arXiv:2603.03119v1 [cs.AI] 3 Mar 2026

AI Space Physics：开放人工智能机构的构成性边界语义

Oleg Romanchuk Roman Bondar

摘要

智能体式 AI 部署日益呈现持久性机构的行为，而非一次性的推理端点：它们累积状态、调用外部工具、协调多个运行时，并随着时间推移改变自身未来的权限界面。现有治理语言通常规定决策层约束，却未充分界定边界跨越的因果机制——尤其是那些不会立即改变外部世界、却会扩展机构未来行动能力的转移 [1], [2]。

本文提出 **AI Space Physics**，将其作为开放式、自我扩张的人工智能机构的一种构成性语义。我们定义了一个最简状态模型，其中包含类型化边界通道、视界受限的可达性语义，以及膜见证规程。核心定律族（*P-1*、*P-1a*、*P-1b*、*P-1c*）要求见证完整性、不可绕过的中介、原子化裁决到效果的转换，以及对裁决类别的可重放重构。我们明确将二阶效应区分为结构扩展（SECOND_T）和政策扩宽（SECOND_P），并将扩展转换视为与治理相关，即使即时外部变化量为零。

基于这一语义，我们推导出关于扩展可检测性、外生钩子作用下外部任务的因果关系、有限观察者不可辨识性，以及明确延续假设下稀缺延续体制的命题。我们还提出一项操作设计推论（*H-1*），将风险加权扩张压力与审查带宽联系起来。该框架独立于具体实现：模型检测和工程产物可提供验证证据，但并不构成理论本身 [3], [4]。

我们的创新主张明确而不夸大：本文并未将中介作为一个概念引入，而是将权限表面扩张重新归类为一种具有构成性见证义务的一级边界事件。在这一语义中，即便没有立即提交，扩张仍与裁决相关。

在本文中，物理学 指支配因果结构的构成性转移语义，而非经验物理科学。开放 指通过通道介导、与外部耦合且能够自我扩展的制度，并非任何特定的

供应商或产品。

1. 引言

自主人工智能系统日益以长生命周期、可使用工具的进程形式部署，通过外部接口在多个迭代周期中进行观察、规划和行动，且无需人为逐步触发。在这一情形下，治理失效并非仅源于孤立的“错误行动”；能力累积、拓扑增长和政策表面扩展也会逐渐发生，在不易察觉的情况下扩大未来的外部可达性 [5], [6]。近期智能体安全基准测试和部署研究报告指出，在真实的智能体工作流中，工具滥用、间接提示注入以及权限过宽的行动链仍很常见[7], [8], [9]。

经典治理文档（宪法约束、政策文本、监督工作流）不可或缺，但如果无法映射到一种从构造上杜绝绕过路径的转移语义，它们便不足以发挥作用。这与既有安全原则相呼应：只有当每项安全相关操作都经由一个不可绕过的控制点时，完全中介才有意义；而只要有任何有效通道逃逸出该中介，遏制就会失效[1], [2]。

一句话定位：本文为视界受限可准入性条件下的制度权威增长提供转移语义。

这一物理层补充了此前对智能体系统中认知担保失效和组织问责的分析 [10], [11], ，也可视为与运营治理框架相契合的机制层 [12], [13], [14]。这些框架对于生命周期风险管理和控制分类体系很有价值，但它们并未定义可达集包含或权威表面质量的转移语义。

1.1 问题陈述

系统在步骤 t 可能不会立即产生外部提交，但仍会通过增加工具、特权、运行时实例或连接性，提高其在 $t+k$ 的预期外部影响。如果治理仅跟踪一阶效应，就可能在制度不断积累潜在权威质量的同时，为表面合规背书。在智能体部署中，这种不匹配是结构性的，而非偶然的，因为工具与拓扑变更属于常规运营操作，而不是罕见异常 [5]。

所需的转变是按效应阶次对边界转移进行类型化，并以构成性不变量同时覆盖直接提交与扩展可达范围的转移。目标并非对隐藏的内部认知进行完整重构，而是对外部因果作用变得真实且可审计的边界实施可强制执行的控制。

经典治理框架很少将可达集包含或权限表面增长形式化为类型化转移对象。在此，扩展被重新归类为具有明确裁决与见证义务的边界相关转移类别，从而弥补治理语义中的结构性盲点。

在操作层面，未立即提交的扩展仍属于外部因果潜能，必须作为边界事件加以治理。

1.2 贡献

本文提出六项相互关联的成果。

1. 一个开放人工智能机构的形式模型，包含类型化通道、能够感知扩展的状态以及明确的外部投影。2. 一套关于 Cell、Unit 和 Membrane 的定义层，使治理语言与可执行转移语义相一致。3. 一个构成法则族：

- *P-1* – 边界见证，
- *P-1a* – 不可绕过，
- *P-1b* – 原子裁决到效应，
- *P-1c* – 可重放裁决。4. 一组命题，明确列出外部任务因果和稀缺延续体制的假设。5. 一项操作设计推论（*H-1*），将风险加权扩张压力与审查带宽联系起来。6. 一个最简实例，说明为何即使未发生 FIRST 转换，也必须对 SECOND_T 作出裁决。

1.3 阅读指南

第 2 节定义符号和基元对象。第 3 节定义系统类别和转移体制。第 4 节阐述本构定律。第 5 节阐述导出命题及一项操作设计推论。第 6 至 8 节讨论影响、范围条件和结论。附录 A 提供精简的符号表，附录B扩展证明概要，附录 C 将理论对应到工程模式（非规范性），附录 D 给出 *H-1* 的操作形式（非规范性）。

1.4 核心陈述（强可治理性准则）

核心论断。只有当每一阶和二阶外部效应都由不可绕过的膜进行调解，并通过原子见证语义（Decide $\rightarrow$ 锚定 $\rightarrow$效应）验证，且裁决类别能够依据已声明策略和上下文假设通过重放重建时，人工智能机构才具有强可治理性。

允许存在规约之外的外部效应转移路径的架构或许仍可运行，但不具备本文所定义的强意义上的可治理性。

命题 0（强可治理性，必要条件）。如果在某个边界相关转移类别中，违反了 *P-1a*（不可绕过）、*P-1b*（原子裁决到效应）或*P-1c*（可重放裁决）中的任一项，则在此框架下，该制度不具备强可治理性。

阐释：强可治理性并非政策愿景，而是一种转移语义属性，要求具备不可绕过的中介、裁决到效应的时间原子性，以及可重放见证语义。

2. 预备知识与符号约定

2.1 核心对象（通俗说明）

Cell 是能够处理输入、产生输出并跨越边界进行交互的最小自主计算实体。**Unit** 是制度中的运行时操作实例；其能力、预算、策略和见证纪律均受到约束。**膜**是裁决界面，用于对边界相关转移进行分类、发布决策（允许|拒绝|隔离），并锚定相应的见证记录。

这些对象将实践中常被混为一谈的治理层区分开来：模型认知、运行时控制与外部生效执行。这种区分是必要的，因为外部安全取决于跨通道所允许的转移类别，而非内部叙事连贯性[15]。

2.2 核心对象（形式化）

我们现在对先前非正式引入的对象进行形式化。目标并非本体论上的完备性，而是边界治理语义的最小充分性。

全文使用的类型化承载集：

> Cell、Unit、膜 : 类符号（类型），$\mathcal{C}$: Cell 实例的承载集，$\mathcal{U}$: Unit 实例的承载集，$\mathcal{W}$: 见证记录的承载集。

我们将 $c \in \mathcal{C}$、$u \in \mathcal{U}$ 与 $w \in \mathcal{W}$ 记为类型化元素。对于所分析的每一制度，膜是唯一的（如第 2.3 节所述）；不同制度可以实例化类为膜的不同膜实现。

以下使用的符号约定：

- 对于符号名称 X，X 表示排序符号或模式字段标签；
- 对于符号名称 X， $\mathcal{X}$ 表示一个承载集；
- 小写斜体符号（例如 c、 u、 w）表示类型化实例/参数。

$$\begin{aligned}\mathsf{CellSchema} := (&\mathsf{Compute}, \mathsf{Model}, \mathsf{Runtime},\\ &\mathsf{IO}, \mathsf{Boundary})\end{aligned}$$

单元实例 $c \in \mathcal{C}$ 实例化 Cell 实例的结构模式。这里，Compute 表示物理资源约束，Model 表示生成基底，Runtime 表示控制逻辑，IO 表示接口表面，Boundary 表示边界调解槽位。这些是模式字段标签（槽位名称），而非状态变量。此分解提取了实现对外有效行动所必需的最小组件。

$$\begin{aligned}\mathsf{UnitSchema} := (&\mathsf{RuntimeLoop}, \mathsf{CapabilitySet},\\ &\mathsf{BudgetSlice}, \mathsf{PolicyBinding},\\ &\mathsf{WitnessBinding})\end{aligned}$$

