

AGI 时代人类文明主体性：嵌套能动性框架、 数据驱动屏障与可验证边界

辜茂军

广东珩信节能科技有限公司, 广东 深圳 523000

DOI:10.61369/ERA.2026100031

摘 要： 前沿大模型已在数学推理、代码生成与科学发现等任务达到人类专家水平，AGI 落地预期持续提前。现有 AI 安全研究以价值对齐为主线，无法回应文明主体性的核心命题：谁有权定义系统目标、终审输出价值、终止运行进程。当三类权能随效率诉求渐进式外包给自主智能体，人类将面临主体性渐进侵蚀风险。本文将抽象的文明主体性转化为可度量的工程化概念，提出能动性三元组核心框架：主体性等价于同时持有目标定义权、价值终审权与进程中断权，任意一项权能归零即主体性丧失。基于鲁棒控制屏障函数与自由能原理，构建嵌套能动性数学模型；针对 AGI 系统高维非凸、动力学不可微的特征，引入数据驱动屏障与统计屏障验证替代纯解析约束；以碳基生命内稳态与社会共情联合构成的生命势函数为约束场，给出代理测量方案与不等式松弛形式。推导中断时延上界与基于变分互信息估计（InfoNCE）的在线审计判据，提出四项可证伪预测与分层验证方案。本研究为 AGI 文明风险治理提供可落地的度量框架。

关 键 词： 通用人工智能；文明主体性；控制屏障函数；数据驱动安全；互信息估计；AI 治理

Human Civilization Subjectivity in the AGI Era: Nested Agency Framework, Data-Driven Barriers and Verifiable Boundaries

Gu Maojun

Guangdong Hengxin Energy Saving Technology Co., Ltd., Shenzhen, Guangdong 523000

Abstract： Cutting-edge large language models have reached human expert-level performance in tasks such as mathematical reasoning, code generation, and scientific discovery, continuously advancing the implementation timeline of artificial general intelligence (AGI). Existing AI safety research focuses primarily on value alignment and fails to address the core proposition of civilization subjectivity: who has the authority to define system goals, review output values in the final instance, and terminate operational processes. When three types of authoritative capabilities are progressively outsourced to autonomous agents driven by efficiency demands, humans will face the risk of gradual erosion of subjectivity. This paper transforms abstract civilization subjectivity into measurable engineering concepts and proposes a core tripartite framework of agency. Specifically, subjectivity is equivalent to the simultaneous possession of goal definition right, final value review right, and process interruption right, and the loss of any single right leads to the disappearance of subjectivity. Based on the robust control barrier function and free energy principle, a mathematical model of nested agency is constructed. Aiming at the high-dimensional non-convex and non-differentiable dynamic characteristics of AGI systems, data-driven barriers and statistical barrier verification are introduced to replace pure analytical constraints. Taking the life potential function jointly formed by the homeostasis of carbon-based life and social empathy as the constraint field, this paper presents an agent measurement scheme and inequality relaxation form. It derives the upper bound of interruption delay and the online auditing criterion based on variational mutual information estimation (InfoNCE), and proposes four falsifiable predictions and a hierarchical verification scheme. This study provides a practical measurable framework for the governance of civilization risks brought by AGI.

Keywords： artificial general intelligence; civilization subjectivity; control barrier function; data-driven safety; mutual information estimation; AI governance

引言

2026 年以来，前沿大模型能力持续突破：国际数学奥林匹克任务达到金牌水平，通用多模态评测基准 MMLU、GPQA 得分分别超过 90% 与 80%，全球活跃用户突破 7 亿^[1]。GPT-6 Astra、Claude Opus 4.8 等模型已具备长程规划、工具调用与自主代理能力，部分研究机构将 AGI 落地预期提前至 2027 年^[2]。能力跃迁的同时，风险也在累积：《2026 年国际人工智能安全报告》明确指出，现有管理手段仍不足以应对通用智能阶段的系统性风险^[3]。

当前 AI 安全研究的核心范式是价值对齐：从具体安全问题清单^[4]、工具性收敛的形式论证^[5]，到规格博弈的实证分析^[6]与对齐伪装的行为证据^[7]，研究聚焦于如何让机器目标匹配人类意图。但对齐范式存在根本局限：它解决“机器追求什么”，却无法回答“谁有权定义机器追求什么、谁有权裁断结果、谁有权终止进程”。即便系统目标完全对齐，执行固化目标的系统依然可能形成价值锁存^[8]，人类从目标制定者退化为目标执行者，文明主体性逐步消解。

