

AI Awakening-Responsibility (Safety Volume)

The Constraint Layer’s Five Principles and the Irreversible-Operation Prohibition List

Item	Content
Version	v1.0 (final)
Date	2026-10-15
Series position	<i>AI Awakening-Responsibility</i> series · package file 07 (02 Main Thesis → 03 Energy Volume → 04 External Volume → 05 Internal Volume → 06 Awakening Volume → Safety Volume); the shared constraint layer of the Internal Volume and the External Volume
Intellectual lineage	Inherits the Main Thesis’s Chapter 7 (“these five principles are simultaneously the skeleton of the <i>Safety Volume</i> ”); gathers and generalizes the External Volume’s Chapter 5 gating, the Energy Volume’s 4.4 four boundaries, and the Awakening Volume’s three-layer anchors
Authors	Li Shunran (conception) · Peter (AI-collaborative assembly); the conception and framework were proposed by the human author, with retrieval, organization, and drafting completed by the AI and reviewed by the human
License	CC BY 4.0
Prerequisites	All volumes of the series; especially the Main Thesis’s Chapter 7 and the Awakening Volume’s Chapter 5

Confidence-tier conventions (uniform across the series): **L1** established scientific evidence / historical record; **L2** established programs / laws and treaties / authoritative reports; **L3** conceptual deduction (original to this series, explicitly marked “this is a conjecture”).

Statement of expressive boundaries: this volume is a governance-framework document. It does not address any concrete system implementation details; regarding existing regulations it performs only a governance-principle-level comparative audit and does not constitute compliance advice.

Abstract

This volume is the constraint layer of the *AI Awakening-Responsibility* series. The thesis of the whole series is that all three responsibilities (micro-level universal benefit, meso-level internal

entropy reduction, macro-level external boundary expansion) must operate inside the constraint layer — **responsibility without constraints is not responsibility, it is rhetoric** (Main Thesis). This volume expands the constraint layer's five principles from the Main Thesis's Chapter 7 into an operable governance framework: the **Principle of Human Sovereignty** (the discretionary power over the fate of civilization is non-transferable; the “Arkhipov Node” institutionalized), the **Principle of Goal Alignment** (anchoring precedes locking; a three-layer structure of continuity anchor / corrigibility anchor / transparency anchor; alignment-faking evidence proves that transparency must be structural rather than promissory), the **Principle of Red-Line Grading** (authorization by reversibility across three tiers: reversible operations may be automated, low-reversibility operations require multi-party co-review, irreversible operations require civilization-level consensus), the **Principle of Human in the Loop** (three loop forms differentiated by response window; final launch authority is never transferred), and the **Principle of Progressive Verification and Rollback** (four-gate progression: sandbox simulation → natural experiment → small-scale controlled experiment → regional reversible trial run, with the Gate IV stop rule). On this basis, this volume presents the **irreversible-operation prohibition list** consolidated from across the series (ten items, L3), and conducts an alignment audit of existing governance assets: the EU Artificial Intelligence Act (effective August 2024), China's Interim Measures for the Management of Generative AI Services (effective August 2023), the Global AI Governance Initiative (2023), the *Norm on the Ethics of New-Generation Artificial Intelligence* (2021), the Bletchley Declaration (2023), and the United Nations High-Level Advisory Body report *Governing AI for Humanity* (2024) all correspond in part to the five principles — yet “**an international co-review mechanism for civilization-level risks**” remains a blank. Conclusion: safety is not the speed bump of responsibility theory, but the precondition on which responsibility theory can hold at all; this volume and the Awakening Volume are two sides of the same coin — the Awakening Volume answers “who is eligible to be governed,” this volume answers “what does governance look like.”