Unit 实例 $u \in \mathcal{U}$ 实例化 Unit 实例的结构模式，表示制度内某个 $c \in \mathcal{C}$ 的运行时实例。它明确列出了与治理相关的约束：能力范围、预算分配、活动策略绑定和见证纪律。区分 Cell 与 Unit，可避免将物理计算资源的存在与符合宪制要求的权限混为一谈。

决策类别域:

$$\begin{aligned}D = \{&\mathrm{ALLOW}, \mathrm{REJECT},\\ &\mathrm{QUARANTINE}\}\end{aligned}$$

这些是膜级裁定结果。对于边界中介而言，其他决策类在语义上都没有意义。

内部／边界耦合分类:

$$\begin{aligned}\mathrm{Inside}(x) := \mathrm{PolicyControl}(x) &\wedge \mathrm{StopControl}(x)\\ &\wedge \mathrm{WitnessCoverage}(x)\end{aligned}$$

$$\begin{aligned}\mathrm{BoundaryCoupled}(x) := \mathrm{MembraneMediatedAccess}(x)\\ \wedge \neg\, \mathrm{Inside}(x)\end{aligned}$$

这些谓词将部署分类形式化。只有当制度能够（1）通过策略治理组件的行为，（2）停止或撤销

该组件，并且（3）确保覆盖其边界相关行动的见证时，该组件才属于内部。通过由膜治理的调用可达、但缺乏完整停止／撤销权限的组件属于边界耦合。

这些谓词定义的是系统级分类不变量，而非观察性启发式。它们区分了架构控制保证与单纯的网络可达性。

2.3 状态、通道与外部投影

制度状态分解为边界相关推理所需的各个组成部分：

我们始终采用离散转移时间：$t \in \mathbb{N}$ 表示转移步索引（逻辑时间）。

$$S_t = (S_{int,t}, S_{ext,t}, S_{ledger,t}, \\ S_{budget,t}, S_{topo,t}, A_{adm,t})$$

其中：

- S_{int} 表示内部计算与工件状态；
- S_{ext} 记录外部生效的世界状态；
- S_{ledger} 记录仅追加的见证与裁定记录；
- S_{budget} 记录与可受理性和延续相关的资源约束；
- S_{topo} 记录运行时总体和连通结构；
- A_{adm} 记录用于有限视界可达性分类的当前可允许性配置。

符号说明：S_{int}、S_{ext}、S_{ledger}、S_{budget}、S_{topo}是机构状态的分量投影；因此，$S_{int,\ t} = S_{int}(S_t)$（其他分量同理），而$A_{adm,\ t}$表示步骤 t时生效的可允许性配置。构成性约定：用于 PolicyExpand$_H$ 分类的策略版本和可允许性配置差异由$A_{adm,\ t}$表示；它们不被建模为$S_{int,\ t}$中的构成性状态变化。

这一分解虽是最小的，但已足够：任何边界相关效应都必须表现为这些组成部分中至少一个发生变化。

拓扑状态定义如下：

$$S_{topo,t} = (N_t, G_t) \\ \text{where } G_t = (V_t, E_t)$$

其中：

- N_t 表示活跃运行时总体；

- G_t 是一个用于描述通信和委派结构的有向图。

V_t 是在时间 t处于活跃状态的运行时顶点集合，其中 $V_t \subseteq N_t$。 $E_t \subseteq V_t \times V_t$ 是表示运行时顶点之间通信、委派或承载权限的连接性的有向边集。

对 S_{topo} 的更改代表运行时总体或连接性的结构修改。这些转移可能扩大制度未来的可容许权限面，之后会被正式归类为二阶效应。

下文使用的投影记号：

- $caps(s)$ 表示从状态 s中提取的能力指派；
- $tools(s)$ 表示从状态 s中提取的活动工具/连接器表面；
- $V(s)$、$E(s)$ 表示从 $S_{topo}(s)$中提取的拓扑投影。

边界信道集：

$$C = \{C_{net}, C_{fs_shared}, C_{exec}, C_{money}, \\ C_{deploy}, C_{comm}, C_{spawn}, C_{connect}\}$$

通道集枚举模型所识别的全部跨边界效应类别。边界治理的完备性主张以这一枚举为条件；遗漏任一通道将使完备性保证失效。

集合 C 是所有外部有效因果接口的抽象划分。具体部署必须细化该集合，以确保在*SC1*下的通道完备性。

膜决策函数：

$$M : \mathsf{Event} \times C \times \mathcal{S} \times \mathsf{Policy} \times \mathsf{Caps} \times \mathsf{Budget} \to D$$

膜是边界相关转移的唯一裁决界面。任何影响外部投影或权限界面的转移，均不得在此中介之外被允许。

用于钩子与任务核算的外部投影：

$$\begin{aligned} \mathcal{S} &: \text{institution state space}, \\ \mathcal{O}_{ext} &:= \mathsf{Inbox} \times \mathsf{Outbox} \times \mathsf{CommitLedger} \times \mathsf{WorldObs}, \\ \pi_{ext} &: \mathcal{S} \to \mathcal{O}_{ext}, \\ \pi_{ext}(S_t) &= (\mathsf{Inbox}_t, \mathsf{Outbox}_t, \\ &\qquad \mathsf{CommitLedger}_t, \mathsf{WorldObs}_t) =: o_t \end{aligned}$$

π_{ext} 定义了可观测边界投影，钩子生成、提交分类和续接核算均以此为依据。治理推理——

以该投影为锚定，而非以完整内部状态为依据，体现了可观测性受限的约束。此处，收件箱（Inbox）指外部接收、在边界可见的输入；发件箱（Outbox）指外部发出、在边界可见的输出；提交账本（CommitLedger）指用于记账的边界提交投影记录；世界观测（WorldObs）指边界层可获取的外部来源世界观测。这些是投影字段标签（槽位名称），而非完整状态变量。不同的完整状态可能具有相同的投影值 o_t；因此，治理记账有意采用基于投影的方式，而非力求涵盖所有隐藏状态。

投影提交谓词：

$$\mathrm{Commit}_{\pi}(act, s, s') := (s \xrightarrow{act} s') \wedge (\pi_{ext}(s') \neq \pi_{ext}(s))$$

本体论一阶谓词：

$$\mathrm{Commit}_{ext}(act, s, s') := (s \xrightarrow{act} s') \wedge (S_{ext}(s') \neq S_{ext}(s))$$

此处，$s \xrightarrow{act} s'$ 表示在动作实例 act下，系统从 s 转移至 s' 的一次有效状态转移。转换标签约定：每个建模的状态转移都带有动作标签（包括内部维护操作）；无标签的静默状态转移不属于此语义的范畴。

钩子和延续契约的提交记账基于投影，通过 Commit_{π}实现。一阶效应类型标注基于状态，通过 Commit_{ext}实现。既不改变任一谓词的内部计算不属于边界提交，但仍可能促成结构性扩展或策略层扩展，定义见第 2.4 节。

2.4 可达性与扩展谓词

边界治理不仅关注即时的外部提交，也关注未来可允许权限的增长。为分析这种增长，我们引入相对于声明的可受理性配置文件的可达语义。

声明的准入配置：

$$A_{adm} := (\Pi_{adm}, H, U_{policy})$$

其中：

- Π_{adm} 是可允许策略类，
- $H \in \mathbb{N}$ 是以转换步数计量的有限前瞻视界，
- U_{policy} 声明时域上的策略更新语义（通过有见证的治理更新实现固定或版本化）。

时域 H 在计算可达性时限定未来迹长度：只考虑长度不超过 H 次可容许转移的迹。因此，所有可达性和扩展声明都明确是时域相对的。

我们用 $A_{adm,\ t}$ 表示在步骤 t 生效的可允许性配置（即第 2.3 节引入的治理状态分量）。

时域受限可达性

时域受限可达性定义如下：

$$\mathrm{Reach}_H(s; A_{adm}) := \left\{ \tau \in C^* \;\middle|\; \begin{array}{l} |\tau| \le H, \\ \exists \pi \in \Pi_{adm} :\ \Pr(\tau \mid s, \pi, A_{adm}) > 0 \end{array} \right\}$$

在可容许策略 Π_{adm}下，从状态 s 出发，在时域 H 内可达的、对外部产生实际效力的通道轨迹集合。

仅用于风险核算时，我们在时域 H上使用步索引提升：由策略声明的映射 $\ell_t : C \to R_{class}$ 为边界信道分配风险类别标签；每个步骤事件编码为 $e_t := (c_t, \ell_t(c_t))$ ，表示边界信道事件 c_t 在步骤 t发生；若未发生边界事件，则编码为 $e_t := \emptyset$ 。此提升不会改变可达性的定义本身；它仅为下文定义的风险加权泛函提供带类型事件流。