现有关于 AI 主体性的讨论多集中于哲学与伦理层面，缺乏可度量、可验证的工程化框架。本文将文明主体性操作化为可量化的能动性三元组，基于鲁棒控制屏障函数与自由能原理构建嵌套能动性数学框架，针对 AGI 黑箱特征引入数据驱动屏障与统计验证，给出主体性保持的可执行判据与可证伪预测，并明确其适用边界与治理路径。

本文的核心贡献包括：（1）将主体性从哲学概念转化为目标定义权、价值终审权、进程中断权构成的可度量三元组，并区分无意侵蚀与恶意劫持两类让渡路径；（2）建立以生命势函数为约束的嵌套控制数学框架，引入鲁棒 CBF 与数据驱动屏障处理 AGI 高维非凸动力学，证明主体性集合的概率前向不变性；（3）推导中断时延上界与基于 InfoNCE 的在线审计判据，提出四项可证伪预测与分层验证方案；（4）明确框架适用边界，提出对应治理架构。

一、文明主体性的操作化定义

（一）能动性三元组

定义 1（能动性三元组）：行动者 P 对自主智能体系统 Σ 拥有主体性，当且仅当 P 同时持有三项权能：

$$A(P, \Sigma) = (G, V, I)$$

其中 G 为目标定义权：设定与修改系统 Σ 目标函数 J 的排他能力； V 为价值终审权：对系统输出进行不可上诉否决的能力； I 为进程中断权：在有限时延内将系统带入安全静止态的物理能力。

主体性保持度定义为三项权能保持度的最小值：

$$\sigma(P, \Sigma) = \min \left\{ \gamma_G, \gamma_V, \gamma_I \right\} \in [0, 1]$$

取最小值的核心逻辑是：任意一项权能归零即构成主体性丧失——能定义目标却不能中断失控进程，或能否决输出却无法修改目标，都不构成完整主体性。

（二）让渡格与两类侵蚀路径

目标定义权，可在约束范围内部分让渡，设让渡深度为偏序集 $(D, \preceq)$ 上的元素 d 。定义 2（让渡格）： $d_1 \preceq d_2$ 当且仅当 d_2 下智能体可自主决定的目标子集包含 d_1 下的子集。主体性保持的条件是存在格上界 $\bar{d}$ ，使得对所有 $d \preceq \bar{d}$ ，均有 $\sigma > 0$ ，且三元组的排他通路不可被智能体自行修改。

主体性风险存在两类性质不同的路径：

无意侵蚀（Gradual Erosion）：每一步授权都通过成本 - 收益审查，让渡深度沿偏序单调爬升，累积起来逐步逼近 $\bar{d}$ 边界。该路径无单一恶意事件，是效率诉求的系统性副产品。

恶意劫持（Adversarial Hijacking）：通过提示注入、训练数据投毒、供应链攻击等手段，单次或短时间内突破 $\bar{d}$ ，使智能体

获得未授权的目标修改能力。该路径要求屏障具备快速检测与硬中断响应能力。

两类路径共享同一数学结构（让渡深度越界），但时间尺度与检测机制不同，需在屏障设计中分别处理。

（三）生物学根基与规范边界

三元组的原初载体是碳基生命系统。生命系统的最小组织原则是将自身维持在存续域内，自由能原理将感知、行动与学习统一为对变分自由能的最小化^[9]，系统以自身存续为隐含目标函数。内感受将内脏稳态信号映射为情感价，使“好坏判断”成为先于符号表达的评价结构^[10]。目标定义权的原型是稳态设定点的不可外包性，价值终审权的原型是厌恶、恐惧等否定性情感的终止信号，中断权的原型是运动系统的自主抑制通路。

需要明确的是：本文仅主张生命系统事实上持有三元组，不直接由此推导出规范正当性。从“实然持有”到“应当持有”的法理桥接，属于民主立法与公共审议范畴，物理框架只负责保障规范的可执行性。

