

AGI-COS 版本地图

文明级原则、治理框架与软件实现的分层说明

子君赋 | 文明跃迁研究组

v2.0 正式提交版 | 2026 年 7 月

一句话定位

AGI-COS 是一套文明级安全运行协议栈：底层规定不可突破的文明原则，中层规定治理、审计和危机接口，上层可以由不同软件与机构实现。它不是单一软件，也不是由 AI 替代人类治理的操作系统。

一、三层结构

层级	回答的问题	典型内容	不可混同
原则层	后 AGI 时代什么不能被突破	人类最终决定权、可逆性、反封闭、未知声明、三元目标门	不是产品功能清单
治理层	多主体怎样共同管理风险与责任	权限、审计、申诉、熔断、回滚、责任锚点	不是单一机构统治
实现层	具体系统怎样落地	感知、日志、模型接口、策略模拟、软件与硬件控制	任何实现都不能反向修改原则

二、版本演进逻辑

阶段	核心任务	状态说明
L0 概念层	定义文明级操作系统与 AGI 放大器位置	概念与原则
L1 宪法层	固定人类最终判断、可逆性与反封闭底线	上位约束
L2 感知审计层	置信度、溯源、未知声明、日志与责任锚	治理/技术接口
L3 文明免疫层	识别结构性失真、级联风险与意义压平	跨系统治理
L4 共同窗口层	不同制度和主体之间的最低危机沟通接口	兼容与降误判

阶段	核心任务	状态说明
L5 实验实现层	在限定场景测试软件、硬件和组织实现	必须受相位纪律约束
L6 长期展望层	跨区域、跨文明、跨形态安全运行	开放研究，不作为启动期义务

三、相位边界

- 启动期只建立人类终审、日志、未知声明、暂停和回滚等最低接口。
- 初期可以在有限场景运行感知与审计模块，不赋予系统文明目标决定权。
- 过渡期才讨论跨机构互操作、责任分担和更复杂的共同窗口。
- 稳定期的全域文明操作框架属于长期方向，不能被包装成当前即可部署的统一系统。

四、评议问题

1. 哪些内容属于文明原则，哪些属于治理设计，哪些属于技术实现？
2. 原则如何被不同制度环境共同接受，而不要求统一价值观？
3. 如何防止软件实现者通过技术细节修改上位规则？
4. 如何设计跨系统熔断而不造成新的级联灾难？
5. 何种证据足以使某一实现从实验版进入更高相位？

从本文继续前行

以下入口按本文的自然问题链承接，不重复正文已有的体系说明；均指向 www.civitas.top 当前稳定文稿页。

本文网站文稿页

<https://www.civitas.top/documents/p05-agicos-003/>

<https://www.civitas.top/go/p05-agicos-003/>

01 AGI-COS全系列总览

按阶段与相位管理危机操作、感知审计、文明免疫、稳定文明、星际运行和长期演化。

<https://www.civitas.top/documents/p05-agicos-001/>

02 AGI-COS技术与治理蓝皮书

按阶段与相位管理危机操作、感知审计、文明免疫、稳定文明、星际运行和长期演化。

<https://www.civitas.top/documents/p05-agicos-004/>

03 AI开始杀敌人，人类凭什么就不是敌人

把自动化决策、AI军事化和系统失控问题转化为可观察、可处置的风险结构。

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

完整公开文库

<https://www.civitas.top/library/>

公众阅读总入口

<https://www.civitas.top/documents/p01-entry-006/>