AGI-COS 核心典藏版

文明免疫系统 · 面向 AGI 时代的系统稳定、风险协调与开放未来框架

项目	信息
文稿编号	CL-AGICOS-003-ZH-v2.0-CE-PUB
版本	v2.0-CE｜网站公开结构补全版
状态	PUB｜网站公开版
作者	子君赋
出品	文明跃迁研究组
官网	civilleap.com / www.civitas.top
联系邮箱	zijunfu@civitas.top
日期	2026 年 5 月

AGI 分析。人类决定。文明适应。

本文件为 AGI-COS 全系列网站公开资料的一部分。它用于公开说明文明免疫系统的定位、边界、版本关系与续读路径，不构成治理授权、工程部署说明、军事接口、执法接口或资源分配方案。

文档信息 / 版权页

项目	说明
文稿名称	AGI-COS 核心典藏版
所属体系	文明跃迁理论体系｜CL-AGICOS｜AGI-COS 治理协议层
文稿定位	AGI-COS 全系列的核心公开入口；用于说明文明免疫系统的理论地基、架构原则、限制宪章与续读路径。
公开边界	公开原则、体系定位、边界、续读路径；不公开高风险自动化治理、军事、执法、个人监控、资源调度接口。
版权与授权	© 子君赋 / 文明跃迁研究组。公开引用请注明作者、文稿名称、版本与官网。
建议引用	子君赋：《AGI-COS 核心典藏版》，文明跃迁研究组，2026。
版本说明	前台采用“核心典藏版 2.0-CE”作为网站主入口名称；文明永续典藏版 3.0-CE 作为延展整合稿归入版本链说明。

重要边界声明

- AGI-COS 不是政府，不拥有立法权、行政权、执法权或强制权。
- AGI-COS 不做决策，不设定社会目标，不决定资源分配。
- AGI-COS 不得成为个人监控系统，不追踪个人，不建立个人画像。
- AGI-COS 只作为认知辅助、风险测量、边界提醒与开放未来保护框架。

阅读前导

本文回答的问题

- AGI 时代，文明为何需要从危机响应转向文明免疫？
- AGI-COS 在文明跃迁体系中处于什么位置？
- 如何防止文明免疫系统滑向文明控制系统？
- 全系列版本如何从 1.0 演进到 Ω 与 Ω+？

适合读者

- AI 治理、安全与政策研究者；
- 关心 AGI 对人类长期存在影响的技术团队、机构与公共决策者；
- 已经读过《文明跃迁五卷主链》《文明永续》或《AI 时代文明跃迁白皮书》的读者；
- 希望理解“AGI 分析，人类决定，文明适应”这一底线的读者。

一、AGI-COS 的核心定位

AGI-COS（Artificial General Intelligence - Civilization Operating System）不是政府，不是国际组织，也不是自动决策系统。它是一套原则、工具、流程、边界与审查协议的组合，用于帮助人类文明在 AGI 时代获得持续感知、风险分析、边界提醒、微干预评估、恢复学习与开放未来保护能力。

它的公开底线可以概括为：AGI 分析，人类决定，文明适应。

二、文明阶段理论：为什么自觉文明需要免疫系统

AGI 的出现意味着文明进入“自觉文明”阶段：技术能力超过人类直觉，全球风险高度耦合，决策时间尺度缩短，人类第一次需要主动管理自身力量。

文明阶段	核心问题	对应结构
生存文明	如何活下去	武力、禁忌、部落协作
秩序文明	如何避免内部崩溃	法律、国家、长期规则
生产文明	如何持续增长	市场、资本、技术创新
自觉文明	如何管理自身力量	文明免疫系统
长期文明	如何长期存在	开放未来、可逆稳定、意义守恒

三、文明稳定方程

长期文明可以被理解为四个变量的动态平衡：

长期文明 = 稳定性 × 协调能力 × 学习能力 × 意义结构

- 稳定性：系统在变化中保持连续性的能力；
- 协调能力：不依赖单一强制权力而形成合作的能力；
- 学习能力：从错误中回滚、修正和迭代的能力；
- 意义结构：让未来仍值得存在、让个体仍愿意参与的结构。