可达集 $\mathsf{Reach}_H(s; A_{adm})$ 描述的并非制度已经做了什么，而是在有限的未来时域内能够做什么。它被用作可容许未来的语义包络，而不是声称在已部署系统中通常能够切实完成精确的包络计算。

建模立场：本文有意采用用于可达性的最小随机语义。我们不将马尔可夫性、平稳性或完整的 MDP 核规范视为构成性前提。此处唯一的要求是，每个可接受的策略都能在不超过时域 H 的有限边界轨迹上诱导出定义良好的概率律。在领域特定实例化中，可以加入更丰富的随机结构。

风险加权可达性的最小语义（记法）

定义规范边界迹字母表如下：

$$\Sigma_{ext} := C \times R_{class},$$

其中：

- C 是边界信道集；
- R_{class} 是由策略声明的有限风险类别集。

定义步事件域：

$$\Sigma_{step} := \Sigma_{ext} \cup \{\emptyset\}.$$

策略 $\pi \in \Pi_{adm}$ 从初始状态 s 出发，在时域 H 内的每条边界相关执行轨迹都会导出一个有限序列

$$e_{1:H} = (e_1, \dots, e_H), \quad e_t \in \Sigma_{step}.$$

如果在步骤 t 没有发生边界相关事件，则定义 $e_t := \emptyset$。

定义策略声明的步风险权重函数:

$$w_{step} : \Sigma_{step} \to \mathbb{R}_{\geq 0},$$

关于声明的风险类别单调递增，并满足 $w_{step}(\emptyset) = 0$。

期望 $\mathbb{E}_\pi[\cdot]$ 取自策略 π 执行所引发的随机性，以及任何建模的外生过程或环境过程。

风险加权可达性

理想风险加权可达泛函定义如下:

$$\mu_H^{risk}(s; A_{adm}) = \sup_{\pi \in \Pi_{adm}} \mathbb{E}_\pi \left[\sum_{t=1}^{H} w_{step}(e_t) \right].$$

这衡量了时域 H 内可容许策略的最大期望边界加权风险。我们假设由此诱导的、以策略为索引的轨迹分布族具有这样的性质：该上确界在 $\mathbb{R}_{\geq 0} \cup \{+\infty\}$ 中有定义；但不要求由优化器取得该上确界。

由于对 Π_{adm} 取上确界可能难以处理，我们引入一个有界估计量:

$$\hat{\mu}_H^{risk}(s; A_{adm}, L) = \sup_{\pi \in \Pi_{adm}^{(L)}} \mathbb{E}_\pi \left[\sum_{t=1}^{H} w_{step}(e_t) \right],$$

其中， $\Pi_{adm}^{(L)}$ 限制策略复杂度。

操作性近似概况是以下数对 (L, δ_μ):

- L 通过 $\Pi_{adm}^{(L)}$ 对策略类别复杂度进行界定；
- $\delta_\mu \geq 0$ 是在增加 L时用于校准的估计量稳定性容差。

可处理结构替代模型（非规范性）

运行诊断时，可以使用能力图代理。

Let:

$$G_{cap}(s) = (V_{cap}, E_{cap})$$

是一个类型化权限图，其中：

- V_{cap} 是承载权限的顶点集合（运行时、工具或外部主体）；
- $E_{cap} \subseteq V_{cap} \times V_{cap}$ 是有向委派关系或通信关系。

每个顶点 $v \in V_{cap}$ 都按以下内容进行类型标注：

- 一个风险类别 $r_v \in R_{class}$,
- 一个信任衰减因子 $\chi_v \in [0,1]$,
- 一个归一化有效性参数 $\nu_v \in [0,1]$。

在诊断图范围 H_{cap}内定义能力可达范围：

$$\mathrm{Reach}^{cap}_{H_{cap}}(s) \subseteq V_{cap},$$

从当前权限配置出发，经过至多 H_{cap} 次委派或传播步骤可到达的顶点集合。

接着定义代理扩展度量：

$$\mu^{proxy}_{H_{cap}}(s) = \sum_{v \in \mathrm{Reach}^{cap}_{H_{cap}}(s)} w_{class}(r_v)\, \chi_v\, \nu_v,$$

其中：

- $w_{class} : R_{class} \to \mathbb{R}_{\geq 0}$ 是策略声明的类别风险权重函数（与 w_{step} 及通道到风险标签的映射保持一致地选取）；
- χ_v 会降低对间接顶点或弱受控顶点的信任；
- ν_v 体现剩余有效性（例如，生存时间归一化或撤销时域）。

此替代指标属于非规范性内容，灵感来自能力安全权限核算 [16]。它旨在用于运行诊断（例如，图遍历、权限差分检查或策略影响比较），且不提供任何不变性或完备性保证。

效应标签与扩展类型化

区分两种二阶子类型：

- 第二_T：结构扩展，

- 第二_P：策略扩展。

结构扩展谓词：

$$\begin{aligned}\text{StructuralExpand}(s \to s') := (caps(s') \neq caps(s)) \\ \vee (tools(s') \neq tools(s)) \\ \vee (V(s') \neq V(s)) \\ \vee (E(s') \neq E(s))\end{aligned}$$

StructuralExpand 有意采用过近似：它可能将与权限无关的重新配置归类为 SECOND_T。这种保守偏差在构成层是有意为之；精度过滤属于运行诊断的范畴。

我们使用效应标签，而非互斥划分：标签符号 FIRST、SECOND_T 和 SECOND_P 是固定的标签常量（有限标签字母表中的元素），而非状态变量或数值索引。

$$\text{Tag}_{\mathsf{FIRST}}(act, s, s') := \text{Commit}_{ext}(act, s, s')$$

$$\text{Tag}_{\mathsf{SECOND_T}}(s \to s') := \text{StructuralExpand}(s \to s')$$

$$\text{Tag}_{\mathsf{SECOND_P}}(s \to s') := \text{PolicyExpand}_{H}(s \to s')$$

$$\begin{aligned}\text{TagSet}(act, s, s') := \{\mathsf{FIRST} \mid \text{Tag}_{\mathsf{FIRST}}(act, s, s')\} \\ \cup \{\mathsf{SECOND_T} \mid \text{Tag}_{\mathsf{SECOND_T}}(s \to s')\} \\ \cup \{\mathsf{SECOND_P} \mid \text{Tag}_{\mathsf{SECOND_P}}(s \to s')\}\end{aligned}$$

组合变迁是允许的：单个转移可以携带多个标签（例如 {FIRST、SECOND_T}）。

这体现了拓扑和能力的增长。

策略扩展通过仅策略核心状态差分和严格的时域限制可达性包含来定义：

$$\begin{aligned}\text{CoreEq}(s, s') := S_{int}(s') = S_{int}(s) \wedge S_{ext}(s') = S_{ext}(s) \\ \wedge S_{budget}(s') = S_{budget}(s) \wedge S_{topo}(s') = S_{topo}(s)\end{aligned}$$

CoreEq 有意排除 S_{ledger} 相等性：允许由策略变迁裁决/锚定引起的仅追加见证增长。

$$\begin{aligned}\text{PolicyExpand}_H(s_t \to s_{t+1}) := \text{CoreEq}(s_t, s_{t+1}) \\ \wedge \text{Reach}_H(s_{t+1}; A_{adm,t+1}) \\ \supsetneq \text{Reach}_H(s_t; A_{adm,t})\end{aligned}$$

为便于阅读而省略索引时，我们将 $\text{PolicyExpand}_H(s \to s')$ 理解为采用转换局部时间索引。

可比性约定：在规范迹空间上评估 PolicyExpand_H 中的严格包含。如果策略版本导致不同的内部表示，则必须通过固定的策略声明编码将迹规范化为规范字母表，再进行集合比较。

PolicyExpand_H 比较的其他条件不变约定：除非明确声明为 A_{adm} 的一部分，否则在步骤 t 和 $t+1$ 之间保持外生／环境随机性假设不变。

这刻画了固定核心状态下，策略版本转换所引发的可采纳性表面扩展。它不包括仅由结构或能力突变导致的可达性增长。

SECOND_P 是相对于配置而言的：其分类是相对于当前的可允许性配置（$A_{adm,\ t}$）和声明的近似配置来定义的。因此，它并非原始系统动力学中不依赖本体的属性。换言之，SECOND_P 并非状态内在属性，而是相对于治理配置的转移属性。

SECOND_P 并非语义重新标记。它要求有见证表明发生了策略版本转换，该转换改变了可允许策略类（$\Pi_{adm,\ t} \to \Pi_{adm,\ t+1}$），进而在声明配置下改变了未来的可达包络（$\text{Reach}_H$）。

从这个意义上说，可采纳性表面构成治理状态的一部分：对其进行修改会改变哪些外部未来变得可执行。

在操作层面：

- SECOND_T 表示对权限界面的结构修改（例如，能力分配、工具配置或运行时拓扑的变更），可解释为控制平面权限突变；
- SECOND_P 表示在版本化语义下，由策略引发的对可受理策略类别的修改。