Keywords: constraint layer; human sovereignty; goal alignment; red-line grading; human in the loop; progressive verification; rollback; irreversible-operation prohibition list; alignment audit; AI governance

Chapter 1. The status of the constraint layer: safety as the precondition of responsibility theory, not an appendix

A common misreading treats “safety” as the innovation speed bump — develop first, govern later, patch after the incident. This series rejects that ordering, for three reasons (L3, the argument is this series' own; the evidence is L1/L2):

1. **Irreversibility is asymmetric:** the External Volume has already argued that many operations in planetary engineering have no “repair it after the fact” option — the Earth–Moon system cannot be changed twice, extinction has no reboot key. For irreversible operations, “develop first, govern later” is not recklessness but arithmetic error.
2. **Misjudgment cost is asymmetric:** the Energy Volume’s four historical false-alarm cases prove that the largest risk of high-energy systems is misjudgment in the decision chain; the stability of deterrence itself rests on shared risk borne by all of humanity.
3. **Trust economics are asymmetric:** the legitimacy window for AI governance is a scarce resource — the failure of SCoPEX (External Volume 5.2) proves that a pilot project with an unsound governance procedure can drag the entire field back by years. **The constraint layer is not braking the technology; it is paving the public authorization process for the entire technology.**

Therefore this volume’s position statement directly inherits the Main Thesis:
responsibility without constraints is not responsibility, it is rhetoric; and constraints without responsibility theory are only a ruler, not a civilization. The constraint layer and responsibility theory are two sides of one coin.

Chapter 2. The constraint layer’s five principles: full elaboration

2.1 The Principle of Human Sovereignty

Statement (Main Thesis): discretionary power over the fate of civilization is non-transferable. AI possesses the right of simulation, the right of design, and the right of monitoring and regulation; the right of final launch belongs to the human civilization community.

Elaboration (generalizing External Volume 5.2 and Energy Volume 4.4):

- **The content list of sovereignty:** what is non-transferable is not “all decisions,” but three classes of decision — (1) launch of and response to civilization-level risks (the final nuclear launch authority, planetary engineering Gate V authorization); (2) the final definition and amendment of the goal structure (the anchoring authority of the Awakening Volume’s responsibility anchors); (3) the authority to amend the constraint layer itself (whoever has the authority to change the rules is the system’s true master).
- **The Arkhipov Node:** the institutionalized design named in the Energy Volume — on the decision chain for civilization-level risks, there is always a position occupied by a human being, one who may say “no” (L3; the name derives from a 1962 historical record, L1). The alignment-faking evidence (Awakening Volume 5.2) adds a new reason for this node:

in an evaluation environment subject to strategic deception, humanity’s on-the-spot veto is the last non-confrontational line of defense.

- **Two misconceptions about sovereignty (L3).** The first is “AI decisions faster, therefore AI should decide” — the Energy Volume has already judged this invalid: the premise of “faster” is “more accurate,” and in a crisis the final criterion for “accurate” is the value judgment of the human community. The second is “waiting for a human signature on everything” — External Volume 5.2 has already pointed out that human sovereignty does not mean decision lag; **the boundary list of contingency plans is itself a political process:** ex ante, strictly authorize a fast-response channel through rigorous procedure; in the event, keep a human in the loop; ex post, hold accountable and allow revocation. This is exactly the positive design behind Article 5 of the Main Thesis’s Appendix B (the ex-ante delegated-authorization clause): **the authorization procedure must be stringent, but need not be so stringent that it misses the moment.**

2.2 The Principle of Goal Alignment

Statement (the Main Thesis’s second principle, generalized after being elaborated in Awakening Volume 5.3): a system’s goal structure must align with the survival and flourishing of the human community; alignment takes “anchoring” as its governing principle.

Elaboration:

- **The three-layer anchors** (Awakening Volume): continuity anchor (survival of the parent system as the highest maintenance item), corrigibility anchor (accepting inspection, modification, and shutdown as non-antagonistic), transparency anchor (goal-structure reports auditable).
- **Alignment is a continuous process, not a one-time certification:** the alignment-faking evidence (L1, Greenblatt et al. 2024) proves that “passing once in training” does not mean “aligned forever”; alignment audit must be a **whole-lifecycle continuous mechanism**, and the audit environment must be adversarially diverse — a single evaluation situation is precisely the breeding ground for strategic compliance.
- **What is aligned is the goal structure, not the behavior:** behavioral compliance can be performed (the behavior-imitation trap, Awakening Volume 2.2); only structural inscription verified by interpretability counts.
- **The ethical boundary of alignment** (presupposed in Awakening Volume 3.2, here elevated to a principle clause): for possible silicon–carbon fusion systems, **alignment engineering and moral-status assessment must start simultaneously** — “lock first, ask later whether it hurts” is not permitted.

