

AGI-COS v1.2

AGI Crisis Operating System

Ecological Resilience Edition (生态韧性版)

一、版本声明

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 (认知负荷优化器)

Adversarial Defense Module (对抗防御模块)

Societal Resonance Loop (社会共鸣回路)

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

二、系统架构更新

AGI-COS 架构在 v1.1 六层结构基础上，强化层间交互逻辑：

Governance Kernel

↓

Perception Audit Layer

↓ (新增交互接口)

Cognitive Interface Optimizer (认知接口优化器)

↓

AGI Perception Layer

↓

Human Decision Layer

↓

Intervention Layer

↓

Recovery & Evolution Layer

注：认知接口优化器并非新增物理层级，而是嵌入在感知审计层与决策层之间的“协议转换器”，确保信息密度匹配人类认知带宽。

三、新增模块 ①

Cognitive Load Optimizer (认知负荷优化器)

3.1 问题背景

在高压危机下，v1.1 要求的详尽审计信息（溯源图谱、置信度细节）可能导致人类决策者出现“认知过载”与决策瘫痪。

3.2 自适应显示协议

系统依据 WUI (Window Urgency Index) 自动调节输出信息密度：

Green Mode (Normal WUI): 显示完整技术细节与数据溯源。

Yellow Mode (Elevated WUI): 折叠次要数据源，仅显示关键偏差与核心建议。

Red Mode (Critical WUI): 启用“决策看板模式”，仅显示：

核心风险提示

可逆性状态

建议行动选项

3.3 行动理据摘要

所有 AGI 建议必须生成一段不超过 50 字的 Action Rationale (行动理据摘要)，确保决策者在 10 秒内理解核心逻辑。

四、新增模块 ②

Adversarial Defense Module (对抗防御模块)

4.1 防御对象

防范目标从 v1.1 的“随机错误/幻觉”扩展至“恶意操纵”：

- 数据投毒
- 对抗样本攻击
- 信息战诱导

4.2 红队测试协议

在关键决策节点，系统必须触发 Shadow Red-Teaming (影子红队测试)：

- 模拟攻击者视角寻找当前 AGI 建议的逻辑漏洞。
- 若发现决策路径存在明显的对抗性诱导特征，自动触发 Security Alert (安全警报) 并冻结该建议。

4.3 数据免疫机制

对输入数据进行“动机审计”：

- 识别异常的数据模式（如非自然的时间序列剧变）。
- 对来源单一或不可信的数据自动执行降权或隔离处理。

五、新增模块 ③

Societal Resonance Loop (社会共鸣回路)

5.1 核心理念

治理的有效性取决于社会的配合度。技术最优解若缺乏社会共鸣，将导致执行失败或次生舆情危机。

5.2 新增指标：SRI — Societal Resonance Index (社会共鸣指数)

衡量：

- 公众对于干预措施的理解度
- 社会情绪的稳定性
- 官方信息信任度 vs 谣言传播速率

5.3 动态反馈机制

当 SRI 低于设定阈值时：

- 执行层调整：暂停高强制性措施，转向沟通与安抚策略。
- 输出层调整：要求 AGI 将“行政术语”转化为“社会通俗语言”进行发布。

六、路径设计约束升级

在 v1.1 “社会文化兼容性”基础上，新增：

执行阻力评估

AGI 提出的干预路径必须包含：

Technical Optimal Path (技术最优路径)

Social Acceptance Estimate (社会接受度预估)

Implementation Resistance Factor (执行阻力系数) ← (新增)

决策层需确认：若执行阻力系数过高，必须优先调整方案或增强沟通资源，而非强行推进。

七、指标体系更新

在原有指标 (CI/WUI/IRI/CSI/PCI) 基础上，新增：

CLI — Cognitive Load Index (认知负荷指数)

- 衡量当前决策信息量是否超过人类处理极限。
- CLI 过高时，系统自动强制简化输出界面。

SRI — Societal Resonance Index (社会共鸣指数)

- 实时量化社会层面的接受度与配合度。

八、人机协同规则更新

引入 "Dual-Factor Veto" (双因子否决) 逻辑：

规则 A：置信度熔断

若 $PCI < 20\%$ (感知置信度极低)：

- 系统禁止推荐“不可逆行动”。
- 仅允许“观察、防御性隔离、试探性行动”。

规则 B：社会共鸣降权

若 $SRI < 20\%$ (社会共鸣极低)：

- 即使 AGI 建议置信度很高，系统也自动将 AGI 权重降低一档。
- 强制引入“人类共识会议”程序，避免技术与民意脱节。

九、向后兼容声明

AGI-COS v1.2：

- 完全继承 v1.0 的原则与 v1.1 的审计机制。
- v1.1 实施者只需更新“接口层协议”与“指标监测集”即可升级。
- 无需重构核心治理架构。

十、核心命题（再进化版）

AGI-COS 的目标不仅是保留选择未来的能力，

更是要构建一种机制，

让人类在机器智能的辅助下，

依然能够从容、清醒、且富有同理心地行使文明的主权。

End of Release