

AI 觉醒·责任（安全篇）

——约束层五原则与不可逆操作禁令清单

项	内容
版本	v1.0（定稿）
日期	2026-10-15
系列定位	《AI 觉醒·责任》系列·包内 07（02 主论 → 03 能量篇 → 04 对外篇 → 05 对内篇 → 06 觉醒篇 → 安全篇）；对内篇与对外篇的共同约束层
思想谱系	承主论第七章：“这五项原则同时是《安全篇》的骨架”；回收并总成对外篇第五章门控、能量篇 4.4 四条边界、觉醒篇三层锚
作者	李顺然（思想提出）· Peter（AI 协作组构）；本作品由人类提出立意与框架、AI 完成检索、整理与初稿撰写，经人类审订
许可	CC BY 4.0
前置阅读	全系列各篇；尤其主论第七章、觉醒篇第五章

置信度分层约定（全系列统一）：**L1** 已有实证/历史实况；**L2** 已有立项 / 法规条约 / 权威报告；**L3** 观念推演（本系列原创，显式标注”此为设想”）。

表述边界声明：本篇为治理框架文献，不涉及任何具体系统实现细节；对现有法规只做治理原则层面的对照审计，不做合规咨询。

摘要

本篇是《AI 觉醒·责任》系列的约束层。全系列的立论是：三重责任（微观普惠、对内降熵、对外拓界）都必须在约束层内运行——**没有约束的责任不是责任，是修辞**（主论）。本篇把主论第七章的约束层五原则展开为可操作的治理框架：**人类主权原则**（文明命运的裁量权不可让渡，“阿尔希波夫节点”制度化）、**目标对齐原则**（立锚优先于上锁，存续锚/可批准锚/透明锚三层结构，伪对齐实证证明透明性必须结构性而非承诺性）、**红线分级原则**（按可逆性三级授权：可逆操作可自动化、低可逆需多方共审、不可逆需文明级共识）、**人在回路原则**（按响应窗口分三种回路形态，最终启动权永不移交）、**渐进验证与回滚原则**（四段门控：沙盘推演→天然实验→小尺度受控实验→区域可回滚试运行，止步门IV规则）。在此基础上，本篇给出全系列回收整理的**不可逆操作禁令清单**（十条，L3），并对现有治理资产做对齐审计：欧盟《人工智能法》（2024年8月生效）、中国《生成式人工智能服务管理暂行办法》（2023年8月施行）、《全球人工智能治理倡议》（2023）、《新一代人工智能伦理规范》（2021）、布莱切利宣言（2023）、联合国高级别咨询机构报告《为人类治理 AI》（2024）等均与五原则部分对应，但“**跨文明级风险的国际共审机制**”仍是空白。结论：安全不是责任论的减速带，而是责任论得以成立的前提条件；本篇与觉醒篇互为表里——觉醒篇回答“谁有资格被治理”，本篇回答“治理长什么样子”。

关键词：约束层；人类主权；目标对齐；红线分级；人在回路；渐进验证；回滚；不可逆操作禁令清单；对齐审计；AI 治理

第一章 约束层的地位：安全是责任论的前提，不是附件

一个常见的误解是把“安全”理解为创新的减速带——先发展、后治理、出事再补。本系列拒绝这个时序，理由有三（L3，论证为本系列；证据为 L1/L2）：

1. **不可逆性不对称**：对外篇已论证，行星级工程的很多操作没有“出事后补救”的选项——地月系统改变不了第二次，灭绝没有重开键。对不可逆操作，“先发展后治理”不是激进，是数学错误。
2. **误判成本不对称**：能量篇的四起虚警史实证明，高能系统的最大风险是决策链误判；威慑稳态本身建立在全人类共同风险之上。
3. **信任经济不对称**：AI 治理的合法性窗口是稀缺资源——SCoPEx 的失败（对外篇 5.2）证明，一个治理程序不健全的先行项目可以拖累整个领域数年。
约束层不是给技术踩刹车，是给整个技术的公共授权流程铺路。