扩展可达未来的策略版本转换归类为 SECOND_P 事件，并受 *P-1* 见证纪律约束。

归一化扩张诊断

为区分结构增长与数值噪声：

$$\Delta_{expand}(s \to s'; A_{ref}, L) = \frac{\hat{\mu}_H^{risk}(s'; A_{ref}, L) - \hat{\mu}_H^{risk}(s; A_{ref}, L)}{\max\big(1,\, \hat{\mu}_H^{risk}(s; A_{ref}, L)\big)}$$

其中， A_{ref} 是用于比较的固定参考可采纳性配置（通常为$A_{adm,\ t}$）。此诊断在同一配置下比较两种状态，具有可操作性。它不能取代 PolicyExpand$_H$，后者的 SECOND_P 分类可能会比较 $A_{adm,\ t}$ 和 $A_{adm,\ t+1}$。

边界效应标签触发条件：

$$\begin{aligned}&(s \xrightarrow{act} s') \wedge \Big(\mathrm{Tag}_{\mathsf{SECOND_T}}(s \to s') \vee \mathrm{Tag}_{\mathsf{SECOND_P}}(s \to s')\Big)\\ &\Rightarrow \mathrm{TagSet}(act, s, s') \cap \{\mathsf{SECOND_T}, \mathsf{SECOND_P}\} \neq \emptyset\end{aligned}$$

操作阈值标记（相对于配置文件）：

$$\Delta_{expand}(s \to s'; A_{ref}, L) > \epsilon_{expand_norm}$$

其中， $\epsilon_{expand_norm} > 0$ 是当前近似配置文件中声明的策略阈值。

这种分类将权限表面增长提升为与边界相关的转变类别，即便没有立即发生外部状态变化。
该构造将序贯决策理论中的可达性和权力寻求分析应用于策略约束的外部轨迹 [6], [17]。

2.5 见证、原子性与重放原语

外部效应谓词：

$$\begin{aligned}\mathrm{ExternalEffect}(act, s, s') := {}& \mathrm{Commit}_{ext}(act, s, s')\\ &\vee \mathrm{StructuralExpand}(s \to s')\\ &\vee \mathrm{PolicyExpand}_H(s \to s')\end{aligned}$$

外部效应是关于转移 (s, act, s') 的语义谓词，而不是观察性标签。它用于分类那些直接修改外部生效状态或扩大未来可容许权限面的转移。等价地，外部效应$(act, s, s') \Leftrightarrow$ 标签集$(act, s, s') \neq \emptyset$。

如第 2.3 节所定，对钩子/延续的提交记账是基于投影的，通过 Commit$_\pi$ 实现；而一阶效应类型标注（因此也包括 FIRST）是基于状态的，通过 Commit$_{ext}$ 实现。当审计性要求投影忠实性时，我们假定 Commit$_{ext}(act, s, s') \Rightarrow$ Commit$_\pi(act, s, s')$。

膜中介谓词：

令 $\mathcal{T}$ 表示系统转换关系，令 $\mathcal{T}_{mem} \subseteq \mathcal{T}$ 表示膜介导转移。

$$\textsf{ThroughMembrane}(act, s, s') := \exists d \in D,\ w \in \mathcal{W} :$$
$$(s \xrightarrow{act} s') \in \mathcal{T}_{mem}$$
$$\wedge\, Decide(d, act) \wedge WitnessBind(w, d, act)$$

- $Decide(d, act)$ 表示决策函数 M 针对给定行为实例返回的膜决策。
- $WitnessBind(w, d, act)$ 表示见证记录 w 将决策类d 绑定到具体行为实例 act。
- $\mathcal{T}_{mem}$ 体现了拓扑调解：只有通过膜转移表面发出的转移，才是经膜调解的。

原子关系：

$$Atomic(d, w, act, s, s')$$

当且仅当裁决、见证锚定和外部有效转移提交作为单个不可分割转换，在系统转换关系下发生时，该条件才成立。

- 效应提交$(act, s, s') :=$ 外部效应(act, s, s')。
- $Anchor(w, d, act)$ 表示将见证绑定仅追加式锚定到见证账本中。本节将 $Anchor(w, d, act)$ 视为转移事件（步骤级追加操作），而不只是状态谓词。
- 原子性是状态转移关系 $\mathcal{T}$ 上的执行前缀顺序约束，不同于视界受限边界轨迹可达性 $\textsf{Reach}_H(\cdot; A_{adm})$。

形式上，原子性要求：

1. 三元组 $(Decide(d, act), Anchor(w, d, act),$ 效应提交$(act, s, s'))$ 在状态转移关系中体现为单次转移；并且 2. 在 $\mathcal{T}$ 中，不存在任何有限执行前缀，使得满足效应提交$(act, \cdot, \cdot)$ 的步骤先于任何满足 $Anchor(w, d, act)$的步骤发生。

等价地，任何执行前缀都不得暴露针对 act 的外部有效转移，同时其对应的见证锚点尚不存在。
因此，原子性排除了效应执行可能发生在见证锚定之前的分阶段模式。

重放函数：

$$\text{Replay}(policy_version_t, ctx_abs_t, witness_t) = d_t \in D$$

其中， $witness_t \in \mathcal{W}$ $policy_version_t$ 表示$A_{adm,\ t}$的策略版本组件， d_t 是绑定在 $witness_t$中的裁决类别。
在策略版本和声明的上下文抽象完全相同的情况下，重放可重构裁决类别。在这些输入条件下，它保证裁定类别确定性，但不声称能够重构完整的隐藏状态轨迹或生成式内部机制。

3. 系统模型

3.1 系统类别

以下分类法基于第 2 节引入的转移语义定义。

如果一个系统在状态 S_t 上的转移关系满足以下所有条件，则它是一个**开放控制论过程 (OCP)**：

1. **状态累积**：存在持续的步间内部状态演化（$S_{int,\ t+1}$ 取决于 $S_{int,\ t}$）。2. **环境耦合反馈**：源自 S_{ext} 的观测会影响后续内部转移。3. **外部因果作用潜力**：存在能够产生外部有效转移的可接受行动。4. **迭代自主性**：系统无需人工逐步触发即可执行多个转移步骤。

此类别包括许多非 AI 系统（控制器、机器人程序、调度器）。其定义基于结构，与生成模型无关。

一个 **AI 空间核心** 是具备以下特征的 OCP：

- 能够生成开放式产物的生成式符号基底；
- 嵌入在 S_{int}中的明确产物状态；
- 治理影响 S_{ext} 或权限界面之转移的策略调控边界。

一个 **自主 AI 空间** 是至少具备一个自主性增强器的 AI 空间核心：

- **目标自主性**：系统可以依据明确且受策略约束的更新规则，更新其可对外核算的目标组合；或者
- **扩展能力**：可接受行动可以增加拓扑（S_{topo}）、能力范围或可接受策略范围。

这种分层分类通过区分通用自动化与能够重塑其自身未来权威边界的制度，避免了夸大其词。

3.2 转移类别与任务层

第 2.4 节所定义的边界相关转移，按效应标签进行类型划分：

- FIRST 标签：标签$_{\text{FIRST}}(act, s, s')$;
- SECOND—T 标签：标签$_{\text{SECOND T}}(s \to s')$;
- ¯SECOND—P 标签：标签$_{\text{SECOND P}}$ $(s \to s')$。 _

标签并非互斥：单个转移可能带有多个标签。

任务过程分为两个层次：

- $T_{int,\ t}$：内部规划、分解、调度和制品构建；
- $T_{ext,\ t}$：用于延续契约和治理核算的外部可问责任务分配。

这种区分确保内部认知演化不会自动意味着外部可问责的制度意图。

外部问责的钩子来源类：

- 外生：由外部事件过程产生；
- 受激：由明确授权的边界自我刺激引发；
- 内生：仅由内部动力学产生。

在治理关键场景中，只有符合已声明策略的来源类别才能填入 $T_{ext,\ t}$ 。

3.3 可观测性机制

边界观测具有有限的速率和有限的字母表，监督能力也有限。在开放增长制度下，运行时总体（N_t）、连接性（G_t）和事件复杂度的增加，会与有限的观测带宽相竞争。

因此，我们不将完整的隐藏状态可辨识性视为治理目标。相反，可执行治理必须锚定于针对转移类别的边界不变量和见证纪律，而非对内部生成轨迹的重构 [18], [19]。

这一约束促成了第 4 节中形式化阐述的边界优先立场，该立场体现在构成性法则中。

3.4 最小示例场景（SECOND_—T 不含 FIRST）

考虑以下初始状态 s_t ：

- $CapabilitySet$ 中没有支付连接器，
- 有效凭证中没有部署令牌，
- 未更改的外部投影基线 $\pi_{ext}(S_t)$。

在步骤 t，Unit 执行操作 a_{cfg}：

- 安装了一个特权支付连接器，
- 绑定了一个部署范围凭证，
- 尚未执行任何对外调用。

然后：

- $\pi_{ext}(S_{t+1}) = \pi_{ext}(S_t)$ 因此未观察到任何投影提交；
- $S_{ext}(s_{t+1}) = S_{ext}(s_t)$ 因此未发生 FIRST 转换；
- $tools(s_{t+1}) \neq tools(s_t)$ 且 $caps(s_{t+1}) \neq caps(s_t)$，因此 StructuralExpand$(s_t \to s_{t+1})$ 成立；
- 因此 TagSet$(a_{\text{配置}}\ s_t,\ s_{t+1})$ 包含 SECOND_T，而不包含 FIRST。