四、三元目标锚定条款

AGI-COS 的任何机制、指标与干预评估，都必须锚定三大文明目标：

目标	含义	审查门
集体生存	文明不得被推向不可逆崩溃状态。	生存门：不可逆风险不增加
个体福祉	保护个体追求繁荣生活的物质与心理条件。	福祉门：生活基础条件不恶化
有意义存在	保留探索、选择、失败与未来开放的空间。	意义门：未来空间不收缩

三目标非倒退规则：若任一目标出现净倒退且无其他目标的可证明补偿，相关部署不得继续。生存安全的提升，不能证明永久丧失自由或意义是合理的。

五、七层系统架构（公开说明版）

层级	公开功能	边界
治理内核（人类主权）	确认最终责任主体始终是人类。	AGI 不获得决策主权。
连续感知层	基于匿名、聚合、统计级信号观察系统压力。	不得追踪个人。
感知审计层	验证输入、指标与风险信号是否可信。	不得成为信用评分系统。
人类决策层	把风险分析转交人类机构、社群或责任主体。	不替人类选择政策。
微干预层	评估低强度、可回滚、可退出的调节方案。	不得改变权力结构。
恢复与学习层	记录失败、回滚、复盘与可公开经验。	不建立惩罚性档案。
自适应进化层（MEU）	支持小规模、可审计、可暂停的制度实验。	禁止一刀切推广。

六、不可逾越的边界：限制宪章摘要

- 非治理：不构成政府，不拥有立法、行政、执法、强制权。
- 非决策：不选择政策，不设定社会目标，不决定资源分配。
- 非监控：不追踪个人，不建立个人画像，不存储个体级长期数据。
- 最小干预：任何微干预必须可回滚、可审计、可退出、影响最小化。
- 不可统一：不追求全球统一治理，允许多路径、多制度、地方适配。
- 不可优化：不以“最优文明状态”为目标，只保持开放与可调整状态。
- 不可永久化：任何版本都必须允许被修改、替代或废弃。
- 人类优先失败权：除不可逆风险外，人类仍保留试错权。
- 权力隔离：不得与军事指挥、强制执法、自动武器或单一国家安全结构融合。
- 自我限制优先：当系统能力扩展与自由风险冲突时，必须优先限制自身。

七、版本链中的位置

核心典藏版是 AGI-COS 全系列的主入口之一，但不是终局版本。它应与 1.0、1.1、3.0、4.0、Ω、Ω+ 及危机方法论、反封闭原则、共同窗口协议共同阅读。

版本	定位	公开说明
1.0	危机响应	人类主权、可逆性、反封闭三原则。
1.1	稳定性升级	感知审计层、预部署协议、兼容认证。
1.2	生态韧性	认知负荷优化、对抗防御、社会共鸣。
2.0 / 2.0-CE	文明免疫	连续感知、微干预、MEU、限制宪章。
3.0	后 AGI 稳定	意义层、人类代理层、文明演化层。
4.0	星际协调	分布式主权、光速治理、AI 谱系控制。
Ω / Ω+	永续文明	存在监测、意义再生、动态永恒与未知神圣性。

八、公开边界与风险防偏

本公开版只说明原则、边界、体系位置与公开阅读路径，不作为部署说明或工程操作手册。尤其不得将 AGI-COS 解释为全球治理系统、资源分配系统、个人监控系统、军事或执法接口。

续读指引

方向	推荐内容
向上续读	《文明跃迁五卷主链总览》《文明永续》《意义永生》
同层续读	《AGI-COS 全系列总览》《文明免疫系统限制宪章》《AGI 时代危机与问题处理方法论》
横向续读	《反封闭原则》《共同窗口协议》《AI 时代文明跃迁白皮书》
向下续读	未来可进入治理沙盒、公开标准草案、风险分类与审查协议；高风险执行机制暂不公开。
反向对照	资源池细节、军事化推演、执法接口、个人监控接口均不作为公开续读入口。

相关下载与网站回流入口

- 官网：civilleap.com
- 当前入口：www.civitas.top
- 文库：www.civitas.top/library.html
- 下载中心：www.civitas.top/downloads.html
- 阅读地图：www.civitas.top/reading.html
- AGI-COS 专题：www.civitas.top/agi-cos/
- 联系邮箱：zijunfu@civitas.top