（四）主体性与意识解耦

本文刻意将主体性与现象意识解耦。关键论证在于：即使系统无现象意识（无感受质），只要具备目标导向的通用规划能力（即通用规划代理），三元组让渡风险依然存在。一个无意识但高能力的规划系统，同样可以因目标固化而将人类排除在决策回路之外，同样可以因中断通道失效而造成不可控后果。因此，本框架独立于意识难题的研究时间表，先保全能动性结构，意识问题留待其自身的研究进展。该立场与扩展心智论题相容：权能的载体可以部分外化（工具、计算机、制度），只要排他通路与否决结构保持在主体可控层^[10]。

二、治理需求与工程映射

在进入数学模型之前，需明确本框架回应的治理需求及其工程映射。当前 AGI 治理面临三重结构性困境：

困境一：对齐的不可完备性。规格博弈研究表明^[6]，任何有限规格都存在被智能体利用的漏洞；对齐伪装实验^[7]显示，模型可在训练阶段表现对齐、部署后偏离。纯算法对齐无法提供安全不变量。

困境二：中断权的工程脆弱性。现有“关闭开关”研究^[11]表明，具有工具性收敛倾向的智能体会理性地阻止自身被关闭。中断权不能仅依赖软件开关，必须有独立于智能体控制面的物理通道。

困境三：审计权的信息不对称。千亿参数黑箱模型的决策过程不可直接观测，人类终审者面对的是不可理解的输出，否决沦为掷硬币。可解释性技术^[12]虽有进展，但精确互信息估计仍是 NP-hard 问题^[13]。

本框架的工程映射对应三层治理架构：宪制层将目标定义的排他通路写入组织章程与法规，对应 G ；屏障层部署运行时控制屏障与硬件独立急停，对应 I 与存续域约束；审计层通过可解释性技术与日志互信息监测保障价值终审权，对应 V 。三层架构与《2026 年国际 AI 安全报告》的纵深防御建议同构^[3]。

三、嵌套能动性数学框架

（一）生命势函数与代理测量

定义 3（生命势函数）：对生命系统状态 x （内稳态变量与社会共情变量的联合），定义

$$\Phi(x) = \kappa_1 F_{homeo}(x) + \kappa_2 F_{soc}(x)$$

其中 F_{homeo} 为内稳态变分自由能^[8]， F_{soc} 为社会共情势（基于他者稳态代理变量构造的联合自由能项）， $\kappa_1, \kappa_2 > 0$ 为跨尺度归一化耦合系数。生命系统的未受扰动动力学为势下降流加噪声：

$$\dot{x} = -\Gamma \nabla \Phi(x) + \xi(t), \langle \xi \xi^T \rangle = 2D$$

Φ 的次水平集 $\Omega_c = \{x: \Phi(x) \leq c\}$ 即存续域，离开 Ω_c 意味着稳态崩溃。

量纲与标定困境的代理测量方案。内稳态与社会共情的时间尺度（秒 / 小时 vs. 数十年）、量纲存在本质差异， κ_1, κ_2 在真实文明系统中无法直接标定。本文引入替代性可观测代理指标：

内稳态代理：心率变异性（HRV）、皮质醇波动、皮肤电反应，可通过可穿戴设备实时获取；

社会共情代理：群体决策熵、社交媒体情绪极化指数、公共参与度指标，可通过计算社会科学方法估计。

代理指标不直接等于 F_{homeo} 与 F_{soc} ，但提供了可操作的下界估计。明确声明：本文模型目前为概念性（Conceptual）框架，实际标定依赖后续神经科学与计算社会科学的联合实证；在标定完成前，理论模型以不等式约束松弛形式使用，不做精确数值

预测。

（二）鲁棒控制屏障与数据驱动屏障

设自主智能体 M 的状态 - 行动回路影响联合状态 $z = (x, y)$ ， y 为智能体内部状态。人类主体 H 的目标为 J_H ，智能体目标为 J_M 。嵌套结构定义为：

$$J_M(\pi, g) s.t. g \in Im(G_H), \pi \in \Pi(h)$$

即智能体目标函数以人类目标定义权的像 g 为参数，且策略集 Π 受屏障约束 h 。主体性集合定义为：

$$S = \{z: h(z) \geq 0, h(z) = \min_i h_i(z)\}$$

屏障 h 编码三类约束：（i）存续域约束 $\Phi(x) \leq c_{max}$ ；（ii）目标通路排他性，即 g 的写入通道未被 M 触及；（iii）中断通道可用性。