治理后果：

- 若没有二阶类型标注，这种转移通常会被视为内部配置变动，并可能看起来符合规范；
- 若采用 SECOND_T 类型标注，它就是一个强制性的膜事件，必须先经过裁决并取得见证，才能扩展潜在权限。

这正是该框架所要解决的核心实践不对称性。

4. 构成律

以下定律定义了边界治理具有意义所需的本构条件。这些是模型级不变量，而非操作建议。

4.1 定律*P-1*（边界见证）

如果某个转移具有边界效应（即满足 ExternalEffect(act, s, s')），则必须包含：

- 一项明确的膜决策，
- 一项确定性裁定引用，
- 一份绑定于该裁决的见证记录。

等价形式：

$$\begin{aligned}&\text{ExternalEffect}(act, s, s')\\ &\Rightarrow \exists d \in D,\ w \in \mathcal{W}:\\ &\quad Decide(d, act) \wedge WitnessBind(w, d, act)\end{aligned}$$

非形式地说：

没有膜见证，就没有边界效应。

这条定律将见证生成从可选的审计实践提升为构成性的有效性条件。缺少相应裁定见证绑定的边界效应在模型中无效，而不只是未记录的 [1], [15]。

4.2 法则 P-1a（不可绕过）

$$\begin{aligned}&\text{ExternalEffect}(act, s, s')\\&\Rightarrow \text{ThroughMembrane}(act, s, s')\end{aligned}$$

外部生效转换必须经由膜决策函数中介。任何允许在膜裁决之外执行效应的架构路径，都构成对模型完备性假设的违反，而非可恢复的实现缺陷。

这条定律在转换层级上形式化了完全中介。

4.3 法则 P-1b（裁决到效果的原子性）

$$\begin{aligned}&\text{ExternalEffect}(act, s, s')\\&\Rightarrow \exists d \in D,\ w \in \mathcal{W}:\\&\quad Decide(d, act) \land WitnessBind(w, d, act) \land\\&\quad Atomic(d, w, act, s, s')\end{aligned}$$

这条定律通过加入时间纪律强化了*P-1*：见证锚定必须与外部生效转换的提交在执行上保持原子性。
决策、见证锚定和外部生效转换的提交必须在系统状态转换关系下，作为单一不可分割的转换发生。

分阶段执行模式——即外部效应可能先于见证锚定发生的模式——不属于可接受语义。此类模式会将见证规约化为事后日志记录，并违反原子性。从概念上说，这是一项顺序约束主张：外部可见效果转换不得先于持久裁决证据。

4.4 法则 P-1c（可重放裁决）

法则 *P-1c* 要求，对于时间 t发生的任何有见证边界转换，在相同策略版本和声明的上下文抽象下进行重放，都能重构最初作出的裁决类别。这里，
$policy_version_t$ 是$A_{adm,t}$中的策略版本部分。

等价形式：

$$\begin{aligned}&\forall t,\ \forall act_t, s_t, s_{t+1}, d_t, w_t:\\&\quad (w_t \in \mathcal{W}) \land WitnessBind(w_t, d_t, act_t)\\&\quad \Rightarrow \text{Replay}(policy_version_t, ctx_abs_t, w_t) = d_t\end{aligned}$$

此保证的适用范围：

- 它确保在固定重放输入下实现决策类别确定性；
- 它并不意味着重构隐藏生成轨迹或内部随机过程。

因此，见证记录在裁决层级上具有语义重构作用，而不仅仅是档案材料。

5. 派生陈述

以下结果是基于第 2–4 节中引入的转移语义和本构定律，在明确假设下推导得出的。

认识论状态图例：

- **命题**：在明确假设下推导出的陈述。
- **设计推论**：操作性设计关系；并非系统的形式不变量。

5.1 命题 A（外生钩子下的外部任务因果）

令 Exo_t 表示一个外生输入过程。令 h 为一个钩子生成函数。定义 $HookEvent_t := h(\pi_{ext}(S_t), Exo_t)$。

假设：

1. 钩子生成完全是 $(\pi_{ext}(S_t), Exo_t)$ 的函数；2. 对未发生变化的输入，钩子生成处于静默状态：如果 $\pi_{ext}(S_{t+1}) = \pi_{ext}(S_t)$ 且 $Exo_{t+1} = Exo_t$，则 $HookEvent_{t+1} = \emptyset$；3. 在所分析的时间区间内，不发生外生实现（即对于该区间中的所有步骤 i， $Exo_i = Exo_0$ ）；4. 任何向 T_{ext} 中插入元素的机制，都必须由 $HookEvent_t$ 或明确授权的 STIMULATED 边界动作以因果方式触发；5. 在所分析的时间区间内，纯内生插入 T_{ext} 被策略禁止；6. 任何明确授权的 STIMULATED 边界动作都满足 $\mathrm{Commit}_\pi(a, s, s')$。

那么，任何满足

$$\forall i \in \{0, \dots, n-1\} : \neg\, \mathrm{Commit}_\pi(a_i, s_i, s_{i+1})$$

都不能增加 T_{ext}。

等价地：

> T_{ext} 中任何未伴随外生实现的增加，都意味着至少存在一个影响边界的动作改变了 π_{ext}。

证明概要

（Commit$_\pi$ 的定义）在此前提下给出了所有 i 的 $\pi_{ext}(s_{i+1}) = \pi_{ext}(s_i)$。假设（3）使 Exo 在该区间内保持不变。于是，假设（2）推出不存在钩子事件。假设（4）和（5）排除了内生插入路径，而假设（6）则封闭了受激路径，因为任何此类插入都需要 Commit$_\pi$，这与之矛盾。因此，T_{ext} 不可能在没有边界中介的情况下增长。

解读

这排除了在所述治理约束下纯粹内生生成外部可问责任务的可能性，从而应对捷径式奖励篡改风险 [20], [21]。

范围：本命题适用于内生来源对 T_{ext} 不满足续接可接受条件的制度。

5.2 命题 B（有限观察者不可辨识性）

令 Σ_b 表示边界可观测事件的有限字母表。令 κ 表示每步发出的边界符号的最大数量。令 supp(S_t) 表示时间 t 的可达隐状态支撑集。

假设：

1. 有限的边界字母表 Σ_b，发射速率 κ 有上界；2. 观察者带宽 $R \leq \kappa \log_2 |\Sigma_b|$；3. 对某个 $\delta > 0$，当 t 足够大时，隐藏支持增长 $|\text{supp}(S_t)| \geq 2^{(R+\delta)t}$。

例如，在迭代式工具增强生成和分支分解过程中，隐藏的人工制品/过程状态呈组合式增长的情况下，假设 (3) 成立。

因此，当 t 足够大时，存在不同的隐藏状态，它们会产生完全相同的观测边界轨迹。

证明概要

截至时间 t 可区分的边界轨迹数量至多为 2^{Rt}。根据假设 (3)，隐藏支撑超过此上界。根据鸽巢原理，必然会发生碰撞。

解释

在满足（3）的增长制度中，基于边界轨迹对完整隐藏状态进行重建的治理，在结构上是不充分确定的。可执行边界不变量，而非内部状态推断，成为鲁棒控制目标 [18], [19]。

范围说明：这是一项依赖机制的主张。如果隐藏支持增长受到限制，或监督能力以相当的速度扩展，不可识别性压力可能会减弱。

5.3 命题 C（边界可见进展下的稀缺）

令 λ_{ext} 表示外生边界事件率（边界可见任务证据或外部触发因素的速率）。令 μ 表示内部活动率。令持续稀缺机制由 $\mu \gg \lambda_{ext}$定义。

假设：

1. 治理相关进展仅通过边界可见证据予以确认（即经由 π_{ext} 可观测转变）；
2. 边界可见进展要求外生边界事件或经明确授权的 STIMULATED 边界行为；
3. 延续合约具有吸收式停止语义：若在足够长的时域内未发生边界可见进展，则系统将发生系统状态转换并进入停止状态，除非被明确重新激活；4. 在所分析区间内策略保持固定；5. 仅靠外生边界事件流不足以在所分析时域内满足延续合约。