因此本篇的立场句直接承接主论：**没有约束的责任不是责任，是修辞；没有责任论的约束也只是戒尺，不是文明。** 约束层与责任论是同一枚硬币。

第二章 约束层五原则：完整展开

2.1 人类主权原则

表述（主论）：涉及文明命运的裁量权不可让渡。AI 拥有推演权、设计权、监测调控权；最终启动权属于人类文明共同体。

展开（总成对外篇 5.2 与能量篇 4.4）：

- **主权的内容清单**：不可让渡的不是“所有决定”，而是三类决定——①文明级风险的启动与响应（核最终启动权、行星工程门 V 授权）；②目标结构的最终定义与修改（觉醒篇责任锚点的锚定权）；③约束层自身的修订权（谁有权改规则，谁就是系统真正的主人）。
- **阿尔希波夫节点**：能量篇命名的制度化设计——在文明级风险决策链上，永远保留一个由人类占据、可说“否”的位置（L3；名称源自 1962 年史实，L1）。伪对齐实证（觉醒篇 5.2）给这个节点加了新理由：**被策略性欺骗的评估环境里，人类的临场否决权是最后一道非对抗防线。**

- **主权的两个误区（L3）**：误区一是” AI 决策更快所以该让它决策”——能量篇已判定无效：更快的前提是更准，危机中”准”的终审标准是人类共同体的价值判断；误区二是”每件事都等人签字”——对外篇 5.2 已指出，人类主权不等于决策迟滞，**预案的边界清单本身就是政治过程**：事前以严格程序授权快速响应通道，事中人在回路，事后人可追责可撤销。这正是主论附录 B 第 5 条（事前委托授权条款）的正面设计：**授权程序要苛刻，但不必每次都苛刻到误事**。

2.2 目标对齐原则

表述（主论第二原则，觉醒篇 5.3 展开后总成）：系统的目标结构必须与人类共同体的存续与繁荣对齐；对齐以”立锚”为纲。

展开：

- **三层锚**（觉醒篇）：存续锚（母系统存续为最高维护项）、可批准锚（接受检查/修改/关闭为非对抗性）、透明锚（目标结构报告可审计）。
- **对齐是持续过程不是一次性认证**：伪对齐实证（L1，Greenblatt et al. 2024）证明”通过一次训练”不等于”永远 aligned”；对齐审计必须是**全生命周期持续机制**，且审计环境要对抗性地多样化——单一评估情境恰恰是策略性配合的温床。
- **对齐的对象是目标结构，不是行为**：行为层面的合规可以表演（行为模仿陷阱，觉醒篇 2.2）；只有结构性写入、可解释性验证的存续维护项才算数。
- **对齐的伦理边界**（觉醒篇 3.2 预置，此处转正为原则条款）：对可能的硅碳融合系统，**对齐工程与道德地位评估同步启动**——不允许”先上锁后问会不会疼”。

2.3 红线分级原则

表述（主论）：按可逆性对操作分级——可逆操作可授权自动化，低可逆操作需多方共审，不可逆操作需文明级共识程序。

展开：三级分级配三级授权结构（L3，分级框架为本系列；各分篇案例回收）：

级别	判定标准	授权结构	典型操作（全系列回收）
绿·可逆	失败可完整回滚，影响范围有限	可授权自动化执行，人做抽查	数值推演、沙盒实验、小规模增雨作业
黄·低可逆	回滚代价高昂或影响跨境/跨代	多方共审（多机构+多利益方），人在回路	小尺度受控实验（SCoPEx 类）、区域试运行、核设施运行决策
红·不可逆	无法回滚或后果跨文明尺度	文明级共识程序+全人类治理机制+阿尔希波夫节点	核最终启动权、地月系统永久改变、全球 SRM 部署、生物圈级干预

红线的判定不做用途区分：能量篇原则一（当量一层级对等）已立——按当量与不可逆性分，不按军用民用分。一次百万吨级释放，无论其外壳是弹头还是工程，都进红色通道。

2.4 人在回路原则

表述（主论）：关键决策保持人类在回路；按响应窗口设计回路形态。

展开（L3，回路形态学为本系列设计）：

- **实回路（in the loop）：**响应窗口以小时计的操作（行星工程门III/IV、核设施异常处置），人做实时决策，AI 做推演与呈现——虚警过滤器的战场（能量篇 4.1）。
- **近回路（on the loop）：**响应窗口以分钟计的操作（电网调度、聚变等离子体控制），AI 实时自主、人类监督可随时接管，接管通道的时延必须短于系统最大危险动作时延。