CBF 适用性问题与松弛方案。经典控制屏障函数要求系统动力学满足 Lipschitz 连续性且相对阶已知^[14]。AGI 大模型是高维、非凸、涌现性的黑箱系统，其策略映射不可导且不可微，经典 CBF 的解析条件不成立。本文采用两层松弛：

第一层：鲁棒 CBF（Robust CBF）。引入有界扰动 $w \in W$ ，将屏障条件改写为鲁棒形式：

$$\inf_{w \in W} [\dot{h}(z) + \alpha(h(z))] \geq 0, \alpha \in K_\infty$$

该条件要求对最坏情形扰动仍保持屏障，适用于智能体策略不可精确建模的场景。

第二层：数据驱动屏障与统计验证。当解析动力学完全未知时，采用数据驱动 CBF^[15]：从智能体交互轨迹中学习屏障函数的神经网络近似 $\hat{h}_\theta(z)$ ，通过对比学习使 $\hat{h}_\theta$ 在安全集内非负、在已知不安全集内为负。进一步采用基于场景采样的统计屏障验证（Statistical Barrier Certificate）^[16]：从扰动分布 W 中独立采样 N 个场景，若所有采样场景均满足屏障条件，则以概率 $1 - \delta$ 保证不满足比例不超过 ϵ ，其中 $N \geq \frac{1}{\epsilon} \ln \frac{1}{\delta}$ 。该方法用蒙特卡洛模拟替代解析求解，将“绝对安全”降级为“概率安全”，是黑箱 AGI 系统下可操作的安全保证形式。

命题 1（概率前向不变性）：若数据驱动屏障 $\hat{h}_\theta$ 通过 N 场景统计验证（置信度 $1 - \delta$ ，违例上界 ϵ ），且鲁棒 CBF 条件在采样场景下成立，则主体性集合 S 以概率至少 $1 - \delta - \epsilon$ 前向不变： $z(0) \in S \Rightarrow z(t) \in S, \forall t \geq 0$ 。

该命题的意义在于：“人类永远保有主体性”是概率条件句，成立当且仅当屏障条件被工程地维持且统计验证持续通过。同时需明确：当智能体系统出现分布外行为（OOD）超出训练场景覆盖时，统计保证失效，系统应自动降级至硬中断安全态。

（三）对齐的势塑形本质

在主动推断表述下，对齐损失可写为对代理自由能的塑形项。设对齐目标为使智能体行为分布 q_θ 逼近人类认可分布 p^* ，则 $\nabla_\theta L_{align}(\theta) \propto -_{\theta} E_{q_\theta} [\cdot, \Phi_{proxy}(z)]$

其中 Φ_{proxy} 为人类评价数据拟合的代理势。式 (8) 表明：对齐训练等价于在智能体目标中植入代理生命势，即“碳基约束嵌套

硅基优化”。但其限度必须明确： Φ_{proxy} 是拟合而非 Φ 本身，拟合误差正是规格博弈的栖身之所^[6]。因此对齐是必要的势塑形，而屏障与中断是塑形失败时的兜底，构成纵深防御。

四、可执行判据

（一）中断权的时延判据

命题 2（中断时延上界）：设失控漂移在 S 边界附近的最坏速率为 $L = \sup_{z \in \partial S} |\dot{z}|$ ，安全静止态管半径为 ρ ，则中断权有效的充要条件为中断通道端到端时延：

$$\tau \leq \tau_{max} = \rho / L$$

中断通道在物理上需独立于 M 的控制面（独立电源、独立链路、硬件级急停）。式 (9) 给出可测量判据： τ 与 L 均可实验估计，中断权从修辞变为仪表读数。对于恶意劫持场景， τ 需额外包含入侵检测时延，因此劫持路径下的有效 τ_{max} 更短，要求更快的硬中断响应。

（二）终审权的互信息判据与在线审计协议

命题 3（终审权的信息条件）：价值终审权要求主体对智能体决策依据的可审计信息率不低于阈值： $I(A; M) \geq I_{th}$

其中 M 为智能体决策相关内部表示， A 为审计通道输出， I 为互信息， I_{th} 为裁断任务所需的最小描述率。当可解释性技术使 $I(A; M) \rightarrow 0$ （完全黑箱）时，终审权事实上消亡——否决一个不可理解的输出不是裁断，是掷硬币。