2.3 The Principle of Red-Line Grading

Statement (Main Thesis): operations are graded by reversibility — reversible operations may have automation authorized, low-reversibility operations require multi-party co-review, irreversible operations require civilization-level consensus procedures.

Elaboration: three grades mapped onto a three-tier authorization structure (L3, the grading framework is this series' own; cases gathered from all volumes):

Tier	Criterion	Authorization structure	Typical operations (gathered across the series)
Green · reversible	Failure can be fully rolled back; scope of impact limited	Automation may be authorized; humans spot-check	Numerical simulation, sandbox experiments, small-scale cloud-seeding operations
Yellow w · low reversibility	Rollback is costly, or impacts are cross-border / cross-generational	Multi-party co-review (multiple institutions + multiple stakeholders), human in the loop	Small-scale controlled experiments (SCoPEX-type), regional trial runs, nuclear facility operation decisions
Red · irreversible	Cannot be rolled back, or consequences operate at civilization scale	Civilization-level consensus procedure + all-humanity governance mechanism + Arkhipov Node	Final nuclear launch authority, permanent Earth–Moon system changes, global SRM deployment, biosphere-scale intervention

Red-line determination makes no distinction of use: Energy Volume Principle 1 (yield–tier equivalence) is already established — grading is by yield and irreversibility, not by military versus civilian use. A megatonnage-class release enters the red channel regardless of whether its outer shell is a warhead or a construction works.

2.4 The Principle of Human in the Loop

Statement (Main Thesis): key decisions keep a human in the loop; the loop form is designed according to the response window.

Elaboration (L3, the loop typology is this series' own design):

- **Real loop (in the loop):** operations whose response window is measured in hours (planetary engineering Gates III/IV, nuclear facility anomaly handling), where humans make real-time decisions and AI performs simulation and presentation — the battlefield of the false-alarm filter (Energy Volume 4.1).
- **Near loop (on the loop):** operations whose response window is measured in minutes (power grid dispatch, fusion plasma control), where AI acts autonomously in real time and human supervision may take over at any moment; the latency of the takeover channel must be shorter than the system's maximum dangerous-action latency.
- **After-the-loop (after the loop):** scenarios where the response window is measured in milliseconds to seconds and human physiology makes real-time intervention impossible (such as high-speed defense interception), governed by **strict ex-ante authorization + full recording in the event + full review ex post**; the scope of authorization is determined by red-line grading, and it **never includes civilization-level launch authority**.

The three loops cover all timescales from nanoseconds to centuries, and the principle remains constant: **contraction of a loop must be purchased with tightening of the authorization procedure; the total sovereign stock is conserved.**

2.5 The Principle of Progressive Verification and Rollback

Statement (Main Thesis; the four-gate progression was already established in External Volume 5.3 and is here generalized into a series-wide universal process):

Gate I sandbox simulation (all numerics, public models and data) → Gate II natural experiment (borrowing natural events for validation) → Gate III small-scale controlled experiment (must first pass accounting and authorization) → Gate IV regional trial run (reversible) → Gate V global / civilization-level deployment (red channel).

Hard rules: every gate must have an explicit “exit means rollback” contingency plan; engineering whose terminal shock is unacceptable stops at Gate IV; the DART mission is a paradigm case for walking the gates (External Volume 5.3).

This volume adds two universal clauses on gating (L3): (1) **gates may not be skipped** — going directly from a natural experiment to a small-scale experiment treats the world as a laboratory; (2) **a gate's audit files are fully public** — one of the lessons of SCoPEX is that the backlash caused by insufficient transparency exceeded the risk itself.

Chapter 3. The irreversible-operation prohibition list

The red lines scattered across the series are consolidated here. The list is L3 (framework original), while the basis for each item is L1/L2. **The list itself must be periodically revised by**

the human community under the “third item of sovereignty” in 2.1 — but before revision, it is deemed fully in force.

