——

——

八、人机协同规则更新

决策权重调整逻辑：

IRI 等级	AGI 权重	人类权重
低	50%	50%
中	35%	65%
高	20%	80%
极高	≤10%	100%

——

——

九、向后兼容声明

AGI-COS v1.1:

- 不改变 v1.0 原则
- 不改变治理哲学
- 仅增加稳定性防护

v1.0 实施者无需重构系统即可升级。

——

——

十、核心命题（更新版）

AGI-COS 的目标不是依赖 AGI 解决危机，

而是在复杂性超过人类直觉的时代，
通过可审计、可回滚、可验证的人机协同，
确保文明始终保留继续选择未来的能力。

当前受控阅读说明

P05-AGICOS-005 | V1.1 | 感知/认知安全

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

高后果判断中，怎样让模型输出带来源、置信度、数据血缘和未知声明，而不是把模型答案当成事实。

当前仍有效什么

跨阶段当前有效。感知审计、数据溯源、Unknown Zone、低置信度采纳理由与预部署准备应进入高后果调用。

不能自动继承什么

固定人机决策权重退出；风险指标不能自动分配治理权。1.2的认知负荷/对抗韧性可作为受控子轨组合。

继续进入

AGI-COS 2.0 | 从单次判断进入持续文明免疫。

AGI-COS 1.0 | 进入危机相位时调用危机生命周期。

分阶段一页入口 | 查看何时组合其它轨道。

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/>