

AGI-COS 2.1 完整版（Canonical Operational Edition）。

本版本定位为：

AGI-COS 从理论体系正式进入“全球可运行阶段”的标准文本。

它是 AGI-COS 1.x + 2.0 的工程化整合升级，
用于 政策递交 / 国际讨论 / MEU 试点 / 危机干预对接。

——

AGI-COS 2.1

Global Civilizational Operating System for the AGI Era

全球 AGI 时代文明操作系统（2.1 完整版）

Version: 2.1 Canonical

Status: Operational Framework

Nature: Coordination Infrastructure

Principle:

AGI analyzes • Humans decide • Civilization adapts

——

PART I — 时代背景与目标

1. AGI 时代的根本问题

AGI 的出现使人类首次进入：

- 决策复杂度超越人类直觉
- 技术加速超过制度反应速度
- 风险呈系统耦合状态

文明面临的核心威胁不再是单点灾难，而是：

- 不可逆路径形成
- 自动化锁定
- 协调失败
- 文明意义萎缩

——

2. AGI-COS 2.1 的使命

AGI-COS 2.1 的目标：

1. 提升全球协调速度
2. 延长文明可逆时间
3. 防止系统性失控
4. 建立文明级免疫结构

PART II — 核心原则（Non-Negotiable Principles）

Principle 1 — Human Sovereignty

AGI 永不拥有决策权。

AGI 负责：

- 分析
- 模拟
- 风险评估

人类负责：

- 价值判断
- 政策选择
- 责任承担

Principle 2 — Reversibility First

所有干预必须：

- 可暂停
- 可退出
- 可回滚

不可逆设计视为高风险状态。

Principle 3 — Anti-Closure

文明不得被单一最优解锁定。

必须允许：

- 多路径发展
- 制度竞争
- 社会实验

Principle 4 — Minimal Intervention

仅在必要时介入。

系统目标不是控制社会，而是维持稳定空间。

Principle 5 — Non-Governance AGI

AGI-COS 不是政府。

它是：

认知与协调基础设施

PART III — AGI-COS 2.1 核心创新模块

MODULE A — First-Mover Safety Architecture

（先行者安全架构）

问题

国家不会首先接受约束。

机制

Simultaneous Activation Protocol

流程：

1. 各方内部批准
2. 文件托管中立机构

3. 达到参与阈值
4. 同步全球激活

效果：

- 消除战略劣势
- 破解囚徒困境

MODULE B — Strategic Ambiguity Layer

（战略模糊层）

AGI-COS 对外定义为：

Global AGI Stability Infrastructure

而非控制体系。

目的：

避免主权防御反应。

MODULE C — Dual-Track Adoption Model

（双轨采用机制）

Track 1 — Industry Track

AI 实验室先采用：

- 风险指标
- 熔断接口
- 审计语言

Track 2 — Government Track

政府随后制度化。

MODULE D — Non-Threat Sovereignty Protocol

（非威胁主权协议）

保证：

- 国家决策权完整
- 无超国家权力
- 无强制执行机构

AGI-COS 仅提供协调能力。

——

MODULE E — Civilizational Time-Buying Mechanism

（文明时间机制）

新增核心指标：

CTI — Civilizational Time Index

衡量：

是否为人类争取更多未来选择。

——

PART IV — AGI-COS 七层运行架构

——

Layer 1 — Continuous Sensing

文明连续感知层

监测指标：

- SSI — 系统压力指数
- CI — 耦合指数
- IRI — 不可逆风险指数
- TCS — 信任校准分数
- MOS — 意义选项空间
- CTI — 文明时间指数

——

Layer 2 — Perception Audit

感知审计层

要求：

- 置信度标注
- 数据溯源
- 未知声明
- 对抗性验证

AGI 必须可审计。

Layer 3 — Micro-Intervention Layer

微干预层

特点：

- 小规模
- 低可见
- 可回滚

避免“大改革风险”。

Layer 4 — Crisis Operating System

危机操作层

触发条件：

IRI 或 SSI 达到阈值。

AGI 提供：

- 情景模拟
- 协调建议

人类作最终决定。

Layer 5 — Decision Downgrade Mode

（红队模式）

极端风险时：

AGI 停止建议行动，仅提供：

- 最坏情景
- 假设压力测试

防止自动化依赖。

——

Layer 6 — Adaptive Evolution Layer

MEU 2.1 最小进化单元

允许：

- 地区治理实验
- 制度试点
- 可退出创新

文明通过实验进化。

——

Layer 7 — Learning & Memory Layer

文明学习层

建立：

Global Civilization Learning Archive

保存：

- 危机经验
- 失败案例
- 治理模型

文明避免重复错误。

——

PART V — 运行协议（Operational Protocols）

——

1. Crisis Attachment Protocol

体系不主动推广。

在危机中被引用并自然采用。

2. Invisible Expansion Protocol

隐形扩散原则：

- 永久公开
- 多节点传播
- 不中心化控制

3. Coordination Acceleration Protocol

核心目标：

提升文明协调速度。

PART VI — AGI 与人类关系模型

AGI 是：

- 认知放大器
- 分析伙伴
- 协调工具

AGI 不是：

- 统治者
- 政府
- 价值裁决者

PART VII — 文明健康模型

文明稳定函数：

Civilization Stability =
Coordination Speed
× Reversible Time
× Trust Calibration

PART VIII — 与文明跃迁三元目标对齐

文明目标	AGI-COS 作用
生存无忧	防止不可逆崩溃
生活幸福	维持稳定环境
生命意义	保留未来可能

PART IX — AGI-COS 2.1 的历史定位

AGI-COS 2.1 不是：

- AI 政策文件
- 技术治理方案
- 理想主义宣言

它是：

Civilizational Coordination Infrastructure

（文明协调基础设施）

类似：

- 科学方法
- 宪政原则
- 互联网协议

PART X — 最终声明

AGI-COS 2.1 的目标不是创造完美社会。

而是：

在智能加速时代，
保证人类文明仍拥有选择未来的能力。

AGI analyzes.
Humans decide.
Civilization continues.