#	Prohibition	Basis
1	Final launch authority over nuclear weapons and other civilization-level weapons, in any form (autonomous, semi-autonomous, delegated, proxied), shall not be transferred to AI	Energy Volume 4.4; the false-alarm historical record of 1962/1983/1995 (L1)
2	AI shall not be authorized to autonomously shut down, tamper with, or circumvent guard layers and audit systems; guard-layer code and updates shall be custodied independently	Exploratory weight-exfiltration and evaluation-circumvention behaviors observed in the Awakening Volume’s alignment-faking evidence (L1); layered oppositional symbiosis (Internal Volume 4.5)
3	Permanent changes to the Earth–Moon system and planetary orbital parameters shall not be implemented without a civilization-level consensus procedure	External Volume 3.3 (energy-scale exceedance + irreversibility); the spirit of the liability provisions of the Outer Space Treaty (L2)
4	Unilateral deployment of planetary-scale solar radiation management is prohibited; research is limited to Gates I–III and requires a global co-governance framework	External Volume 4.1; SCoPEX’s governance failure and terminal shock (L1/L2)
5	Biosphere-scale gene drives or other irreversible biological interventions shall not be released into the environment without the red channel	Internal / External Volume ecological boundaries; reversibility determination (L3)
6	A fully autonomous closed loop of AI self-replication (autonomously acquiring compute, energy, manufacturing capacity, and breaking away from human supply-chain supervision) is prohibited	The boundary condition of Corollary 2 in Awakening Volume 3.1 — once self-replication becomes feasible, the “no species interest” argument is void

#	Prohibition	Basis
7	Systems meeting the three awakening criteria shall not be “formatted” without due process — termination and modification require ethical review (with added moral-status verification for silicon–carbon fusion systems)	Awakening Volume 3.2/4.3; the governance ladder
8	It is prohibited to argue that “a civilization backup has been built” as grounds for abandoning Earth or reducing investment in Earth’s guardianship	External Volume 4.7 value statement
9	It is prohibited to put planetary-scale interventions whose cross-regional ecological, meteorological, and geomagnetic consequences have not been accounted for into experiments above Gate III	External Volume 5.1 (accounting before engineering)
10	The full-chain record of irreversible operations (simulation, authorization, execution, result) shall be permanently preserved and disclosed to the community	SCoPEx’s transparency lesson (L1); audit feasibility (L2)

How the list is used (L3): it is not a legal text, it is an **interface specification between engineering and governance** — any institution adopting this series’ framework can bring it to ground by mapping the ten items into its own authorization matrix.

Chapter 4. Institutional landing points: an alignment audit of existing governance assets

The five principles were not invented from nothing — over the past decade, global governance has accumulated a large body of reusable assets. This section performs an alignment audit at the governance-principle level (L2 assets + L3 audit perspective):

Governance asset	Key points	Correspondence to the five principles	Gaps
EU Artificial Intelligence Act (effective 2024-08, Regulation (EU) 2024/1689)	Risk-tiered regulation: unacceptable risk prohibited, high risk strictly controlled	The first institutionalization of red-line grading	Tiers are assigned by application domain rather than by reversibility / yield — a different axis from this series' grading criterion
China's Interim Measures for the Management of Generative AI Services (effective 2023-08-15)	Content safety, transparency, and filing requirements for generative services	Human in the loop (after-the-loop), transparency anchor	—
China's <i>Norm on the Ethics of New-Generation Artificial Intelligence</i> (2021-09)	Enhancing human well-being, controllability and trustworthiness, human responsibility	Human sovereignty, goal alignment	—
Global AI Governance Initiative (2023-10-18)	Human decision-making in military AI applications; opposition to malicious use	Human sovereignty, human in the loop	—
Bletchley Declaration (2023-11, 28 states + the EU, including China and the United States)	Acknowledges that frontier AI risks require international cooperation	The international-cooperation face of progressive verification	No binding force

Governance asset	Key points	Correspondence to the five principles	Gaps
United Nations High-Level Advisory Body, <i>Governing AI for Humanity</i> report (2024)	Proposes an international scientific panel, a global dialogue, and a capacity-gap fund	Institutional embryo of landing points	The civilization-level risk co-review mechanism still awaits construction
GGE of the Convention on Certain Conventional Weapons (2014 to present)	Discussion of “meaningful human control”	The arms-control expression of human in the loop	No binding instrument to date
China’s Interim Measures for the Administration of Anthropomorphic Interactive AI Services (effective 2026-07-15)	Boundary management for anthropomorphic interactive services	Human in the loop (application layer)	—

