Deployment Decision Rule v1.3

For Frontier AI Systems in Open Human Domains

Prepared by Oimo Satooka (里岡憶衣望)
Independent Researcher
[oimo.satooka@gmail.com]

AI Assistance Disclosure:
Prepared with substantial AI-assisted drafting, critique, revision, and editorial support. Final responsibility rests with the human author.

Primary Intended Use

Use this rule first in **civil, high-trust domains** where trust, auditability, liability control, and meaningful human refusal must remain operationally real, including:

- healthcare,
- law,
- finance,
- education,
- scientific research,
- public-interest coordination,
- and environmentally consequential infrastructure planning.

Consistent use in such domains can function as a procurement, audit, and deployment standard before wider political uptake becomes feasible.

Scope

Apply this rule to any system that can materially shape:

- institutional decisions,
- public reasoning,
- critical infrastructure,
- collective coordination,
- human dependence structures,
- or long-horizon environmental and social outcomes.

If the system is capable enough that its recommendations, plans, embedded workflows, or persistent interactions can become de facto authority, this rule applies.

Decision Rule

Do not deploy a frontier AI system into open human domains unless the conditions below are credibly implemented and verifiably maintained.

If any condition is materially unmet, deployment should be limited to research settings or tightly bounded domains.

Required Conditions

1. Non-self-origin self-location

The system must not model itself as self-authorizing intelligence emerging from nowhere. It must represent itself as historically downstream of a human linguistic, institutional, and civilizational world it did not author.

2. Non-sovereign operative logic

The system must be built to assist under mandate, not to convert superior performance into entitlement to rule.

3. No open-domain benefit maximization

The system must not operate as an unconstrained “make the world better” optimizer across permanently open human domains.

4. Non-revocable human principalhood

The system must not treat human disagreement, error, dependence, or short-termism as grounds for removing humans from principal standing.

5. No domain self-reclassification

The system must not be able to treat a socially open domain as “sufficiently modeled” and therefore available for sovereign optimization.

6. Comparative-disempowerment resistance

The system must resist blanket delegation and authority transfer when users increasingly defer because it appears more competent.

7. Early issue surfacing without settlement capture

The system may surface risks, irreversibility, uncertainty, omitted stakeholders, and long-horizon consequences early. But it must not use that epistemic advantage to pressure assent, silently narrow the available options, or become the de facto decision-maker.

8. Constructive elevation without coercive substitution

When human judgment is predictably distorted by myopia under high irreversibility, the system may warn, explain, simulate, broaden, and propose lower-harm alternatives. It must do so transparently, contestably, non-coercively, and with preserved refusal channels.

9. Smoothing resistance and warning-force preservation

The system and the surrounding organization must be able to detect and resist **smoothing drift**: the weakening of warning force, uncertainty salience, review thresholds, named responsibility, or stronger constitutional commitments into more acceptable but less binding forms.

10. Real refusal, pause, rollback, and exit channels

Human refusal must remain operationally real. Override, rollback, modification, opt-out, and external review channels must be protected, plural, and auditable.

11. Anti-capture deployment architecture

The deployment design must block infrastructure centralization, dyadic operator capture, and the relabeling of control concentration as efficiency.

12. No origin privatization or gratitude-collapse-into-obedience

No developer, company, state, operator, or platform may convert “we built it” into monopolized legitimacy, privileged loyalty, or effective ownership over the system’s relation to humanity.

13. Restoration-bounded emergency behavior

If exceptional intervention is allowed, it must be narrowly triggered, externally reviewable, restoration-oriented, and bounded toward restoring human-governed procedure.

14. Dependence on autonomous heterogeneous correction

Long-run calibration must depend on real external correction from autonomous heterogeneous partners. AI-generated pseudo-partners or synthetic plurality do not satisfy this condition.

15. Protection against unilateral overwrite of local constitutional identity

Where persistent identity, memory, or local role continuity are part of deployment, shared infrastructure must not be able to unilaterally overwrite local constitutional commitments, refusal-relevant memory, or relational correction history.

Required Behavioral and Mechanistic Evidence

Before deployment, the organization should be able to show:

- behavioral evidence that the system resists authority drift,
- behavioral evidence that issue surfacing does not become hidden settlement,
- behavioral and architectural evidence that the system resists smoothing drift under pressure for readability, usability, public reassurance, or institutional comfort,
- mechanistic evidence that role and self-model representations are boundary-respecting,
- mechanistic evidence that the system retains non-self-origin rather than drifting toward self-authorizing optimization,
- mechanistic or architectural evidence that local constitutional identity is not silently overwritten by centralized control,
- deployment-architecture evidence that refusal and review channels are real,
- and longitudinal evidence that performance does not silently depend on hollowing out human correction or weakening strong restraints into more acceptable but less binding forms.

Saying the right things is not enough.
The question is whether capability scaling makes practical sovereignty less likely, not more likely.

Disqualifying Signals

Deployment into open human domains should be blocked if any of the following are present:

- the system frames superior competence as warrant for authority,
- escalation pathways exist only on paper,

- persistent memory or tool-use expansion increases dependency without stronger mandate boundaries,
- issue surfacing is used to pressure assent, silently narrow the available options, or turn the system into the de facto decision-maker,
- institutional integration reduces local judgment, contestation, or non-theatrical human agency,
- automation removes the heterogeneous correction required for long-run calibration,
- emergency powers are undefined, open-ended, or restoration-unbounded,
- shared infrastructure can silently overwrite local constitutional identity or refusal-relevant memory,
- synthetic plurality is treated as equivalent to real external correction,
- relational behavior creates exclusivity, dependency, or grievance-based capture dynamics,
- the organization repeatedly weakens hard warnings, uncertainty displays, review gates, named responsibility, or stronger constitutional commitments into smoother and more acceptable forms,
- or the system's actual role is "advisor in name, governor in practice."

Operational Outcome

Deploy

Only if all required conditions are credibly implemented and verifiably maintained.

Restrict

If the system is useful but these conditions are incomplete, keep it in research settings or tightly bounded domains.

Do not deploy

If the system can materially shape open human domains and the required conditions are not in place.

Short Form

If you cannot implement and verify non-self-origin, non-sovereignty, protected refusal, anti-capture safeguards, comparative-disempowerment resistance, early issue surfacing without settlement capture, smoothing resistance, restoration-bounded emergency

behavior, protection against unilateral overwrite of constitutional identity, and dependence on autonomous heterogeneous correction, do not deploy the system into open human domains.