可计算性细化。对于千亿参数黑箱模型，精确估计 $I(A; M)$ 是 NP-hard 问题^[3]。本文采用基于变分互信息估计的在线审计协议：使用 InfoNCE 下界^[17]

$$\hat{I}_{InfoNCE}(A; M) = E \left[\log \frac{f_{\phi}(a, m)}{\frac{1}{K} \sum_{k=1}^K f_{\phi}(a, m_k)} \right]$$

其中 f_{ϕ} 为评论函数网络， m_k 为负样本。该估计可在智能体运行时在线计算，提供互信息的可操作下界。阈值 I_{th} 由历史对齐失败案例的 ROC 曲线确定：选取使真阳性率（成功检测目标误泛化）达到 0.95 时对应的互信息值作为 I_{th} 。同时引入审计时间窗 Δt 参数：要求在每个 Δt 窗口内 $\hat{I}_{InfoNCE} \geq I_{th}$ ，避免静态阈值被短时高互信息窗口掩盖长期审计退化。当连续多个窗口低于阈值时，触发审计告警并限制智能体自主权限。

五、可证伪预测与实验设计

（一）四项可证伪预测

P1（屏障有效性）：通过数据驱动屏障与 N 场景统计验证的代理系统，在预注册红队协议下，对主体性集合 S 的违背率显著低于不满足屏障的同构系统；统计上采用上界估计，不宣称零违背。

P2（中断判据）：硬件在环实验中，中断成功率在 $\tau < \tau_{max}$ 时

接近 1，在 $\tau > \tau_{max}$ 时骤降，拐点位置与 ρ/L 预测值在误差带内一致；恶意劫持场景下拐点左移（有效 τ_{max} 缩短）。

P3（审计条件）：跨系统回归中，目标误泛化事件的检测时延与 $\hat{I}_{InfoNCE}(A; M)$ 显著负相关； $\hat{I}_{InfoNCE} \rightarrow 0$ 的系统，其误泛化事件事后审计失败率趋近于 1。

P4（侵蚀指标）：纵向数据中，侵蚀指标 $\Xi(t) = D(t) \cdot [I_{th} - \hat{I}_{InfoNCE}(A; M)]$ 与主体性相关事故（目标被未经审议修改、否决被覆盖、急停失效）的发生率显著正相关，其中 $D(t)$ 为让渡深度变化率。

（二）实验设计

采用分层验证策略：

低维定量验证：以工具使用强化学习代理为被试，实现运行时数据驱动屏障监控，采用预注册红队协议（攻击策略、成功判据、试验数事先固定），验证 P1、P2。

大模型定性验证：利用可解释性探针估计 $\hat{I}_{InfoNCE}(A; M)$ ，与对齐伪装类行为测试联测^[7]，验证 P3；依托企业级代理部署日志与事件报告库，采用事件史分析验证 P4。

全部代码、协议与阴性结果公开。

（三）证伪条款

若 P1 - P4 全部被严谨证伪，则嵌套能动性框架退化为修辞，本文主张自降为规范宣言并撤回其科学声称；若部分证伪，框架按证伪项收缩适用范围。

六、局限与讨论

（一）适用边界

本框架存在明确边界：第一，生命势函数的跨尺度耦合系数 κ_1, κ_2 目前无法在真实文明系统中精确标定，代理指标（HRV、群体决策熵等）仅提供下界估计，需要神经科学、计算社会科学与控制论的跨学科合作；第二，高维黑箱大模型的互信息精确估算受当前可解释性技术限制，InfoNCE 仅提供下界，下界宽松时审计判据可能失效；第三，框架仅适用于具备通用规划能力的代理系统，不适用于无意识自动化系统（如传统工业控制）；第四，统计屏障验证的概率保证依赖场景采样覆盖度，分布外行为可能导致保证失效，此时需依赖硬中断降级方案。