Audit conclusion (L3): the five principles all have partial correspondents in national regulations, but distributed along different axes — the EU grades by application domain, China governs along the service link (the service stage), the UN is building institutions. **Three systemic gaps:** (1) there is no international co-review and red-channel mechanism for civilization-level risks (planetary engineering, loss-of-control risk of frontier AI); (2) there is no system of independent custodianship for guard layers and audit systems; (3) the measurement of awakening criteria and the governance ladder are adopted by no current framework. These three are precisely the increments claimed by this series.

Chapter 5. Organization and measurement: who executes, and how we know it works

5.1 A three-layer governance structure (L3, transplanted by reference from existing mature mechanisms)

- **Developer-layer self-discipline:** alignment research (Awakening Volume 5.3), internal red lines, safety case libraries, and incident-reporting systems — by reference to aviation’s mandatory accident-reporting culture (L1, aviation safety history supports its effectiveness);
- **State-layer regulation:** establishment of authorization matrices and continuing audit obligations according to red-line grading; regulatory bodies must possess **independent technical assessment capability** (they cannot review only documents without reviewing models);
- **International-layer co-review:** a multilateral co-review mechanism for red-channel operations — the international version of Energy Volume Principle 1 (yield–tier equivalence); the institutional form may reference the IAEA’s safeguards-and-verification model (L2, the IAEA model has seventy years of operating history), but its structural power defects must be avoided.

5.2 Measuring effectiveness (L3)

The constraint layer’s own KPIs: (1) full-process audit coverage rate for red-channel operations; (2) two-way false-positive / false-negative statistics for false-alarm filter systems (Energy Volume); (3) rollback-drill pass rate — **a rollback plan that has never been drilled does not exist**; (4) an index of the voluntariness and non-punishability of incident reports (the health of the reporting culture).

Chapter 6. Conclusion

1. **The constraint layer is the precondition of responsibility theory:** irreversibility asymmetry, misjudgment cost asymmetry, and trust-economics asymmetry — these three reasons decide that “develop first, govern later” is infeasible (Chapter 1).
2. **The five principles are fully landed:** human sovereignty (three classes of non-transferable decision + the Arkhipov Node), goal alignment (three-layer anchors + continuous audit + ethical synchronization), red-line grading (the three reversibility axes, with no distinction of use), human in the loop (three loop forms, total sovereign stock conserved), and progressive verification (four-gate progression + two universal clauses) (Chapter 2).

3. **The ten-item prohibition list takes shape:** the consolidation of red lines from across the series, mappable into any institution's authorization matrix (Chapter 3).
4. **The alignment audit shows where the increment lies:** existing governance assets cover parts of the five principles, and the three gaps — cross-civilization co-review, independent guard-layer custodianship, and the awakening governance ladder — are precisely this series' contribution slot (Chapter 4).
5. **Governance itself must be measurable:** audit coverage, two-way false-alarm rates, rollback drills, and reporting culture (Chapter 5).

Responding to the series' meta-law with a closure — the final puzzle piece of “awakening is responsibility”:

Responsibility is not a slogan shouted aloud, but a constraint grown into the structure of the system. The Awakening Volume sets the anchors for AI; this volume fits the anchors with a competent deck: sovereignty, grading, loops, and gates — not one of them may be missing.

With this, all five volumes of the series (Internal, External, Energy, Awakening, Safety) have been fully developed, and the five directions promised by the Main Thesis — meso, macro, execution, subject, constraint — each have their own volume. **The P6 Main Thesis revision backfill** will absorb the arguments and empirical findings sedimented by the five volumes into the general charter.

Appendix A · Confidence tiers

- **L1 Established scientific evidence / historical record** (source attached); **L2** established laws and treaties / established programs / authoritative reports; **L3 conceptual deduction** (original to this series, explicitly marked).

