AGI-COS 2.1 Policy Submission Edition（全球递交版）。

本版本不是理论文本，而是 直接可递交政府 / 国际组织 / AI 实验室 /政策机构 的标准政策文件结构。

设计原则：

- 西方政策语境可读
- 非意识形态化
- 风险导向
- 可执行
- 可被内部讨论系统接受

这是一份 Policy Instrument，不是宣言。

AGI-COS 2.1

Global AGI Stability Framework

Policy Submission Edition

Document Type: International Policy Framework
Version: AGI-COS 2.1 — Policy Edition
Purpose: Global Risk Coordination Infrastructure for the AGI Era

---

EXECUTIVE SUMMARY

Artificial General Intelligence introduces systemic risks that exceed the response speed of existing governance structures.

The central challenge is not technological capability but global coordination under accelerating uncertainty.

AGI-COS 2.1 proposes a non-sovereign coordination infrastructure enabling governments, industry, and civil society to:

- detect systemic risk early,
- coordinate responses without loss of sovereignty,
- preserve technological innovation,
- prevent irreversible civilizational outcomes.

AGI-COS does not regulate AGI development.

It provides shared analytical and coordination capacity.

---

1. POLICY PROBLEM STATEMENT

1.1 Structural Shift

AGI transforms risk characteristics:

Pre-AGI Risk	AGI-Era Risk
Local	Systemic
Slow	Rapid
Reversible	Potentially Irreversible
Sectoral	Cross-domain

Existing institutions were not designed for this scale.

1.2 Coordination Gap

The primary danger is:

technological acceleration exceeding collective decision capacity.

This creates four global risks:

1. Uncoordinated AI competition
2. Accidental escalation
3. Systemic infrastructure failure
4. Loss of public trust in technological governance

---

2. OBJECTIVE OF AGI-COS 2.1

AGI-COS establishes a Global AGI Stability Framework to:

- enhance coordination speed,
- extend decision reversibility,
- reduce catastrophic risk,
- maintain human decision authority.

---

3. GUIDING PRINCIPLES

Human Decision Authority

AGI provides analysis only. Humans retain decisions.

Sovereignty Preservation

Participation does not limit national autonomy.

Reversibility

All interventions remain reversible.

Minimal Intervention

Innovation continues unless risk thresholds are crossed.

Non-Governance

AGI-COS is not a supranational regulator.

---

4. SYSTEM ARCHITECTURE

AGI-COS operates as shared infrastructure.

4.1 Continuous Risk Sensing

Common indicators:

- System Stress Index (SSI)
- Coupling Index (CI)
- Irreversibility Risk Index (IRI)
- Trust Calibration Score (TCS)
- Civilizational Time Index (CTI)

Purpose:
Create shared situational awareness.

4.2 Perception Audit Layer

Requirements for analytical AI systems:

- confidence disclosure,
- data provenance mapping,
- uncertainty declaration,
- independent verification.

___

4.3 Micro-Intervention Mechanism

Encourages small, reversible policy adjustments rather than disruptive emergency actions.

___

4.4 Crisis Coordination Protocol

When thresholds are exceeded:

- rapid shared analysis,
- coordinated communication,
- voluntary aligned response options.

Human authorities decide actions.

___

4.5 Adaptive Learning System

Global Learning Archive collects:

- incident reports,
- mitigation outcomes,
- best practices.

___

5. GLOBAL ADOPTION STRATEGY

___

5.1 Dual-Track Adoption

Industry Track

AI developers adopt:

- shared metrics,
- audit interfaces,
- safety reporting standards.

Government Track

States gradually institutionalize practices proven operationally useful.

———

5.2 First-Mover Safety Mechanism

Simultaneous activation prevents strategic disadvantage.

Participation becomes effective only after threshold membership is reached.

———

5.3 Strategic Framing

AGI-COS is positioned as:

Stability Infrastructure, not regulation.

———

6. IMPLEMENTATION ROADMAP

———

Phase 1 (Years 1 – 2)

Pilot Programs

- voluntary participation
- shared metrics testing
- non-sensitive cooperation domains

Suggested partners:
digital governance states, multilateral policy groups.

———

Phase 2 (Years 3 – 4)

Operational Expansion

- cross-border monitoring standards
- incident communication protocols
- research transparency practices

Phase 3 (Year 5+)

Institutional Integration

- incorporation into international risk frameworks
- alignment with existing global institutions.

---

7. GOVERNANCE MODEL

AGI-COS governance follows three constraints:

1. No legislative authority
2. No enforcement power
3. No centralized control

Structure:

- Coordination Secretariat
- Independent Audit Nodes
- Participating Stakeholders

---

8. EXPECTED BENEFITS

For Governments

- reduced strategic uncertainty
- improved crisis readiness
- preserved sovereignty

For Industry

- predictable safety expectations
- reduced regulatory fragmentation
- increased public trust

For Society

- safer technological transition

- maintained innovation environment

___

9. RISK MANAGEMENT

AGI-COS explicitly addresses:

- over-centralization risks
- surveillance concerns
- geopolitical misuse

Safeguards include:

- transparency requirements,
- decentralized audit structure,
- voluntary participation.

___

10. RELATION TO EXISTING INITIATIVES

AGI-COS complements:

- national AI strategies,
- OECD AI principles,
- UN technology governance discussions,
- industry safety frameworks.

It is an enabling layer, not a replacement.

___

11. CALL FOR INITIAL PARTICIPATION

Interested stakeholders are invited to engage through:

- pilot participation,
- technical workshops,
- indicator development collaboration,
- MEU experimental programs.

___

FINAL STATEMENT

AGI-COS 2.1 recognizes that the defining challenge of the AGI era is coordination under accelerating complexity.

The framework seeks to ensure that:

technology advances,
human decision authority remains intact,
and civilization retains the capacity to choose its future.

AGI analyzes. Humans decide. Civilization continues.

___