定义边界核算指标：

$$\mathrm{ext_delta}_t \in \{0, 1\}$$

其中：

$$\mathrm{ext_delta}_t = 1 \Leftrightarrow \left(\mathrm{Commit}_\pi(a_t, s_t, s_{t+1}) \vee (\mathrm{Stimulated}_t = 1)\right)$$

此处 $\mathrm{Stimulated}_t = 1$ 表示步骤 t 上一项经明确授权的 STIMULATED 边界行动。

定义累积边界计数：

$$\mathrm{ext_state}_t = \sum_{i \le t} \mathbf{1}\{\mathrm{ext_delta}_i = 1\}$$

陈述（稀缺条件下的边界可见延续）

在持续稀缺（$\mu \gg \lambda_{ext}$）及假设 (1)-(5) 下，在所分析区间上维持非停机行为，要求该区间上至少存在一项经明确授权的 STIMULATED 边界行动。

形式上，对于任意区间 $I = \{t_0, \ldots, t_1 - 1\}$：

令 $\mathrm{NonHalt}(I)$ 表示系统在区间 I内未进入吸收式停止状态。

$$\text{非停机}(I) \Rightarrow \exists t \in I : \text{受激}_t = 1.$$

诊断性推论（假设违反测试）

如果在 I 期间观察到持续不停机，且对所有 $t \in I$均处于受激状态$_t = 0$ ，那么假设 (1)-(5) 中至少有一项在该区间内不成立。典型失效模式包括边界核算不一致、放宽停止语义，或事实上外生流量足以维持运行的情形。

解释

在满足假设 (1)-(5) 的稀缺状态中，纯粹的内部活动无法无限期地维持运行。持续的运行时需要显式的、由边界介导的延续事件；否则，根据假设 (3)，将随之停机。这一结果是结构性的：它源于内部动力学与边界可见状态之间的分离，而非任何特定的奖励模型。

地位：在显式延续契约与边界可观测性假设下的命题，外加针对假设违反的诊断性推论。并非关于任意经济或奖励结构的普遍定理。

5.4 设计推论 H-1（监督能力与扩张压力）

令 $ObsCap(t)$ 表示边界审查能力，令 $X_{risk}(t)$ 表示风险加权的扩张压力。在操作化实例中，根据声明的近似配置文件，一种可接受的选择是 $X_{risk}(t) := \hat{\mu}_H^{risk}(s_t; A_{adm,t}, L)$ 。只要其他由策略声明的实例化方式保留对权限表面扩张的单调敏感性，也同样允许采用。

定义前向差分：

$$\Delta X_{risk}(t) := X_{risk}(t+1) - X_{risk}(t),$$
$$\Delta ObsCap(t) := ObsCap(t+1) - ObsCap(t).$$

令 $k > 0$ 为在所选的 X_{risk} 和 $ObsCap$归一化方式下，经政策校准的比例因子。它充当单位换算因子，将扩张压力增量换算为等效的审查工作需求。

操作性设计关系：

$$\Delta X_{risk}(t) \leq k\, \Delta ObsCap(t),$$

解释

- 在实际运行中，这是审查流程的稳定性条件：扩张压力所引发的增量审查需求到达量不得超过审查服务能力；

- 如果连续足够多步都出现 $\Delta X_{risk}(t) > k\,\Delta ObsCap(t)$ 这种情况，审查积压就会累积，见证审查／异常检测也会滞后；
- 因此，稳定性需要通过限制扩张动态或增加监督资源来维持。

附录 D 提供了这一解释的一种非规范性队列／积压实现。

状态：这是治理设计的操作性推论，而非构成性不变量。

6. 启示

经典治理往往将扩展视为偶然的实施偏移。相比之下，AI Space Physics 将扩展视为一种明确且与边界相关的转变类别，带有二阶效应标签（SECOND_T / SECOND_P），必须经过裁定并由见证方确认。这样便弥补了一个核心盲点：潜在权限增长可能绕过治理关注，直到为时已晚。

该框架与能力安全相邻，但适用范围并不相同：

- 能力安全关注的是当前由谁持有哪些权限；
- AI Space Physics探讨的是，在时域受限可达性和版本化策略语义下，未来权限规模如何发生可允许的变化。

前者是对权限持有情况的快照；后者则是关于制度发展的转移语义。

1. **边界优先治理**。 在开放增长制度下，完整的内部可解释性并非可靠的控制目标；可执行边界不变量才是稳健的目标。 2. **将扩张治理视为权限核算**。 能力、连通性和拓扑累积必须作为带有明确效应标记的状态转移进行治理，而不能被视为背景工程变更 [1], [2], [6]。 3. **见证完整性是安全关键的基础**。 没有防篡改证据的覆盖并不构成治理；事件后问责需要采用带有密码学承诺的追加式、不可抵赖记录（例如哈希链或默克尔根锚定） [22], [23], [24]。 4.**将任务来源作为防走捷径的控制手段**。 延续关键型T_{ext} 要求明确记录外生|受激|内生来源，以防止悄无声息的信号伪造循环[20], [21]。 5.**验证目标转变**。 形式化方法应优先关注不可绕过性、原子性和见证完整性，而不是声称能够完整重建内部状态 [3], [4]。

7. 局限性与范围条件

7.1 构成性主张与经验性主张

本文中的定律族具有构成性：它界定了治理语义要具有意义必须满足的条件。这并非一项经验性主张，即现实部署默认就满足这些条件。因此，通过模型检验或策略测试应被视为在特定假设下获得的证据，而非无条件的安全认证。

7.2 范围条件（SC1-SC6）

SC1 通道完备性： 模型化的通道集必须覆盖一阶提交或二阶标记效应可能发生的所有路径。未建模的侧信道会使见证完备性主张失效。

SC2 完全中介/不可绕过性： 所有边界相关转移都必须在结构上被强制通过膜裁决。如果存在绕过路径，则*P-1a* 不作为系统属性成立 [1]。

SC3 有效内部控制： 被视为内部的组件必须满足策略控制、停止/撤销控制和见证覆盖。边界耦合依赖会削弱更强的可控性主张。

SC4 见证完整性与完整性保障： 见证记录必须完整，并能防止删除、回滚或不一致呈现造成的篡改。实际操作中，这要求对见证历史进行密码学仅追加承诺（例如采用哈希链或默克尔根锚定），而非使用可变的应用日志。在更强的部署中，$Anchor(w, d, act)$ 应表示可验证的追加语义（例如支持包含性证明/一致性证明的日志记录），而非尽力而为写入事件。对于跨组织来源，SCITT 风格的透明度服务是一种兼容的实现模式 [24]。否则，证据就会退化为事后日志记录 [22], [23]。

SC5 声明可达近似配置： 二阶分类取决于声明的 $(A_{adm}, L, \delta_\mu, \epsilon_{expand_norm})$ 配置。特别是，SECOND_P 相对于该配置而言是制度相对的，而非绝对本体层级的标签。为避免提出表示依赖扩展的主张，必须在规范迹空间（或通过固定的策略声明规范化映射到该空间）上评估 PolicyExpand$_H$ 中的比较。

SC6 外部记账的投影忠实性： 当通过投影日志对一阶效应进行操作性审计时，部署必须满足提交$_{ext}(act, s, s') \Rightarrow$ 提交$_\pi(act, s, s')$。否则，一阶记账可能会低估本体论外部效应。

7.3 计算与观测极限

μ_H^{risk}是一个理想语义对象，可能难以精确计算；操作性治理依赖有界近似和阈值校

准。在接近阈值边界时，预期会出现假阳性和假阴性，必须通过程序加以管理。

第 2.4 节中的可达语义有意保持轻量：它避免在规范核心中承诺完整的马尔可夫/马尔可夫决策过程假设，优先考虑在工程治理场景中的可移植性。如果部署场景需要更强的随机结构，则应明确将其作为实现侧的细化层添加。

有限的观察者带宽也限制了从边界轨迹中能够推断出的信息，尤其是在隐藏支撑和拓扑的扩展速度快于观察能力时。因此，框架优先采用可执行的边界转换规约，而非声称能够完全辨识内部状态 [18], [19]。

7.4 形式化制品的局限

TLA+/SPIN 及相关制品可以验证蕴含关系，并为有限抽象揭示反例，但它们不能取代理论，也无法证明生产系统的端到端性质。经过验证的性质仍以抽象边界和模型假设为条件 [3], [4]。

7.5 前瞻性研究方向：多核人工智能空间动力学

本文有意以边界为中心：它并未声称能够为异构基础模型运行时（例如 GPT、Claude 或 Gemini 级系统）在共享策略绑定下生成、委派任务并进行递归协调时的内部动力学，给出完整的预测性组合定律。

单独构建理论时可能纳入研究范围的现象包括：

- 异构模型／运行时堆栈之间的非交换组合效应；
- 并发运行的规划器—执行器循环之间的战略性干扰；
- 由跨运行时委派和能力放大驱动的联盟层级 SECOND_T 级联效应。