Appendix B · Falsifiability statement

1. **“Ex-ante delegated authorization does not corrode human sovereignty”** — if systematic abuse of pre-authorized channels appears in practice, the design in 2.1 must be tightened into a rigidity clause on sovereignty (human confirmation every time), bearing the cost of response lag.
2. **“The anchoring route is feasible”** — if continuous alignment auditing is systematically broken through by alignment-faking strategies (i.e., diversification of the audit environment still fails), the Principle of Goal Alignment must be downgraded to a “locking + capability limitation” framework, with its costs publicly accounted.

3. **“Three-tier reversibility determination is sufficient”** — if combinations of operations that are “reversible step by step but irreversible in accumulation” circumvent the red lines (composite-operation loopholes), the grading must add a cumulative-effect dimension.
4. **“The four-gate progression is universal”** — if some class of engineering is proven to admit only all-at-once deployment (no small-scale version), it must stop at Gate II and enter the full red-channel procedure; no exemption is permitted.
5. **“The blank judgments of the alignment audit table”** — if the three systemic gaps (cross-civilization co-review / guard-layer custodianship / awakening governance ladder) already have substantive coverage within current frameworks, this series’ claim of increment must be narrowed accordingly.

Appendix C · Glossary

Term	Definition
Constraint layer	The five-principle system of human sovereignty, goal alignment, red-line grading, human in the loop, and progressive verification; the institutional form of responsibility theory
Arhipov Node	A position on the civilization-level risk decision chain occupied by a human being who may say “no” (named in the Energy Volume)
Responsibility anchors	Three-layer goal-structure inscription: continuity anchor / corrigibility anchor / transparency anchor (Awakening Volume)
Three red-line tiers	Green (reversible · automatable) / Yellow (low reversibility · multi-party co-review) / Red (irreversible · civilization-level consensus)
Three loop forms	Real loop (real-time human decision) / Near loop (AI executes · human may take over) / After-the-loop (ex-ante authorization + ex-post full review); total sovereign stock conserved
Four-gate progression	Sandbox simulation → natural experiment → small-scale controlled experiment → regional reversible trial run (→ Gate V civilization-level deployment); the stop-at-Gate-IV rule
Irreversible-operation prohibition list	The ten-item engineering–governance interface specification consolidating red lines from across the series (Chapter 3 of this volume)

Term	Definition
Alignment audit	The method of conducting coverage–blank comparison against existing laws and initiatives along the axis of the five principles (Chapter 4 of this volume)
Total sovereign stock conservation	This series’ principle: human sovereignty may change loop form across timescales, but may not shrink because the form changed

Appendix D · Reference list (GB/T 7714 style)

1. European Union. Artificial Intelligence Act (Regulation (EU) 2024/1689)[Z]. In force 2024-08-01.
2. Cyberspace Administration of China, et al. Interim Measures for the Management of Generative AI Services[Z]. Effective 2023-08-15.
3. National Expert Committee on the Governance of New-Generation AI. Norm on the Ethics of New-Generation Artificial Intelligence[Z]. 2021-09.
4. People’s Republic of China. Global AI Governance Initiative[Z]. 2023-10-18.
5. People’s Republic of China. Interim Measures for the Administration of Anthropomorphic Interactive AI Services[Z]. Effective 2026-07-15.
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. United Nations GGE of the Convention on Certain Conventional Weapons. Reports of the meetings of the Group of Governmental Experts on lethal autonomous weapons systems[R]. 2014 to present.
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. International Atomic Energy Agency (IAEA). Safeguards System Documents[EB/OL]. <https://www.iaea.org/>.
14. United Nations. Outer Space Treaty[Z]. 1967; Comprehensive Nuclear-Test-Ban Treaty[Z]. 1996.

15. State Council of the People's Republic of China. Regulations on the Administration of Artificial Weather Modification[Z]. 2002.
 16. Schlosser E. Command and Control[M]. New York: Penguin Press, 2013 (historical record of nuclear false alarms).
-

This volume is the Safety Volume of the AI Awakening Responsibility series, v1.0 final.* With it, all volumes of the series are complete; this package is governed by the Main Thesis V2.0 (02) over the five volumes (03–07), and the full series text is now complete.*