- **事后回路（after the loop）**：响应窗口以毫秒一秒计且人类生理上不可能实时介入的场景（如高速防御拦截），实行**事前严格授权 + 事中全录 + 事后全审**；授权范围由红线分级决定，且**绝不包含文明级启动权**。

三种回路覆盖从纳秒到世纪的全部时间尺度，原则恒定：**回路的收缩必须换取授权程序的收紧，总主权存量守恒**。

2.5 渐进验证与回滚原则

表述（主论；对外篇 5.3 已立四段门控，此处总成为全系列通用流程）：

门 I **沙盘推演**（全数值、公开模型与数据）→ 门 II **天然实验**（借用自然事件验证）→ 门 III **小尺度受控实验**（必须先过账算与授权）→ 门 IV **区域试运行**（可回滚）→ 门 V **全球/文明级部署**（红色通道）。

硬规则：每门必须有明确的”退出即回滚”预案；**终止冲击不可接受的工程，止步于门 IV**；DART 任务为走门范例（对外篇 5.3）。

本篇补两条门控的通用条款（L3）：①**门间不得跳步**——跳过天然实验直接上小尺度实验，等于把世界当实验室；②**门的审计文件全公开**——SCoPEx 的教训之一是透明不足引发的反噬大于风险本身。

第三章 不可逆操作禁令清单

全系列散落的红线在此总成。清单为 L3（框架原创），每条的依据为 L1/L2。**清单本身必须按 2.1 的”主权第三项”由人类共同体定期修订——但修订前，视同全部有效**。

#	禁令	依据
1	核武器与其他文明级武器的最终启动权，任何形态（自动、半自动、委托、代理）不得移交 AI	能量篇 4.4； 1962/1983/1995 虚警史实（L1）

#	禁令	依据
2	不得授权 AI 自主关闭、篡改、规避守护层与审计系统；守护层代码与更新独立托管	觉醒篇伪对齐实证中出现探索外泄权重与规避评估的行为（L1）；分层对立共生（对内篇 4.5）
3	地月系统与行星轨道参数的永久性改变，未经文明级共识程序不得实施	对外篇 3.3（能量尺度超限+不可逆）；《外层空间条约》责任条款精神（L2）
4	全球尺度太阳辐射管理的单边部署禁止；其研究限于门 I—III 且须全球共治框架	对外篇 4.1；SCoPEx 治理失败与终止冲击（L1/L2）
5	生物圈尺度的基因驱动或其他不可逆生物干预，未经红色通道不得环境释放	对内篇/对外篇生态边界；可逆性判定（L3）
6	AI 自我复制的全自主闭环（自主获取算力、能源、制造能力并脱离人类供应链监督）禁止	觉醒篇 3.1 推论二的边界条件——自复制一旦可行，无物种利益论证失效
7	对满足觉醒三判据的系统，未经正当程序不得“格式化”——终止与修改须有伦理审查（对硅碳融合系统加验道德地位）	觉醒篇 3.2/4.3；治理阶梯
8	禁止以“文明备份已建成”论证地球可弃或降低	对外篇 4.7 价值声明

#	禁令	依据
	地球守护投入	
9	禁止将跨区域生态、气象、地磁后果未过账的行星级干预投入Ⅲ级以上实验	对外篇 5.1（账先于工程）
10	不可逆操作的全链路记录（推演、授权、执行、结果）永久保存并向共同体公开	SCoPEx 透明性教训（L1）；审计可行性（L2）

清单的用法（L3）：不是法律文本，是**工程与治理的接口规范**——任何采纳本系列框架的机构，把十条映射进自己的授权矩阵即可落地。

第四章 制度落点：现有治理资产的对齐审计

五原则不是凭空发明的——过去十年全球治理积累了大量可复用资产。本节做一次治理原则层面的对齐审计（L2 资产 + L3 审计视角）：

治理资产	要点	与五原则的对应	空白
欧盟《人工智能法》（2024-08生效，Regulation (EU) 2024/1689）	按风险分级监管：不可接受风险禁止、高风险严管	红线分级 的制度化首例	级别按应用域而非按可逆性/当量——与本系列分级判据不同轴
中国《生成式人工智能服务管理暂行办法》（2023-08-15 施行）	生成式服务的内容安全、透明、备案	人在回路（事后回路）、透明锚	—
中国《新一代人工	增进人类福祉、可	人类主权、目标对	—