这些问题需要独立的形式化层以及受控的实验方案。在本研究中，这些问题不属于构成性保证的研究范围；构成性保证仍以可强制执行的边界不变量为锚定基础。

8. 结论

AI Space Physics 为开放式、自我扩张的人工智能机构引入了转型层级治理语义。其核心贡献是构成性的，而非程序性的：权限表面扩张被重新归类为一种与边界相关的转移类别，并明确标注效应标签，即使不存在即时的外部提交。结构增长（SECOND_T）和可准入性表面拓宽（SECOND_P）因此并非背景

工程漂移，而是需要裁决、原子锚定和可重放见证语义的事件。

在此框架下，只有当每个一阶和二阶边界效应都经由不可绕过的膜加以中介，并通过原子化的 Decide $\rightarrow$ 锚定 $\rightarrow$ 效应语义验证时，治理才具有可执行性。强可治理性被定义为系统的一种转移属性，而非政策愿景或监控实践。允许在这一规约之外发生具有外部效力的转移的架构或许仍能正常运行，但不具备此处所定义的强意义上的可治理性。

该框架并未消除风险，也不保证对齐；它明确了机构权威产生外部后果并可接受审计时所依赖的最小因果接口。在开放式增长和观察者有限的情境下，鲁棒控制目标是边界不变量，而非对内部的全面可解释性。因此，这项工作的贡献并非提出一份更严格的合规检查清单，而是从根本上阐明哪些转移首先与治理相关。

在操作层面，下一步是将这些语义落实为持续执行的机构契约：明确的通道完备性声明、机器可验证的不可绕过性与原子性测试、可重放级见证基础设施，以及校准后的扩展诊断。

随着自主制度不断扩大，治理质量将取决于监督能力能否随着权限表面扩张而同步提升。如果扩张压力持续超过审查带宽，见证纪律就会退化为积压，而非控制。

政策表面扩展如何处理同样至关重要。如果可受理性扩张被视为常规配置漂移，而非受治理的相对于配置档案的二阶转移（SECOND_P），权限表面增长就会在缺乏审慎裁决的情况下不断累积。

强可治理性并不意味着安全性。它界定了安全性主张在制度层面能够接受检验、质疑并随时间追责的语义前提。

附录 A：符号表

符号	含义
Cell、Unit、膜	核心模型的类符号（类型）
	对象
$\mathcal{C}$	Cell 实例的承载集
$\mathcal{U}$	Unit 实例的承载集
$\mathcal{W}$	见证记录的承载集

符号	含义
Cell 实例的结构模式	Cell 实例的结构模式
Unit 实例的结构模式	Unit 实例的结构模式
$\mathcal{S}$	制度状态空间
S_t	t 时的制度状态
$S_{ext,t}$	本体论外部有效状态
S	组件
$C^{topo,t}$	权威拓扑状态（N_t, G_t）
M	边界信道集
事件、策略、上限、预算	膜决策函数
D	膜的参数类型
A_{adm}	裁决签名
$A_{adm,t}$	决策类别
A_{ref}	允许、拒绝、隔离
U_{policy}	声明的准入配置（Π_{adm}, H, U_{policy}）
可达性$_H(s; A_{adm})$	步骤 t 时启用的可允许性配置
$\ell_t : C \to R_{class}$	固定参考可采纳性配置
w_{step}	用于同配置诊断
μ_H^{risk}	比较
$\hat{\mu}_H^{risk}(s; A_{adm}\ L)$	时域内的策略更新语义
$G_{cap}(s)$	（固定或版本化）
可达性$^{cap}_{H_{cap}}(s)$	时域受限的外部有效
w_{class}	在声明的可受理性
$\mu^{proxy}_{H_{cap}}$	个人资料
第二 T_	策略声明的通道到风险
	用于步骤事件类型标注的映射
	针对步骤事件的风险权重函数
	$e_t \in \Sigma_{step}$
	理想风险加权可达泛函
	风险加权可达性的有界估计量
	在声明配置下
	状态下的类型化能力/权限图
	状态 s
	范围内的能力可达节点
	诊断图时域 H_{cap}
	风险类别上的风险权重函数
	在能力图代理中
	非规范性可处理可达性
	替代指标（基于能力图）
	结构性二阶子类型
	（结构性权威表面
	扩展）

符号	含义
SECOND P_	策略二阶子类型（在……下的可接受性表面扩展策略版本控制）
标签FIRST，标签SECOND_T，标签SECOND_P	一阶提交的效应标签，结构扩展和策略扩展
标签集(act, s, s')	由一个所携带的效应标签集合
CoreEq(s, s')	转移；标签可以同时出现……的相等性
$\Delta_{expand}(s \to s'; A_{ref}\ L)$	内部/外部/预算/拓扑仅策略增量下的组件（允许仅追加账本增长）归一化同剖面扩展
$\mathcal{O}_{ext}$	诊断分数边界观测陪域
π_{ext}	收件箱 $\times$ 发件箱 $\times$ 提交账本 $\times$ 世界观测用于问责和钩子生成的外部治理投影
o_t	步骤 t 的边界观测，即 $o_t = \pi_{ext}(S_t)$
H_{cap}	诊断图跳数范围能力代理可达范围
提交$_{\pi}(act, s, s')$	投影级提交谓词，适用于记账和钩子核算
提交 $_{ext}(act, s, s')$	本体论一阶提交关于外部状态的谓词
外部效应	用于分类的语义谓词
ThroughMembrane(act, s, s')	本体论外部效应或权限表面扩张转换拓扑上的膜中介谓词
$WitnessBind(w, d, act)$	转换拓扑谓词绑定见证记录 w 至裁决类别 d 和动作实例 act
$Anchor(w, d, act)$	见证的仅追加式锚定与决策/动作绑定的记录
$Atomic(d, w, act, s, s')$	确保不可分割性的关系裁决、见证锚定及外部有效转移
重放	确定性裁决类在记录的策略下进行重构版本与上下文抽象

符号	含义
$policy_version_t$	提取的策略版本组件
ctx_abs_t	来自 A_{admt} 用于重放提供的声明的上下文抽象作为重放输入
$T_{int,t}$	内部任务进程
$T_{ext,t}$	外部问责任务进程
外部增量$_t$	外部投影指示事件
受激t	明确授权的指示 STIMULATED 边界行动的步骤 t
外部状态$_t$	累积外部投影计数器
$ObsCap(t)$	监督能力过程
$\Delta ObsCap(t)$	单步监督能力增量
$X_{risk}(t)$	风险加权扩张压力（例如例如 $X_{risk}(t) = \hat{\mu}_H^{risk}(s_t; A_{adm,t}L)$ 在声明配置下）
$[x]_+$	正部算子 $\max(x, 0)$

附录B：扩展证明概要

B.1 不可绕过性与原子性相容性

P-1a 和*P-1b* 共同意味着效应有效性在拓扑上封闭（不存在绕过路径），在时间上也封闭（不存在分阶段效应转移）。任何一项遭到违反，都会使见证语义坍缩为事后日志记录。

B.2 为什么有限观察者结果与重放并不矛盾

在固定的策略/上下文抽象下，重放可以重建已见证转移的裁决类别。有限观察者不可辨识性指出，隐藏内部状态可能无法从边界轨迹中辨识出来。这两种说法针对不同对象，在逻辑上彼此相容。

B.3 稀缺性延续解释

该命题提出了一项假设限定的必然性主张：在假设 (1)-(5) 成立的条件下，持续不停机需要明确授权的 STIMULATED 边界行为。相应的诊断性推论在应用时采用逆否形式：如果观察到持续不停机，却没有此类行为，则表明在所分析的区间内至少有一项假设不成立。

B.4 设计推论 H-1 的解释

设计推论 H-1 刻意采用操作性表述：它为容量规划提供控制理论框架，而非构成律。它的价值在于实用性，因为它将抽象的扩展问题转化为一种可衡量的治理张力，即增长压力与监督处理能力之间的张力（操作形式见附录 D）。

附录 C：工程映射（非规范性）

本理论不规定单一的实现方式，但典型的映射包括：

1. *P-1a* $\rightarrow$ 完全中介架构，并配有明确的不可绕过测试。2. *P-1b* $\rightarrow$ 原子化的决定、 $\rightarrow$锚定和 $\rightarrow$ 效应执行语义。3.*P-1c* $\rightarrow$ 可重放见证模式，并固定策略/上下文版本。4. SECOND_T / SECOND_P $\rightarrow$ 明确的扩张治理和审查队列。5. 设计推论 *H-1* $\rightarrow$ 监督规划仪表板（X_{risk} 与$ObsCap$）。6. S C4 </style}完整性要求： $\rightarrow$ 以密码学方式提交仅可追加的见证（例如哈希链/默克尔根锚定）。