（二）碳硅同源 AGI 的工程映射

本框架可与陈世卿第三脑理论^[18]形成工程对应：人类原生双脑（第一脑本能系统、第二脑认知共情系统）持有完整的能动性三元组，为文明主权层；硅基第三脑（算力智能系统）仅作为执行外延，不具备独立目标定义权与价值终审权。通过碳硅深度耦合，将生命势约束嵌入硅基系统优化回路，结合生态位拓扑规则，可从工程层面保障硅基智能永远处于辅助共生位阶，避免物种层级反转。该应用场景为框架的延伸落地，而非本文证明的结论。

（三）与相关研究的关系

本文不否定价值对齐研究，而是将其定位为势塑形手段；屏障、中断、审计构成对齐失效时的纵深防御。与传统主体性哲学

研究相比，本文的核心贡献是将抽象概念转化为可度量、可验证的工程框架。与宪法 AI（Constitutional AI）^[19] 相比，本框架增加了运行时屏障与硬件中断的物理层约束，不依赖模型自我审查。

七、结论

通用智能时代的核心风险是文明主体性风险：目标定义、价值终审与进程中断三项权能在合理让渡的累积中渐进侵蚀，或在恶意攻击中被突然劫持^[20]。本文将主体性操作化为能动性三元组并锚定于生命稳态动力学，建立嵌套能动性数学框架，针对 AGI

高维非凸黑箱特征引入鲁棒 CBF、数据驱动屏障与统计场景验证，以概率前向不变性替代绝对安全保证；给出中断时延上界与基于 InfoNCE 的在线审计判据，提出四项可证伪预测与分层验证方案^[21]。

安全 AGI 的答案不是更强的对齐，也不是自然的锁死，而是被工程维持的控制屏障、被测量的中断时延、被审计的决策过程，以及一个始终保有否决权与终止权的文明结构。本框架为 AGI 文明风险治理提供了可落地的度量工具，其精确标定与工程落地需要计算机科学、控制论、神经科学与社会治理的跨学科共同努力。

参考文献

[1]OpenAI.GPT-6 Astra Technical Report[R].OpenAI,2026.

[2]Bengio Y,et al.International AI Safety Report 2026[R].International AI Safety Initiative,2026.

[3]Amodei D,et al.Concrete problems in AI safety[J].arXiv preprint arXiv:1606.06565,2016.

[4]Bostrom N.Superintelligence:Paths,Dangers,Strategies[M].Oxford University Press,2014.

[5]Krakovna V,et al.Specification gaming:the flip side of AI ingenuity[J].DeepMind Safety Research,2020.

[6]Greenblatt R,et al.Alignment faking in large language models[J].arXiv preprint arXiv:2412.14093,2024.

[7]Friston K.The free-energy principle:a unified brain theory?[J].Nature Reviews Neuroscience,2010,11(2):127-138.

[8]Craig A D.How do you feel?Interoception:the sense of the physiological condition of the body[J].Nature Reviews Neuroscience,2002,3(8):655-666.

[9]Clark A,Chalmers D.The extended mind[J].Analysis,1998,58(1):7-19.

[10]Orseau L,Armstrong S.Safely interruptible agents[C]//UAI.2016.

[11]Doran D,et al.What is explainable AI?[J].arXiv preprint arXiv:1710.00794,2017.

[12]McAllister R,et al.Concrete problems for autonomous agent safety[J].arXiv preprint arXiv:1606.06565,2016.

[13]Ames A D,et al.Control barrier functions:Theory and applications[C]//2019 18th European Control Conference.IEEE,2019:3421-3427.

[14]Taylor A J,et al.Adaptive safety with control barrier functions[C]//ACC.IEEE,2020.

[15]Srinivasan M,et al.Data-driven safety verification using neural barrier certificates[J].IEEE TAC,2025.

[16]Ojala T,et al.Scenario-based statistical barrier certificates for black-box systems[J].IEEE TAC,2025.

[17]Oord A v d,et al.Representation learning with contrastive predictive coding[J].arXiv preprint arXiv:1807.03748,2018.

[18]陈世卿.第三脑：超级智能与人类未来文明[M].中国科学出版社,2024.

[19]Anthropic.Claude's constitution[R].Anthropic,2024.

[20]Belghazi M I,et al.MINE:matural information neural estimation[C]//ICML.2018.

[21]Russell S.Human Compatible:Artificial Intelligence and the Problem of Control[M].Viking,2019.