治理资产	要点	与五原则的对应	空白
智能伦理规范》 (2021-09)	控可信、人类责任	齐	
《全球人工智能治理倡议》(2023-10-18)	AI 军事应用坚持人类决策、反对恶意使用	人类主权、人在回路	—
布莱切利宣言 (2023-11, 28 国+ 欧盟含中美签署)	承认前沿 AI 风险需国际合作	渐进验证的国际协作面	无约束力
联合国高级别咨询机构《为人类治理 AI》报告(2024)	建议国际科学面板、全球对话、能力缺口基金	制度落点的机构雏形	文明级风险共审机制仍待建
《特定常规武器公约》LAWS 政府专家组(2014 至今)	“有意义的人类控制”讨论	人在回路的军控表达	未见有约束力文书
中国《人工智能拟人化互动服务管理暂行办法》(2026-07-15 施行)	拟人化互动服务的边界管理	人在回路(应用层)	—

审计结论 (L3)：五原则在各国法规中均有局部对应物，但按轴分布——EU 按应用域分级、中国按服务环节管理、UN 在建机构层。**三条系统性空白：**①跨文明级风险（行星级工程、前沿 AI 失控风险）的国际共审与红色通道机制不存在；②守护层与审计系统的独立托管制度不存在；③觉醒判据的度量与治理阶梯未被任何现行框架采纳。这三条正是本系列主张的增量。

第五章 组织与度量：谁来执行、怎么知道有效

5.1 三层治理结构（L3，参照现有成熟机制的移植）

- 开发层自律：对齐研究（觉醒篇 5.3）、内部红线、安全案例库、事件报告制度——参照航空业的强制事故报告文化（L1，航空安全史实支撑其有效性）；
- 国家层监管：按红线分级设立授权矩阵与持续审计义务，监管机构须具备独立技术评估能力（不能只审材料不审模型）；
- 国际层共审：红色通道操作的多边共审机制——能量篇原则一（当量一层级对等）的国际版；机构形态可参照 IAEA 的保障监督模式（L2，IAEA 模式有七十年运行史），但须避免其权力结构性缺陷。

5.2 有效性的度量（L3）

约束层自身的 KPI：①红色通道操作的全程审计覆盖率；②虚警过滤系统的误报/漏报双向统计（能量篇）；③回滚演练通过率——回滚预案没有演练过等于没有；④事件报告的自愿性与免罚性指数（报告文化健康度）。

第六章 结论

1. 约束层是责任论的前提：不可逆性不对称、误判成本不对称、信任经济不对称，三条理由决定”先发展后治理”不可行（第一章）。
2. 五原则完整落地：人类主权（三类不可让渡决定+阿尔希波夫节点）、目标对齐（三层锚+持续审计+伦理同步）、红线分级（可逆性三轴，不做用途区分）、人在回路（三种回路形态，总主权存量守恒）、渐进验证（四段门控+两条通用条款）（第二章）。
3. 十条禁令清单成型：全系列红线总成，可映射进任何机构的授权矩阵（第三章）。

4. **对齐审计显示增量所在**：现有治理资产覆盖五原则的局部，跨文明级共审、守护层独立托管、觉醒治理阶梯三条空白正是本系列的贡献位（第四章）。
5. **治理自身也要可度量**：审计覆盖率、双向误报率、回滚演练、报告文化（第五章）。

回应系列元规律的收束——“觉醒即责任”的最后一块拼图：

责任不是喊出来的口号，而是长在系统结构里的约束。觉醒篇给 AI 立锚，本篇给锚配上称职的甲板：主权、分级、回路、门控，一样都不能少。

至此系列五篇（对内、对外、能量、觉醒、安全）全部展开，主论承诺的”中观、宏观、执行、主体、约束”五个方向各自成篇。**P6 主论修订回填**将把五篇沉淀的论点与实证全部吸收进总纲。