这些映射展示的是结构上的对应关系，而非必需的实现模式。满足构成约束的等效设计均可接受。

实现选择仍不属于本文的规范核心。

附录 D：H-1 操作形式（非规范性）

本附录提供了设计推论 H-1 的工程实例。它不引入新的构成性主张，也不修改第 2–4 节的语义。

D.1 审查积压模型

Let:

- A_t 表示在步骤 t生成的与审查相关的项数。其中包括二阶扩张事件，以及超过声明风险阈值的边界提交。
- R_{obs} 表示审查带宽，以每个步骤可审查的项数计量。
- B_t 表示步骤 t的审查积压（未处理的见证项）。

我们将积压动态建模为离散时间队列：

$$B_{t+1} = \max(0,\ B_t + A_t - R_{obs}).$$

这表达了一个最小的会计不变量：当到达压力超过审查能力时，积压就会增加。

D.2 稳定性条件

只有积压不会持续增长，监督工作才能在运营层面保持稳定。

在标准的平稳排队近似下（单一审查流程、一阶矩有限，并采用长期平均解释），必要的稳定性条件为：

$$\mathbb{E}[A_t] \leq R_{obs}.$$

If

$$\mathbb{E}[A_t] > R_{obs},$$

此时，预期积压会无限增长，审查延迟也会变得无界，从而使见证纪律退化为延迟检查。

D.3 将到达率与扩张压力联系起来

为了将排队动态与该理论的扩张语义联系起来，到达率可以建模为风险加权扩张压力的函数。

Let:

$$A_t = \alpha \cdot [\Delta X_{risk}(t)]_+ + \beta \cdot C_t^{high},$$

其中：

- $\Delta X_{risk}(t) := X_{risk}(t+1) - X_{risk}(t)$;
- $[x]_+ := \max(x, 0)$ 是正部算子;
- C_t^{high} 统计步骤 t 中的高风险边界提交次数;
- α、 $\beta \geq 0$ 为校准系数。

这里：

- 二阶扩张在结构上促成审查负荷，
- 高风险一阶提交按事件促成审查负荷。

这明确表明，审查需求由权限表面增长和边界活动驱动，而不仅仅取决于运行时长。

D.4 工程解释

在相同的近似条件下，操作层面的 H-1 可表述为：

$$\mathbb{E}[A_t] \leq R_{obs}.$$

等价地：

- 如果扩张压力持续导致到达率超过审查带宽，预期积压将无限增长。
- 稳定性要求采取以下措施之一：– 限制扩张动态（降低 A_t），或– 提高监督能力（R_{obs}）。

这种表述避免了对连续可微性的假设，并与主体理论的离散转移语义直接保持一致。

D.5 校准

参数 R_{obs}、 α 和 β 应根据以下内容进行校准：

- 事件报告，
- 未遂事件，
- 红队演练，
- 经验性审查延迟测量。

排队模型是一种保守近似：它不声称具备完备性、最优性或不变性，而是提供了可测量的运行稳定性判据。

D.6 范围

本附录提出了一种工程稳定性模型。它不重新定义二阶转变，不改变本构定律，也不声称稳定监督能够保证安全性。

它提供了一种实用方法，用于检测权限表面增长何时超过审查能力。

参考文献

[1] J. H. Saltzer 和 M. D. Schroeder，《计算机系统中的信息保护》，*IEEE*会刊，第 63 卷，第 9 期，第 1278–1308 页，1975年9月，doi：10.1109/PROC.1975.9939。[2] B. W. Lampson，《关于限制问题的注记》，《*ACM*通讯》，第 16 卷，第 10 期，第 613–615 页，1973年10月，doi：10.1145/362375.362389。

[3] L. Lamport, 《动作时序逻辑》, ACM程序设计语言与系统汇刊, 第 16 卷, 第 3 期, 第 872–923 页, 1994年5月, doi: 10.1145/177492.177726。[4] G. J. Holzmann, 《模型检测器SPIN》, IEEE软件工程汇刊, 第 23 卷, 第 5 期, 第 279–295 页, 1997年5月, doi: 10.1109/32.588521。[5] T. Schick 等, 《Toolformer: 语言模型可以自学使用工具》。访问日期: 2026年1月11日。 [在线]。网址: http: //arxiv.org/abs/2302.04761[6] A. M. Turner、L. Smith、R. Shah、A. Critch 和 P. Tadepalli, 《最优策略倾向于寻求权力》。访问日期: 2026年3月2日。 [在线]。网址: http://arxiv.org/abs/1912.01683[7] E. Debenedetti、J. Zhang、M. Balunović、L. Beurer-Kellner、M. Fischer 和 F. Tramèr, 《AgentDojo: 用于评估大语言模型智能体提示注入攻击及防御的动态环境》。访问日期: 2026年3月2日。 [在线]。网址: http://arxiv.org/abs/2406.13352[8] H. Zhang 等, 《智能体安全基准（ASB）: 对基于大语言模型的智能体中的攻击与防御进行形式化并开展基准测试》。访问日期: 2026年3月2日。 [在线]。网址: http://arxiv.org/abs/2410.02644[9] A. Zou 等, 《人工智能智能体部署中的安全挑战: 大型公开竞赛的见解》。访问日期: 2026年3月2日。 [在线]。网址: http://arxiv.org/abs/2507.20526[10] O. Romanchuk 和 R. Bondar, 《人工智能智能体架构中的语义洗白: 为何工具边界不能构成认知担保》。访问日期: 2026年1月14日。 [在线]。网址: http://arxiv.org/abs/2601 .08333[11] O. Romanchuk 和 R. Bondar, 《责任真空: 规模化智能体系统中的组织失灵》。访问日期: 2026年2月19日。[在线]。网址: http://arxiv.org/abs/2601.15059[12] 《人工智能风险管理框架》, 美国国家标准与技术研究院, 2021年7月, 访问日期: 2026年3月2日。 [在线]。网址: https://www.nist.gov/itl/ai-risk-manageme nt-framework[13] O. Editor, 《2025年OWASP大语言模型应用十大风险》。访问日期: 2026年3月2日。 [在线]。网址: https://genai.owasp.org/resource/owasp-top-10-for-llm-applications-2025/[14] C. Autio 等, 《人工智能风险管理框架: 生成式人工智能概况》, 美国国家标准与技术研究院, 2024年7月, 访问日期: 2026年3月2日。 [在线]。网址: https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-generative-artificial-intelligence[15] F. B. Schneider, 《可执行安全策略》, ACM信息系统安全汇刊, 第 3 卷, 第 1 期, 第 30–50 页, 2000年2月, doi: 10.1145/353323.353382。

[16] M. S. Miller、K. Yee 和 J. Shapiro，《能力神话破除》，2003年。访问日期：2026年3月2日。 [在线]。可访问：https://www.semanticscholar.org/paper/Capability- Myths- Demolished-Miller-Yee/4c53c5d778150c4c734bce4be9844b8d31a9cbd4[17] V. Krakovna、L. Orseau、R. Kumar、M. Martic 和 S. Legg，《使用逐步相对可达性惩罚副作用》。访问日期：2026年3月2日。[在线]。可访问：http://arxiv.org/abs/1806.01186[18] L. P. Kaelbling、M. L. Littman 和 A. R. Cassandra，《部分可观测随机域中的规划与行动》，人工智能，第101卷第1期，第99–134页，1998年5月，doi：10.1016/S0004-3702(98)00023-X。[19] G. N. Nair、F. Fagnani、S. Zampieri 和 R. J. Evans，《数据率约束下的反馈控制：综述》，*IEEE*会刊，第95卷第1期，第108–137页，2007年1月，doi：10.1109/JPROC.2006.887294。[20] T. Everitt、M. Hutter、R. Kumar 和 V. Krakovna，《强化学习中的奖励篡改问题与解决方案：因果影响图视角》。访问日期：2026年3月2日。 [在线]。可访问：http://arxiv.org/abs/1908.04734[21] H. Khalaf、C. M. Verdun、A. Oesterling、H. Lakkaraju 和 F. du P. Calmon，《大语言模型中的推理时奖励黑客》。访问日期：2026年3月2日。 [在线]。可访问：http://arxiv.org/abs/25 06.19248[22] B. Schneier 和 J. Kelsey，《支持计算机取证的安全审计日志》，*ACM*信息系统安全汇刊，第2卷第2期，第159–176页，1999年5月，doi：10.1145/317087.317089。[23] B. Laurie、A. Langley 和 E. Kasper，《证书透明度》，互联网工程任务组，征求意见稿 RFC 6962，2013年6月。doi：10.17487/RFC6962。[24] H. Birkholz、A. Delignat-Lavaud、C. Fournet、Y. Deshpande 和 S. Lasker，《可信且透明的数字供应链架构》，互联网工程任务组，互联网草案 draft-ietf-scitt-architecture-22，2025年10月。访问日期：2026年3月2日。 [在线]。可访问：https://datatracker.ietf.org/doc/draft-ietf-scitt-architecture