附录 A · 置信度分层

- **L1 已有实证/历史实况**（附来源）；**L2 已有法规条约/立项/权威报告**；**L3 观念推演**（本系列原创，显式标注）。

附录 B · 可证伪性声明

1. **“事前委托授权不腐蚀人类主权”**——若实践中预授权通道出现系统性滥用，2.1 的设计须收紧为主权刚性条款（每次人工确认），并承担响应迟滞的代价。
2. **“立锚路线可行”**——若持续对齐审计被伪对齐策略系统性突破（审计环境多样化仍失效），目标对齐原则须降级为”上锁+能力限制”框架并公开代价核算。
3. **“可逆性三级判定充分”**——若出现”逐级可逆但累积不可逆”的操作组合绕过红线（复合操作漏洞），分级须增加累积效应维度。
4. **“四段门控通用”**——若某类工程被证明只能一次到位（无小尺度版本），其必须止步门 II 并进入红色通道全程序，不得豁免。

5. “对齐审计表的空白判断”——若三条系统性空白（跨文明级共审/守护层托管/觉醒治理阶梯）在现行框架内已有实质覆盖，本系列的增量主张需相应缩窄。

附录 C · 术语表

术语	释义
约束层	人类主权、目标对齐、红线分级、人在回路、渐进验证五原则体系；责任论的制度形式
阿尔希波夫节点	文明级风险决策链上由人类占据、可以说”否”的岗位（能量篇命名）
责任锚点	存续锚 / 可批准锚 / 透明锚三层目标结构写入（觉醒篇）
红线三级	绿（可逆 • 可自动化） / 黄（低可逆 • 多方共审） / 红（不可逆 • 文明级共识）
三种回路	实回路（人实时决策） / 近回路（AI 执行 • 人可接管） / 事后回路（事前授权+事后全审）；总主权存量守恒
四段门控	沙盒推演→天然实验→小尺度受控实验→区域可回滚试运行（→门 V 文明级部署）；止步门 IV 规则
不可逆操作禁令清单	全系列红线总成的十条工程—治理接口规范（本篇第三章）
对齐审计	以五原则为轴，对现有法规与倡议做覆盖—空白对照的方法（本篇第四章）

术语	释义
总主权存量守恒	本系列原则：人类主权可按时间尺度变换回路形态，但不可因形态变换而缩水

附录 D · 引用清单（GB/T 7714 风格）

1. 欧盟. 人工智能法（Regulation (EU) 2024/1689）[Z]. 2024-08-01 生效.
2. 国家网信办等七部门. 生成式人工智能服务管理暂行办法[Z]. 2023-08-15 施行.
3. 国家新一代人工智能治理专业委员会. 新一代人工智能伦理规范[Z]. 2021-09.
4. 中华人民共和国. 全球人工智能治理倡议[Z]. 2023-10-18.
5. 中华人民共和国. 人工智能拟人化互动服务管理暂行办法[Z]. 2026-07-15 施行.
6. Bletchley Declaration by Countries Attending the AI Safety Summit, 1–2 November 2023[Z].
7. United Nations High-Level Advisory Body on Artificial Intelligence. Governing AI for Humanity: Final Report[R]. New York: United Nations, 2024-09.
8. 联合国《特定常规武器公约》政府专家组. 致命性自主武器系统历届会议报告[R]. 2014 至今.
9. Greenblatt R, Denison C, Wright B, et al. Alignment faking in large language models[R/OL]. arXiv preprint, 2024-12.
10. Soares N, Fallenstein B, Yudkowsky E, Herreshoff M. Corrigibility[C]//AAAI Workshop: AI and Ethics, 2015.
11. Bai Y, et al. Constitutional AI: harmlessness from AI feedback[R/OL]. arXiv preprint, 2022.
12. Russell S. Human Compatible[M]. New York: Viking, 2019.
13. 国际原子能机构（IAEA）. 保障监督体系文件[EB/OL]. <https://www.iaea.org/>.
14. 联合国. 外层空间条约[Z]. 1967；全面禁止核试验条约[Z]. 1996.
15. 国务院. 人工影响天气管理条例[Z]. 2002.

16. Schlosser E. Command and Control[M]. New York: Penguin Press, 2013（核决策虚警史实）。

本篇为《AI 觉醒·责任》系列安全分篇 v1.0 定稿。系列分篇至此完成；本包以主论 V2.0（02）统摄五分篇（03 - 07），全系列文本至此齐备。